

Automated Information Warfare

For and Against Saturation Attacks

Jwen Fai Low

School of Information Studies

McGill University

Montreal, Quebec, Canada

2023

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

© Jwen Fai Low, 2023

Abstract

Information societies exist under constant threat of information warfare that can be carried out against military and civil targets. Breaching electronic hardware security has proven successful in bringing down power networks, nuclear research programs, and stock exchanges. Similar outcomes, however, can also be achieved by targeting humans that are crucial biological components of socio-technical systems. We speak of influence/information operations (info ops), where information systems' users are both the most vulnerable point of attack and the target of an attack itself. Though a subfield of information warfare, info ops is broad and encompasses a host of methods targeting human cognition. This thesis focuses on informational saturation attacks, the tactic of overwhelming a target with a deluge of information.

E-commerce is an unconventional target for saturation attacks. But, the fact that hundreds of millions of dollars are spent per year by Amazon to combat counterfeit product listings demonstrates the potential for such attacks. While deceptive listings are likely the result of multiple independent scammers seeking financial gain, a powerful malicious actor could conceivably amplify the production of deceptive listings on its rivals' e-commerce platforms to damage their economic interests. We developed a better performing entity matching framework to aid in countering saturation attacks against e-commerce databases.

When mounting a defensive response against saturation attacks, one should take care to prevent indirect self-harm. Consider the case of fake and satirical news. Indiscriminate removal of all non-truthful news would result in the elimination of satirical news, which have an established role of raising political awareness and engagement. By developing an effective discriminator for satire and fake news, we do not combat fake news more effectively *per se*, but we empower decision makers to craft more nuanced anti-misinformation policies that does not harm their own wider interests, e.g., a healthy civil society, in the pursuit of security.

Saturation attacks can also manipulate the perception of popularity for information which are subjective opinions, not just established facts. In such cases, suppression of opinions being covertly backed by saturation attacks risks accusations of censorship. At

other times, the situation may demand that a population be proactively inoculated against an impending saturation attack. The only viable tactic in both scenarios may be to deploy saturation attacks on one's own populace. To simulate swarms of humans and motivated propagandists, i.e., bots, attempting to maximize the adoption of their opinion regarding an issue on an online social media platform, we present an agent-based model for information diffusion. The model's design is informed by an extensive study carried out by us on the opinion dynamics of an actual contested issue on Twitter: wearing masks to protect against COVID-19. The model also incorporates concepts with empirical support such as bandwagoning and peer pressure in modeling the susceptibility of humans to opinion manipulation. By showing the success and failure conditions for saturation attacks on social media, our information diffusion model enables planning of defensive measures against saturation attacks without the ethical quandary of testing out the interventions in the real world.

In summary, this thesis presents four original research projects involving saturation attacks polluting various information environments and offered possible ways to address these attacks. The contribution of our work is the advancement of the nascent field of *social cybersecurity*, which is concerned with safeguarding the weakest link in socio-technical systems of the Information Age, humans.

Résumé

Les sociétés de l'information vivent sous la menace constante d'une guerre de l'information. La violation de la sécurité du matériel électronique a permis de faire tomber des réseaux électriques, des programmes de recherche nucléaire et des bourses. Toutefois, des résultats similaires peuvent également être obtenus en ciblant les êtres humains, qui sont des composants biologiques essentiels des systèmes socio-techniques. On parle d'opérations d'influence/d'information (info ops), où les utilisateurs des systèmes d'information sont à la fois le point d'attaque le plus vulnérable et la cible de l'attaque elle-même. Bien qu'il s'agisse d'un sous-domaine de la guerre de l'information, les opérations d'information sont vastes et englobent une multitude de méthodes ciblant la cognition humaine. Cette thèse se concentre sur les attaques par saturation informationnelle.

Le commerce électronique est une cible peu conventionnelle pour les attaques par saturation. Bien que les listes de produits contrefaits soient probablement le résultat de plusieurs escrocs indépendants à la recherche d'un gain financier, un acteur malveillant puissant pourrait vraisemblablement amplifier la production d'inscriptions trompeuses sur les plateformes de commerce électronique de ses rivaux afin de nuire à leurs intérêts économiques. Nous avons mis au point un cadre d'appariement d'entités plus performant pour aider à contrer les attaques de saturation contre les bases de données de commerce électronique.

Lors de l'élaboration d'une réponse défensive contre les attaques par saturation, il convient de veiller à éviter l'automutilation indirecte. Prenons le cas des fausses nouvelles et des nouvelles satiriques. La suppression sans discernement de toutes les nouvelles non véridiques entraînerait l'élimination des nouvelles satiriques, dont le rôle de sensibilisation et d'engagement politique n'est plus à démontrer. En développant un discriminateur efficace pour la satire et les fausses nouvelles, nous ne combattons pas les fausses nouvelles plus efficacement en soi, mais nous donnons aux décideurs les moyens d'élaborer des politiques de lutte contre la désinformation plus nuancées qui ne nuisent pas à leurs propres intérêts généraux, par exemple une société civile saine, dans la poursuite de la sécurité.

Les attaques par saturation peuvent manipuler la perception de la popularité d'informations

qui sont des opinions subjectives. La suppression d'opinions soutenues secrètement par des attaques par saturation risque d'être accusée de censure. Dans d'autres cas, la situation peut exiger l'inoculation proactive d'une population contre une attaque par saturation imminente. Dans les deux cas, la seule tactique viable peut consister à déployer des attaques par saturation sur sa propre population. Pour simuler des essais d'utilisateurs réguliers et de propagandistes motivés tentant de maximiser l'adoption de leur opinion, nous présentons un modèle de diffusion de l'information basé sur des agents. La conception du modèle s'appuie sur notre étude approfondie de la dynamique de l'opinion sur une question contestée sur Twitter : le port de masques pour se protéger contre le COVID-19. Le modèle intègre également des concepts empiriques modélisant la susceptibilité des humains à la manipulation de l'opinion. En montrant les conditions de succès et d'échec des attaques par saturation, notre modèle permet de planifier des mesures défensives sans le dilemme éthique de tester les interventions dans le monde réel.

En résumé, cette thèse présente quatre projets de recherche originaux impliquant des attaques par saturation polluant divers environnements d'information et propose des moyens possibles pour faire face à ces attaques. La contribution de notre travail est l'avancement du domaine naissant de la *cybersécurité sociale*, qui se préoccupe de la sauvegarde du maillon le plus faible des systèmes socio-techniques, les êtres humains.

Dedication

Love ya, mom and dad! ❤️

Acknowledgement

This thesis is partly supported by the Discovery Grants (RGPIN-2018-03872) from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chairs Program (950-232791 and 950-230623), and also by the MITACS-Accelerate program (IT11024).

My heartfelt thanks goes out to everyone who supported me throughout my quest to pursue a doctorate degree. Benjamin Fung (supervisor), who took on the burden of mentoring me after the departure of Charles-Antoine Julien (ex-supervisor) from McGill, deserves the greatest share of my gratitude as he has been liberal in his supervisory style and with his funds as well, both which are crucial in giving me the space to develop my ideas at my own pace. Many times Benjamin has, unprompted, introduced me to opportunities that would help me develop my skills as a researcher and also further my academic career — teaching positions, extensive network of contacts that span Canada and the globe, etc. — and I am grateful for this proactive approach even if I do not always avail myself of them. Charles-Antoine Julien has shown great conviviality in easing me into Montreal life and his generosity in disbursement prior to his departure has ensured that I remain warm and fed as I near the final stages of my PhD journey, and these acts of kindness have not been forgotten.

Farkhund Iqbal (co-supervisor), though geographically distant, has also provided vital financial and supervisory support that I am thankful for. Rebekah Willson, for willingly donating her time and energy by being on my supervisory committee, has my gratitude. Similarly, I would also like to thank France Bouthillier for serving as my internal examiner and Tuan Le for serving as my external examiner. The co-authors for my manuscripts, Shih-Chia Huang, Pulei Xiong, and Claude Fachkha, have my thanks for lending me their expertise. And finally, my thanks to all the people in the wider School of Information Studies and McGill communities keeping everything running smoothly in the background, who, by virtue of performing their jobs well, will go unnoticed.

Contribution of authors

I, the author of this thesis, conceptualized the studies, carried out all the work necessary to obtain the results, and served as the lead author for the four major chapters in this thesis (Chapters 2, 3, 4, and 5) that correspond to four individual papers that are either published, under review, or in the process of submission.

My supervisor, Benjamin C. M. Fung, who is affiliated with the School of Information Studies at McGill University, has provided invaluable support for all these endeavors. Benjamin C. M. Fung is supported by the Discovery Grants (RGPIN-2018-03872) and CREATE Grants (554764-2021) from the Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chairs Program (950-232791 and 950-230623). Another frequent collaborator is my co-supervisor, Farkhund Iqbal from Zayed University, United Arab Emirates, who is supported by Cluster project (#R16083) and Provost Research Fellowship grant (#R20093) from Zayed University, United Arab Emirates. Ebrahim Bagheri from Ryerson University is supported by the CREATE Grants (554764-2021) from NSERC. Shih-Chia Huang is with the National Taipei University of Technology, Taiwan. Pulei Xiong is with the National Research Council Canada. Claude Fachkha is with the University of Dubai, United Arab Emirates.

Of the four manuscripts in this manuscript-based thesis, two have been successfully published, one (Manuscript 2) in *Expert Systems with Applications* and the other (Manuscript 3) in *IEEE Access*. Of the remaining two, one is under review at *Electronic Commerce Research and Applications (ECRA)* and the other is being formatted for submission to *IEEE Transactions on Cybernetics*.

Manuscript 1 - Chapter 2

For the entity matching paper (Chapter 2), Benjamin C. M. Fung aided me in project conceptualization, validation, supervision, and review editing. Pulei Xiong also provided review editing support. The germ of the paper was a 2018 Mitacs collaboration between Desjardins and my then-advisor Charles-Antoine Julien. The project lead from Desjardins

Lab, Mansib Rahman, suggested that my research interests would converge with the Desjardins's problem of deduplicating customer records across their different business units. Regular monthly reports on the project's progress were submitted to the primary liason from Desjardins Lab, Afuad-Abrar Hossain. Unfortunately, getting access to a slice of Desjardins customer records proved challenging. No data from Desjardins was ever obtained at the end of the Mitacs collaboration period. Ultimately, the study came to rely exclusively on publicly available entity resolution datasets that were initially intended only for prototyping a solution for Desjardins. The paper is under review at *Electronic Commerce Research and Applications (ECRA)*:

Jwen Fai Low, Benjamin CM Fung, Farkhund Iqbal, and Pulei Xiong. Better Entity Matching with Transformers through Ensembles. *Electronic Commerce Research and Applications*, 2023

Manuscript 2 - Chapter 3

The paper on distinguishing between satire and fake news (Chapter 3) benefited from the helpful insights and review editing of my co-authors, Benjamin C. M. Fung, Farkhund Iqbal, and Shih-Chia Huang. The paper has been published in *Expert Systems with Applications* in 2022:

Jwen Fai Low, Benjamin C. M. Fung, Farkhund Iqbal, and Shih-Chia Huang. Distinguishing between fake news and satire with transformers. *Expert Systems with Applications*, 187:115824, January 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.115824. URL <https://www.sciencedirect.com/science/article/pii/S0957417421011891>

Manuscript 3 - Chapter 4

My co-authors for the paper on charting stance and themes of the mask-wearing discourse on Twitter (Chapter 4) are Benjamin C. M. Fung and Farkhund Iqbal. As before, my co-authors helped me with their useful insights and review editing. The paper has been published in *IEEE Access* in 2023:

Jwen Fai Low, Benjamin C. M. Fung, and Farkhund Iqbal. Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets. *IEEE Access*, 11:87009–87033, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3298557

Manuscript 4 - Chapter 5

Jwen Fai Low, Benjamin C. M. Fung, and Farkhund Iqbal. Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets. *IEEE Access*, 11:87009–87033, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3298557

The paper describing an agent-based model of bot-backed co-opetitive information diffusion on social media (Chapter 5) benefited from the help of Benjamin C. M. Fung, Farkhund Iqbal, and Claude Fachkha. The paper is being edited for submission to *IEEE Transactions on Cybernetics*:

Jwen Fai Low, Benjamin C.M. Fung, Farkhund Iqbal, and Claude Fachkha. Carthago Delenda Est: Co-opetitive Indirect Information Diffusion Model for Influence Operations on Online Social Media. *IEEE Transactions on Cybernetics*, 2023

Contents

Abstract	i
Résumé	iii
Dedication	v
Acknowledgement	vi
Contribution of authors	vii
Abbreviations	xiv
List of Figures	xxi
List of Tables	xxiii
1 Introduction	1
1.1 Research context: weaponized librarians	1
1.2 Research motivation	2
1.3 Research problem: informational saturation attacks	4
1.4 Research objective	7
1.5 Novel contributions	8
1.6 Thesis organization	10
2 Better Entity Matching with Transformers through Ensembles	11
2.1 Introduction	12
2.2 Preliminaries and Related Work	14
2.2.1 Entity resolution	14
2.2.2 Machine learning, deep learning, word embeddings	15
2.2.3 Transformers	15

2.2.4	State-of-the-art EM	17
2.2.5	Architectures of transformers as they pertain to <i>AttendEM</i>	18
2.3	AttendEM: a Framework for Entity Matching	21
2.3.1	Problem definition	22
2.3.2	Process overview	22
2.3.3	Text preprocessing	23
2.3.4	Aggregator tokens: dual-CLS and NCLS	23
2.3.5	Ensemble	24
2.4	Framework Assessment	25
2.4.1	Datasets	25
2.4.2	Experiment setup	25
2.4.3	Results	26
2.4.4	Implications	27
2.4.5	Ablation study	29
2.5	Conclusion and Future Work	30
3	Distinguishing between Fake News and Satire with Transformers	35
3.1	Introduction	36
3.2	Preliminaries and related work	38
3.2.1	Defining satire and fake news	38
3.2.2	Automatically classifying satire and fake news	39
3.2.3	Transformers	41
3.3	Identifying satire and fake news with transformers	44
3.3.1	Single-CLS, TB-CLS and multi-CLS for information aggregation	44
3.3.2	Unsupervised pre-training	46
3.3.3	Text preprocessing	48
3.4	Evaluation and discussion	50
3.4.1	Datasets	50
3.4.2	Metrics and comparison basis	53
3.4.3	Metrics and comparison basis for generalizability datasets	54
3.4.4	Results discussion	55
3.4.5	Results discussion for generalizability datasets	58
3.4.6	Potential improvements	59
3.5	Conclusion	61

4	Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets	62
4.1	Introduction	62
4.2	Background	64
4.2.1	COVID-19 conversation on social media	64
4.2.2	Mask-wearing conversation on social media	64
4.2.3	Competing views on social media	66
4.3	Methods	66
4.3.1	Data collection	66
4.3.2	Classification	67
4.3.3	Automated summaries with BART	77
4.3.4	Likes and retweets as proxies for the silent majority	78
4.3.5	Word importance	79
4.4	Analysis	80
4.4.1	Co-occurrence of themes	80
4.4.2	Co-occurrence of themes and stances	81
4.4.3	Word importance	82
4.4.4	Temporal trends	83
4.5	June anomaly: censorship?	102
4.5.1	Twitter’s history of censorship	102
4.5.2	Evidence from our dataset	103
4.6	Discussion	104
4.7	Future Work	105
4.8	Conclusion	106
5	Carthago Delenda Est: Co-opetitive Indirect Information Diffusion Model for Influence Operations on Online Social Media	107
5.1	Introduction	108
5.2	Background	110
5.2.1	Information diffusion models	110
5.2.2	Information operations	120
5.3	The Model — Diluvion	131
5.3.1	Stances and themes	132
5.3.2	Memory and attention	134
5.3.3	Engagement	135
5.3.4	Humans	143
5.3.5	Bots	146

5.3.6	Temporal resolution	148
5.3.7	Graph structure	149
5.3.8	Parameterization	150
5.4	Validation	154
5.5	Simulations	156
5.6	Results and discussion	159
5.6.1	Polarization	161
5.6.2	Impact of bots and bot ratios	168
5.6.3	Stance conversion and retention strategies	170
5.6.4	Themes conservation and propagation	172
5.6.5	Embedded vs. peripheral bots	174
5.7	Limitations and future work	177
5.8	Conclusion	178
6	Final Conclusion	180
6.1	Summary	180
6.2	Discussion	182
6.3	Limitations and future work	184

Abbreviations

β	Standardized (regression) beta coefficient.
<i>p</i> -value	<i>P</i> value, a measure of statistical evidence.
<code>poll_tweet</code>	Name: <code>poll_tweet</code> .
<code>read_conversion</code>	Name: <code>read_conversion</code> .
<code>twitter_sim</code>	Name: <code>twitter_sim</code> , <i>twitter_sim</i> .
2CLS	Dual-CLS.
3D	Three-Dimensional.
5G	Fifth Generation of Mobile Communication.
AB	Abt-Buy.
ABM	Agent-Based Model.
ACE2	Angiotensin-Converting Enzyme 2.
ACM	Association for Computing Machinery.
AdamW	Adam with Decoupled Weight Decay.
AG	Amazon-Google.
AGI	Artificial General Intelligence.
AI	Artificial Intelligence.
AIDS	Acquired ImmunoDeficiency Syndrome.
AIMS	Advanced Impact Media Solutions.
al-Qaeda	the Base.
AMID	Agent-based Multiple Influences Diffusion.
API	Application Programming Interface.
AttendEM	Name: AttendEM.
AUC-ROC	Area Under Curve-Receiver Operating Characteristic.
Auto-GPT	Name: Auto-GPT.
BAKE	Name: BAKE.

BART	Bidirectional and Auto-Regressive Transformer.
BC	Before Christ.
BERT	Bidirectional Encoder Representations from Transformers.
BJP	Bharatiya Janata Party, Indian People's Party.
BotHunter	Name: BotHunter.
BPE	Byte-Pair Encoding.
BTC	Bitcoin.
C	Counter-spreader.
CDC	Centers for Disease Control and Prevention.
CDD	Cyberdissident Diplomacy.
CENTCOM	USCENTCOM, United States Central Command.
Character.AI	Name: Character.AI, character.ai.
ChatGPT	Name: ChatGPT.
CLS	A special classification token.
CNN	Cable News Network.
Consv.	Conservation.
ConvBERT	Name: ConvBERT.
COPD	Chronic Obstructive Pulmonary Disease.
COVID-19	Coronavirus Disease 2019, Covid.
CQP	Code Quality Principle.
CSSE	Center for Systems Science and Engineering.
CTRL	Conditional Transformer Language.
DA	DBLP-ACM.
DAEM	Deep Active Entity Matcher.
Dall-E-2	Name: Dall-E-2, Dall-E 2, DALL-E 2, DALL·E 2.
DBLP	DataBase systems and Logic Programming (original), Digital Bibliography & Library Project (backronym).
DeBERTa	Decoding-Enhanced BERT with Disentangled Attention.
DeepER	Name: DeepER.
DistilBART	Name: DistilBART.
DistilBERT	Name: DistilBERT.
DK	Daley-Kendall.

DoS	Denial-of-Service.
DOT	Digital Outreach Team.
DS	DBLP-GoogleScholar.
dual-CLS	Name: dual-CLS.
EA	Entity Alignment.
EB	Entity Blocking.
EM	Entity Matching.
EMTransformer	Name: EMTransformer.
ER	Entity Resolution.
ER-Magellan	Name: ER-Magellan.
exBAKE	Name: exBAKE.
F1 score	F-score, a measure of accuracy.
FakeBERT	Name: FakeBERT.
FakeNewsNet	Name: FakeNewsNet.
fastText	Name: fastText.
FFP	Filtering FacePiece.
FFR	Filtering Facepiece Respirator.
FIFO	First-In First-Out.
FlauBERT	French Language Understanding via Bidirectional Encoder Representations from Transformers.
FNC	Fake News Challenge.
FZ	Fodors-Zagats.
GAN	Generative Adversarial Network.
GloVe	Global Vectors.
GossipCop	Name: GossipCop.
GPT	Generative Pre-trained Transformer.
GPU	Graphics Processing Unit.
H	Hibernator/Hesitated.
H1N1	Influenza A Haemagglutinin (HA) Subtype 1 Neuraminidase (NA) Subtype 1.

Hamas	Harakat al-Muqawamat al-Islamiyya.
HCV	Hepatitis C Virus.
HHG	Hierarchical Heterogeneous Graph.
Hi	High.
HierGAT	Hierarchical Graph Attention Transformer.
HIV	Human Immunodeficiency Virus.
HuggingFace	Name: HuggingFace.
I	Infected/Ignorant.
IA	iTunes-Amazon.
IBM	International Business Machines.
IC	Independent Cascade.
ID	Identification.
IEEE	Institute of Electrical and Electronics Engineers.
IM	Influence Maximization.
info ops	Information Operations.
iPhone	Name: iPhone.
IRI	International Republican Institute.
IS	Islamic State, see ISIS.
ISCR	Ignorant-Spreader-Counterspreader-Recovered.
ISIS	Islamic State of Iraq and Syria, Islamic State (IS), Islamic State of Iraq and the Levant (ISIL).
ISSRR	Ignorant-Spreader-Spreader-Recovered-Recovered.
IT	Information Technology.
iTunes	Name: iTunes.
JHU	John Hopkins University.
KN95	Name: KN95.
LAST	A special token identifying text span type.
LIAR	Name: LIAR.
LIS	Library and Information Science.
LIWC	Linguistic Inquiry and Word Count.

LT	Linear Threshold.
Mar	March.
MAS	Multi-Agent System.
Md	Medium.
ML	Machine Learning.
MLM	Masked Language Modeling.
N95	Name: N95.
NAFO	North Atlantic Fellas Organization.
NATO	North Atlantic Treaty Organization.
NCLS	Name: NCLS.
NED	National Endowment for Democracy.
neg	Negative.
NELA-GT	News Landscape-Ground Truth.
neu	Neutral.
NLP	Natural Language Processing.
NLTK	Natural Language Toolkit.
NSP	Next Sentence Prediction.
OCCRP	Organized Crime and Corruption Reporting Project.
Odnoklassniki	OK, Classmates.
OEV	Operation Earnest Voice.
OneVoice	Name: OneVoice.
OOT	Out-of-Topic.
OpenAI	Name: OpenAI.
OWID	Our World in Data.
pos	Positive.
PropOrNot	Name: PropOrNot.
PTSD	Post-Traumatic Stress Disorder.
PyTorch	Name: PyTorch.
R	Recovered/Removed.

ReLU	Rectified Linear Units.
RNN	Recurrent Neural Network.
RoBERTa	Robustly Optimized BERT Approach.
RT	Retweet.
RTS	Real-Time Strategy.
S	Susceptible/Spreader/Stifler.
SARS	Severe Acute Respiratory Syndrome.
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2.
SB19	Name: SB19.
SciBERT	Name: SciBERT.
SDN	Software Defined Networking.
SEP	A special separator token.
Seq2SeqMatcher	Name: Seq2SeqMatcher.
SHIR	Susceptible-Hesitated-Infected-Removed.
SIHR	Susceptible-Infected-Hibernator-Removed.
Sim	Simulation.
SimSents	Name: SimSents, Similar Sentences.
SIO	Stanford Internet Observatory.
SIR	Susceptible-Infected-Recovered.
SIRS	Susceptible-Infected-Recovered-Susceptible.
SIS	Susceptible-Infected-Susceptible.
SMSim	Name: SMSim.
SN	Social Network.
SOTA	State-of-the-Art.
SRQ	Stance in Replies and Quotes.
SSRS	International Communications Research (ICR) and Social Science Research Solutions.
StarCraft	Name: StarCraft.
SVM	Support Vector Machine.
TB-CLS	Title-Body-CLS.
TF	Term Frequency.
TF-IDF	Term Frequency–Inverse Document Frequency.

TheMarker	Name: TheMarker.
TM	Text Matching.
TPU	Tensor Processing Unit.
UAV	Unmanned Aerial Vehicle.
UI	User Interface.
UK	United Kingdom.
URL	Uniform Resource Locator.
US	United States.
USAID	United States Agency for International Development.
USE	Universal Sentence Encoder.
V15	Victory 2015.
VADER	Valence Aware Dictionary for sEntiment Reasoning.
VGG	Visual Geometry Group.
Vkontakte	VK, InContact, in contact.
VLDB	Very Large Data Bases.
Vuln.	Vulnerable.
w/o	Without.
WA	Walmart-Amazon.
WHO	World Health Organization.
WordPiece	Name: WordPiece.
WPM	WordPiece Model.
XLNet	Generalized Autoregressive Pretraining for Language Understanding.
XSum	Extreme Summarization.
YouTube	Name: YouTube.
ZunZuneo	Name: ZunZuneo, slang for a Cuban hummingbird's tweet.

List of Figures

2.1	Example of an entity matching problem between two datasets	14
2.2	<i>AttendEM</i> 's workflow	32
3.1	Fake news and satire classification framework	45
4.1	Counts of co-occurring themes	80
4.2	Counts of co-occurring stances and themes	81
4.3	Word importance to different stances and themes of mask-wearing tweets .	82
4.4	Time of day statistics for tweet counts, stances, and themes	84
4.5	Temporal evolution of stance and theme proportions	85
4.6	Temporal evolution of user, tweet, conversation, and total COVID-19 case counts	88
4.7	Temporal evolution of stance and theme counts	89
4.8	Correlation coefficients between tweets and pandemic statistics	91
4.9	Rolling averages of ratios between likes, retweets, and replies grouped by stance	93
4.10	Rolling averages of counts for likes, retweets, and replies grouped by stance	94
4.11	Rolling averages of ratios between likes, retweets, and replies grouped by theme	95
4.12	Rolling averages of counts for likes, retweets, and replies grouped by theme	96
4.13	Rolling averages of counts for likes, retweets, and replies	97
5.1	Example of indirect influence exerted upon an agent	144
5.2	Engagement metrics: real-world data vs simulations with 1,000 agents . . .	156
5.3	Temporal evolution of stance distribution among 1,000 agents	161
5.4	Temporal evolution of follower and followee counts split by an agent's stance	162
5.5	Information flow in the embedded bots scenario	163
5.6	Conversion in the embedded bots scenario	164
5.7	Information flow in the peripheral bots scenario	165

5.8	Conversion in the peripheral bots scenario	166
5.9	Temporal evolution of individual agents' stances in embedded and peripheral scenarios	166

List of Tables

2.1	Statistics of ER-Magellan benchmark datasets	25
2.2	F1 scores on the ER-Magellan datasets	27
2.3	Runtimes for different ensembles and text preprocessing methods. Default uses single-CLS, SimSents and Summed use dual-CLS while Aligned uses NCLS. <i>AttendEM</i> is “Aligned + SimSents + Summed”.	33
2.4	F1 scores for different ensembles. Default uses single-CLS, SimSents and Summed use dual-CLS while Aligned uses NCLS. <i>AttendEM</i> is “Aligned + SimSents + Summed”.	34
3.1	Example of tokenization under single-CLS, TB-CLS, and multi-CLS schemes	46
3.2	Pre-trained weights	48
3.3	Summary statistics for text features found in target and generalizability datasets	50
3.4	Examples from the “Fake News vs Satire” dataset released by Golbeck et al. [187]	51
3.5	Our framework’s performance compared against the state-of-the-art on Golbeck et al. [187]	56
3.6	Scores for DistilBERT-based models in distinguishing between fake and satirical news on Golbeck et al. [187]	57
3.7	Scores for DistilBERT-based models in distinguishing between fake and satirical news on Horne and Adali [220]	58
3.8	Scores for DistilBERT-based models in distinguishing between fake and satirical news on Volkova et al. [482]	59
4.1	DistilBERT classification performance for tweet stances and themes	75
4.2	Summary statistics for the ten largest Twitter conversations in the mask-wearing dataset	100
5.1	Simulation start and end states for 20 scenarios	160

Chapter 1

Introduction

1.1 Research context: weaponized librarians

Information studies, alternatively known as library and information science (LIS) and under which the sub-fields of library science and information science are subsumed, is a field chiefly concerned with meeting the information needs of information seekers [52, 78, 379, 447]. Information professionals serve as intermediaries in addressing information needs. These intermediaries wield great power as they “serve as gatekeepers, or filters, for information systems and collections” [379]. A majority of information professionals are known as *librarians* in the era where physical media, predominantly print media, such as books and periodicals, were the primary storage and access points for information and such information containers are often housed in centralized locations known as libraries. Microfilms, video tapes, and compact discs were kept as well, but their presence in libraries are brief relative to the entire span of the history of libraries [357, 501], and their importance were soon overshadowed by online databases.

Information digitization and the proliferation of wired and wireless networks — the internet — gave birth to online databases and enabled remote access of digitized information. People’s information needs are increasingly being met via electronic interfaces and in a decentralized manner, even if the information repositories are still centralized computer servers. According to a Pew Research Center report on online civic engagement released in December 6, 2022, “the overwhelming majority of people surveyed across 18 countries¹ report using the internet, at least occasionally”, reaching a median of 93% [503]. Portable internet-connected computing devices — smartphones — often serve as handy information portals, with a 18-country median of 85% who report owning such devices from the same

¹Canada, the US, Netherlands, Sweden, Greece, Germany, Italy, the UK, Belgium, France, Spain, Hungary, Poland, Israel, South Korea, Malaysia, Singapore, and Japan

Pew survey [503]. Smartphones and the internet also enable access to platforms where information is broadly disseminated and the roles of information generator, provider, and seeker often blur together — social media — and a 19-country median (with the addition of Australia) of 72% report using social media [503]. In societies where the internet, smartphones, and social media are pervasive, information is primarily cyber-mediated information.

The gatekeepers of cyber-mediated information are no longer flesh-and-blood human librarians but algorithms — automated librarians, in essence. Recommendation engines and search engines are the guise taken by librarians in cyberspace. Not only do such engines populate, in an individualized manner, the feed that social media users scroll through [41, 70], these engines also help users of e-commerce websites retrieve product listings that are most relevant or interesting to the users [122], and these engines also suggest websites that they deem most relevant to their users' queries [151]. Algorithms wield overwhelming power in deciding the type of information that people are exposed to. Algorithms are in essence electric shepherds guiding the flock of information seekers.

Adversarial forces have recognized modern societies' dependence on electric shepherds and these forces are also constantly probing for vulnerabilities that have come into existence due to this newfound dependence. The goal is to exploit the vulnerabilities. The goal, in other words, is to conduct information warfare (see Section 1.2). Overcoming the shortcomings of existing automated librarians and our desire to pre-emptively enhance the capabilities of these librarians as defensive weapons against information warfare — in particular against informational saturation attacks (see Section 1.3) — are major motivating factors driving our research program.

1.2 Research motivation

To be a unipolar hegemon is to maintain constant warfare; potential threats must be nipped in the bud or else the hegemonic status will be lost [183]. In securing itself from external threats, a state would “focus on survival, including territorial integrity and political independence, embedded in the sovereignty of states” [206]. The notion of achieving security against external forces can extend “beyond territory and political autonomy” to defending against other states encroaching upon the “values that a state self-identifies as vital interests” [206]. One of the means for achieving security is balancing, i.e., undermining one's rivals. Balancing can be categorized into military and nonmilitary, and negative and positive [206]. Positive balancing acts improve a state's capabilities relative to its rivals, e.g., strategic technology transfer to allies, while negative balancing acts degrade the capabilities

of its rivals, e.g., economic sanctions.

Information warfare opens up a new front to bloodlessly undermine or even subjugate an adversary. It is therefore a form of negative non-military balancing that states engage in. Near the dawn of the 21st century, in 1995, strategists have recognized the ability for information warfare to severely damage entities dependent on information infrastructures. Civil and military targets are both fair game in full-spectrum warfare: “In states or groups with high technomic capability, the target set for information warfare at the strategic level is wonderfully rich: telecommunications and telephony, space-based sensors, communications relay systems; automated aids to financial, banking, and commercial transactions; supporting power production and distribution systems; cultural systems of all kinds; and the whole gamut of hardware and software that constitutes how the adversary knows and what the adversary believes” [457].

Information warfare should not be confused with cyberwarfare. Information warfare can be conducted in the absence of cyber-mediated information. Information printed on paper or broadcasted over the airwaves are not cyber-mediated but are still valid attack vectors in information warfare; “even preindustrial agrarian or primitive societies are vulnerable to information warfare” [457]. Conversely, cyberwarfare can target computing hardware that is not directly involved in the production, storage, or transmission of information to the public, e.g., electrical grids [451]. Although significant overlap exists between information warfare and cyberwarfare [404], they remain distinct concepts².

The perception that cyberwarfare and information warfare are interchangeable is due to the extreme prevalence of cyber-mediated information in modern societies, as discussed earlier in Section 1.1. And due to the central role algorithms play in this thesis, discussions of information warfare are almost invariably about automated information warfare. The concept of social cybersecurity can also be a source of confusion since it is a *subset* of *both* cyberwarfare and information warfare. The term “social cybersecurity” arose from the need to demarcate itself from orthodox cybersecurity. While orthodox cybersecurity is “focused on machines, and how computers and databases can be compromised”, social cybersecurity is “focused on humans and how these humans can be compromised, converted, and relegated to the unimportant” [87].

Routes for informationally impairing adversaries in modern, electronically intercon-

²Cybernetics itself is a transdisciplinary field meant to study “abstract principles of organization in complex systems” regardless of what the systems consist of, be they “physical, technological, biological, ecological, psychological, social, or any combination of those” [215], although in popular usage and in the modern sense, the abuse of the “cyber” prefix has led to its evolution into a prefix that implies the involvement of electronic computing devices, usually connected in networks [437], e.g., cybersecurity [115]. This thesis uses the popular definition.

nected societies are manifold. The attacks need not target electronic hardware. They can target the biological hardware — humans. Social engineering bypasses security systems by manipulating the system’s users to compromise the system itself [266], e.g., divulging passwords due to phishing attacks. Disinformation campaigns can compromise infrastructures such as electrical grids via false pricing attacks [229]. Automation via algorithms has also enabled the scope of attack to be greatly widened. For instance, in the visual domain, an AI-generated image of an explosion near the Pentagon circulated on social media by verified (i.e., trustworthy) news sources on May 22, 2023 deceived many, causing a “brief dip in the stock market” [316, 377]. This thesis focuses on these sorts of saturation attacks against humans: floods of potentially deceptive information compromising people’s cognitive processes.

1.3 Research problem: informational saturation attacks

Saturation attacks is a tactic of inundating a target to overwhelm its defenses. The tactic is domain-agnostic (in the same way that scale-free networks are) and has been employed to great effect across many domains: unmanned aerial vehicle (UAV)/drone swarms [410] (as practiced by Houthis against Saudi Arabia [417] and Russia against Ukraine [154, 268]); anti-ship missiles [507, 513], cruise missiles [487], and rockets (Hamas against Israel’s iron dome [164]); Denial-of-Service (DoS) attack in Software Defined Networking (SDN) [253, 488]; sensors [47, 289, 434]; and cryptography (quantum [391] and classical [113]).

The saturation point is defined relative to the defender’s capabilities. We illustrate this concept with two disparate problem domains, SDN and anti-ship missiles. For anti-ship missiles, a saturation attack is successful if “the missile quantity [arriving] at the target simultaneously is larger than the complete channel number which enemy’s antiaircraft firepower channel and electronic countermeasure can resist” and results in target destruction [513]. In SDN, “a large number of anomalous packets” that is capable of “[overloading] the buffer memory of network devices” constitutes a successful DoS attack [487]. In both cases, if the offense (an arbitrary quantity of missiles or anomalous packets) exceeds the defense (an arbitrary capacity of missile countermeasures or buffer memory), saturation has been achieved.

Although saturation attacks can undoubtedly be orchestrated in the information sphere owing to human susceptibility to repetition as exemplified by phenomena such as bandwagoning [44, 355] and the spiral of silence [363, 459], there has never been a concerted effort to study massed tactics for compromising human cognition under the umbrella of saturation attacks. The breadth of possible methods for carrying out saturation attacks is

not an argument against adopting a unifying framework, as there have been precedents for meta-studies of seemingly disparate domains, such as unifying the study of dynamic processes of complex socio-technical systems “as diverse as epidemic outbreaks, information spreading and Internet packet routing” under the framework of network models [476]. Thus far, we found only a single group of authors who, via an information- and game-theoretic perspective, have taken saturation attacks into consideration when planning for information security [76, 262, 263]. A shortcoming of the the aforementioned group of authors’ conceptualization is their presupposition that perpetrators of saturation attacks are necessarily disseminators of misinformation; competing opinions of subjective truth values were never considered. This restrictive view extends to seeing saturation attacks as one-sided, even though a saturation attack can serve as defense against another saturation attack, e.g., a UAV swarm contesting against another UAV swarm [410].

The study of informational saturation attacks deserves greater imagination and greater attention from researchers. Greater imagination is required because unconventional vectors of saturation attack would catch a target unaware and unprepared. Our earlier examples of false pricing attack on electrical grids [229] and widespread circulation of fake images to attack the stock market [316, 377] are just a fraction of the many possibilities for the weaponization of mass information in service of goals that in conventional cyberwarfare thinking would have required computer hacking. Greater attention is required because commonly employed mental heuristics make humans particularly susceptible to saturation tactics. Repetition makes messages more credible [323, 323], repetition makes users appear more trustworthy [279, 285, 397], repetition induces others to repeat an action [44, 355], repetition suppresses others from voicing their personal beliefs [363, 459], and repetition saps the will of humans to engage in an issue [76] or with a platform [23, 346].

In this thesis, we will explore conventional and unconventional informational saturation attacks.

Conventional saturation attack targets are information systems where information exchange is the end purpose of the system. Social media are conventional in that people use these platforms to inform and be informed: sharing news, expressing opinions, conversing with other social media users, etc. Conventionality is also about fitting the general public’s conceptions. These conceptions are shaped by the prevailing current of thought in academic circles and in news media, at least within the Anglosphere. At present, they have portrayed social media as the threat du jour. In the 2020 seminal work by Carley introducing social cybersecurity as an emerging science [87], the preoccupation of the author was with online information sharing platforms, i.e., social media. Traditional news media have whipped

up hysteria³ over the supposed manipulation of the 2016 US presidential election in favor of Trump that was carried out over social media through a Russian foreign influence campaign, a hysteria that has not subsided since. Although a recent 2023 investigation into the 2016 election interference concluded that “only 1% of users accounted for 70% of exposures” to Russian disinformation accounts and that they “did not detect any meaningful relationships between exposure to posts from Russian foreign influence accounts and changes in respondents’ attitudes on the issues, political polarization, or voting behavior”, the researchers also noted that “it would be a mistake to conclude that simply because the Russian foreign influence campaign on Twitter was not meaningfully related to individual-level attitudes that other aspects of the campaign did not have any impact on the election, or on faith in American electoral integrity” [145]. Chapters 4 and 5 will address the threat of saturation attacks — motivated actors flooding an information space with specific opinions — on social media.

As social media usage has become near indispensable in modern society [31], a threat from social media is an omnipresent threat. An omnipresent threat offers no escape, so a heightened fear of the threat naturally develops. Hysteria ensues. Words uttered in jest and carefree sharing of news online take on outsized significance due to this heightened fear. They become threatening information that demand to be removed. This magnifies existing abuses of censorial power, producing a chilling effect on free expression, the absence of which can damage the functioning of liberal democracies [16, 226]. Even prior to 2016, protecting a democracy against pro-democracy movements such as Occupy requires “an assault on democracy in the name of democracy” [458], as the covert war on dissent waged by the US that Edward Snowden laid bare so aptly demonstrates [184]. The repression prompted one academic to argue that the US harbors a “corporate-state surveillance apparatus” with an “authoritarian nature” [184]. In 2023, the situation has not improved: careless words on social media invite visits from the Secret Service [162], and the US Army Protective Services Battalion monitors and geolocates social media users to protect “current as well as former high-ranking military officers from ‘assassination, kidnapping, injury or *embarrassment*,’ according to Army records” (emphasis ours) [62]. Chapter 3 will address minimizing collateral damage to civic engagement, namely sharing news online, when defending against saturation attacks.

Unconventional saturation attack targets are information systems where information are

³The term “hysteria” is chosen because scholars and journalists were enthusiastic as opposed to terrified over the prospect of remotely inciting regime change over social media prior to 2016, e.g., [27, 221, 438, 468, 498], but 2016 marked a sudden outburst of research interest in securing society against the threat of manipulation from social media, as a survey counting the number of academic publications on countermeasures against influence operations has found [114].

not the end purpose of the system but act as facilitators for the true end purpose. An example is networked transactional databases. People think in terms of online shopping/retail (e-commerce) systems or insurance claims systems, rarely do they think of them as a set of transactional databases. A conventional way of attacking an unconventional target is to focus on the most obviously socially shared information component of the system, for instance, the spamming of deceptive reviews/opinions on e-commerce websites [43, 228, 277, 278, 332, 400]. An unconventional way of attacking an unconventional target is to focus on the less obvious social components. Many of the most popular and valuable e-commerce companies, such as Amazon, eBay, and Etsy, have large, thriving marketplaces filled by third-party sellers (Etsy and eBay started as exclusively third-party whereas Amazon saw third-party sales eclipsing first-party over the years [252]). Even traditional brick-and-mortar retailers, such as Walmart, have introduced third-party marketplaces when establishing their online presences. Third-party sellers are sufficiently numerous to constitute a community [453], and the sellers' product listings are in essence one form of the sellers' socially shared information [94], making product listings a viable saturation attack vector. Chapter 2 will address defending against saturation attacks targeting e-commerce product listings.

1.4 Research objective

To secure susceptible humans against exploitation via saturation attacks, we aim to expand the extant knowledge of unorthodox saturation attacks and improve existing defensive capabilities against the attacks, with a particular focus on civil targets whose information can be processed by natural language processing (NLP) techniques, specifically texts on electronic medium.

The sub-objectives, each represented by an individual research project and as a chapter in this thesis, are:

- Improving defenses against saturation attacks that damage information systems by imposing additional friction, and therefore costs, to transactions.
- Minimizing collateral damage to civic engagement caused by defensive measures taken against saturation attacks that boost the believability of false rumors.
- Developing an understanding of information propagation dynamics of a topic contested by different crowds.

- Leveraging the newfound understanding to create a more accurate model of competitive and cooperative information diffusion that can be used to simulate interventions against saturation attacks.

The evolutionary trajectory of the sub-objectives can be described as going from tactical to strategical, from reactive to proactive, and from discriminative to non-discriminative. The earlier sub-objectives are tactical in that they engage with information at a low level, i.e., detailed implementation of specific techniques that combat saturation attacks. The later sub-objectives features abstracted information and adopt a high-level strategical overview of the fight, i.e., how a mix of different agents employing various methods can enable or counter saturation attacks. Earlier sub-objectives are reactive in that they are meant to be deployed when an information environment is already subjected to saturation attacks. Later sub-objectives offer the option of warding off potential saturation attacks with saturation attacks, adopting the mantra of a best defense being a good offense. Earlier sub-objectives are discriminative as they deploy classifiers that seek to separate credible information from deceptive ones. Later sub-objectives, although categorizing information into stances and themes, treat all information as subjective opinions and take no position on whether a particular type of information is inherently worthy of suppression or promotion, only the position that the informational interests of one side should be maximized.

1.5 Novel contributions

The original scholarship and distinct contributions to the existing body of knowledge are represented by the research projects reported in this thesis. Each of the project is in service of meeting a research sub-objective laid out in Section 1.4. The contributions, listed in the same order as the corresponding sub-objectives, are:

- An improvement over existing frameworks that relied on transformers to tackle the sub-problem of entity matching (EM) under the problem of entity resolution (ER). Our framework, *AttendEM*, achieved state-of-the-art (SOTA) results by reusing a single transformer architecture multiple times to create distinct models and asynchronously constructing ensembles of out of those models. The framework is agnostic towards the transformer architecture used, as demonstrated by us swapping between DeBERTa and ConvBERT. Identifying duplicate entities is a pre-requisite for taking actions that target duplicate entities, i.e., defending against a saturation attack.
- A novel framework for distinguishing between satire and fake news. At the time of publication, we were the first to apply a transformer, specifically DistilBERT, to the

specific task of distinguishing satire from fake news. The framework first pre-trains the transformer language model within it with news dataset(s) that help align the model to the type of language used in news articles. The pre-trained model is then fine-tuned to perform the task of discriminating between fakes and satires. Our framework achieved SOTA results. Identifying satire from fake news is a pre-requisite for preventing actions targeting fake news from affecting satirical news as well, i.e., preventing collateral damage from defensive measures taken against saturation attack.

- A study on the mask-wearing discourse on Twitter at the outset of the SARS-CoV-2 pandemic. The topic studied is unique in that it lacks entrenched opinions during the start of the data collection period unlike other topics, e.g., vaccination and climate change. The study is conducted from a perspective where all information (tweets) exchanged in the discourse can be categorized into stances and themes. Stance represents the position of a tweet towards the issue of mask-wearing. Themes are representations of different aspects influencing the stance, e.g., whether a person's opposition against masks is backed by authoritative medical sources. The study considered the engagement metrics e.g., number of likes, of different stances and themes, including ratiometrics. An unexpected finding of the study is evidence of decision makers behind the social media platform breaking centralized discussions to the detriment of one stance. This study improved our understanding of information diffusion dynamics of contested topics as mask-wearing was vigorously debated on Twitter, and an improved understanding is crucial to developing an accurate model for simulating saturation attacks as well as interventions against the attacks.
- An agent-based model for co-opetitive information diffusion on social media, *Diluv-sion*. The model's distinguishing feature is its emphasis on indirect influence, where information can reach a target without a friendship tie linking the source and the target, and where auxiliary information in the form of engagement metrics can influence a target's decision to adopt a stance. Two distinct classes of agents exist in the model: bots, who represent motivated propagandists, and humans, who represent naive users. The strengths of the *Diluv-sion* model are demonstrated in simulations of orthodox information operations, e.g., maximizing adoption of one stance; creating echo chambers; inducing polarization; and unorthodox info ops, e.g., simultaneous support of multiple stances as a Trojan horse tactic for the dissemination of a theme. The model incorporated findings from our Twitter study in its development and is capable of simulating saturation attacks as interventions against saturation attacks.

1.6 Thesis organization

To get a sense of the high-level connections bridging the remaining chapters of this thesis, especially the chapters featuring the four original research projects/papers, we refer the reader to Section 1.4, where we discussed the evolutionary trajectory of the research sub-objectives that the papers set out to accomplish.

The remainder of this thesis is structured as follows:

- Chapter 2 discusses the entity matching framework, *AttendEM*.
- Chapter 3 discusses the fakes versus satires discriminator framework.
- Chapter 4 discusses our study on the COVID-19 mask-wearing discourse on Twitter.
- Chapter 5 discusses the information diffusion model, *Diluvision*.
- The final chapter, Chapter 6, summarizes the key contributions made by the research projects presented in this thesis, discusses the overarching implications of these works, and proposes future research directions.

Chapter 2

Better Entity Matching with Transformers through Ensembles

We begin our study of informational saturation attacks by examining a basic and intuitive countermeasure — identification — against the most intuitive form of saturation — homogeneous saturation, i.e., the same thing repeated multiple times. Identification is an intuitive countermeasure because we cannot even begin to devise solutions to a problem if we are not made aware of the problem’s existence in the first place. And, homogeneous saturation is the most intuitive form of saturation because basic duplication, sans embellishments and modifications, requires the least amount of effort.

There are a variety of textual domains where homogeneous saturation can occur, e.g., social media and mainstream media. Identification methods for saturation attacks are readily apparent for conventional domains. For instance, paraphrase identification (e.g., [224, 380]) can be applied to countering repeated opinions on social media. In this chapter, we focus on a solution for an unconventional attack vector: entity matching for e-commerce product listings.

Information influence operations against e-commerce platforms are particularly “insidious, hard to attribute, and even more difficult to fix” [346]. By introducing additional friction to transactions on a platform, one can cause economic slowdown, similar to how sanctions and counter-sanctions damage economies by disrupting the smooth operation of businesses [116]. Already, friction as embodied by issues such as “cart abandonment” (placing items in virtual shopping carts but never completing the checkout and purchase process) costs e-commerce platforms such as Amazon millions to billions of dollars [346].

One vector of attacking e-commerce websites is by saturating them with duplicate product listings. A Transformer language model has been shown to be capable of automatically

generating product descriptions [524], which can potentially ease the creation of duplicate listings en masse. Amazon spending millions of dollars every year to combat deceptive listings [23] demonstrates how a saturation attack can present platforms with a difficult dilemma: escalate anti-deceptive listings expenditures or risk deceptive listings driving away customers. Our research in this chapter aims to fight this form of saturation attack against e-commerce sites by identifying potential duplicates. Our findings are generalizable as the framework was tested on domains other than e-commerce.

2.1 Introduction

As e-commerce grows in popularity and continues displacing brick-and-mortar retail sales, especially during the COVID-19 pandemic as indicated by the US Census Bureau [72], so too does the problem of counterfeit listing on e-commerce sites, in which consumers are presented with listings that attempt to deceive the customers into purchasing non-genuine products. The scale of the counterfeit problem afflicting one leading e-commerce site, Amazon, necessitated the deployment of automated tools incorporating machine learning and an investment of over \$900 million to tackle the problem, which succeeded in blocking 4 billion bad — “fraudulent, infringing, counterfeit, at risk of other forms of abuse, or presenting significant product quality concerns” — listings in 2021 alone [23]. Even if a listing is genuine, duplicates can be part of concerted efforts by malicious actors to flood marketplaces so that their products gain better visibility [56, 125].

The specific problem of detecting counterfeit offers (deceptive listings) have received little attention in scholarly works. However, as “determining offers referring to the same product is a special case of object matching” [29], advances in the more mature field of object matching can have spillover benefits for counterfeit detection.

Object matching is often referred to as *entity matching* (EM), where the objective is to determine if two similar records across databases should resolve into one unique real-world entity. Figure 2.1 shows an example of EM where the left dataset has two candidate pairings found in the right dataset with the second pair being a valid match, i.e., the pair consists of duplicates. As one can see, a typical instance of the EM problem is characterized by the presence of two entities with multiple columns of text, where algorithms may struggle to identify relevant or informative text spans. Algorithms may also lack the capacity to capture all available information of each entity and each column.

Comparing all possible pairs for large databases can be infeasibly time-consuming, hence the existence of an *entity blocking* (EB) phase prior to EM that aims to reduce the number of candidate pairs through comparatively computationally cheap algorithms. The combined

EB and EM phases is called *entity resolution (ER)*.

In this paper, we present *AttendEM*, a classifier framework built to accommodate pre-trained transformer models. The framework is meant to be used during the EM phase of ER and as such has been extensively validated against widely used EM benchmark datasets. Results show that *AttendEM* offers improved performance over previous EM classifiers for a majority of the datasets.

The key contributions made in our paper are as follows:

- *A transformer-based classifier framework for EM.* Our implementation explores pre-trained transformer models and optimization strategies that previous works on transformer-based EM solutions did not, namely domain-agnostic solutions for preserving informative text spans and methods for keeping the classifier from being “distracted” by the less relevant parts of the extra preserved information.
- *Text preprocessing strategies.* Many transformers cannot accommodate all the rich textual information from the multiple columns/attributes that an entity may have, so we investigated if prioritizing the inclusion of sections which are likelier to be informative can improve classifier performance.
- *Information aggregation design for transformers.* We investigated if a single aggregator token, as is standard, is sufficient for aggregating all information within a text span composed of two entities or if spreading out the aggregation over two or more tokens can confer a performance advantage by potentially avoiding the unintentional omission of important information.
- *Attention in the classification head.* The classification head used to process the outputs from the transformer block typically uses only feedforward layers; we examined how adding a self-attention layer to the classification head could improve performance, especially when there is information from multiple aggregator tokens instead of a single token to process.
- *Ensembles.* Different text preprocessing and information aggregation strategies have shortcomings and advantages that could result in performance variance across datasets. Rather than devising a way to select the appropriate strategy for each problem, we investigated the performance improvement that can be found from ensembling the strategies, as the sum of multiple models can often be greater than its parts.
- *Comprehensive empirical evaluation over major benchmark datasets.* *AttendEM* is validated using the same ER-Magellan datasets that were used to evaluate the previous

name	category	modelno	price	title	category	modelno	price	Match
hp c6020b coated paper 1 roll ...	stationery	c6020b	49.88	hp coated paper 36 inches ...	photo paper	c6980a	27.99	✗
				hp coated paper 36 inches ...	roll paper	c6020b	29.99	✓

Figure 2.1: Example of an entity matching problem between two datasets.

state-of-the-art (SOTA) transformer-based EM classifiers. An additional ablation study demonstrates how each of the enhancements introduced within our framework may have contributed to the improved performance.

- *Performance improvement.* *AttendEM* performs better on a majority of datasets than SOTA EM solutions, namely Ditto [282] (average improvement of 0.21% compared against Ditto’s own reported results, 3.93% using DAEM’s Ditto results, 2.99% using HierGAT’s Ditto results), DAEM [223] (average of 0.53%), and HierGAT [515] (average of 0.54%). As stated in the abstract, the improvements observed in *AttendEM* are comparable to two other solutions that claimed to have outperformed Ditto, HierGAT [515] and DAEM [223]. The mean of their percentage improvements over Ditto, calculated using their respective replications of Ditto results, were 2.46% for HierGAT (compared to *AttendEM*’s 2.99%) and 3.42% for DAEM (compared to *AttendEM*’s 3.93%).

2.2 Preliminaries and Related Work

2.2.1 Entity resolution

Entity resolution is “matching records that refer to the same entity across databases” [104] and has also been called deduplication or record linkage. Entity resolution is not to be confused with coreference resolution, which is concerned with finding out whether references within a single document point to the same entity (e.g., whether “I” and “Trump” are the same entity in the following sentence: *Trump was quoted saying, “I would vote for Biden.”*).

While many ER solutions have EB and EM phases, not all do. For instance, there are some unsupervised methods, often approaching the ER problem from an information retrieval and graph-theoretic perspective, that achieve sufficient accuracy in identifying duplicates in a single step for certain datasets such that the need for separate EB and EM phases is obviated, e.g., [200, 247, 259, 522, 530].

The framework introduced in our work, *AttendEM*, is meant only for the EM phase of ER. Having an EM solution that is decoupled from EB allows *AttendEM* to be freely paired

with the many existing EB solutions found in the literature, e.g., [142, 431, 490].

While classification can be performed with the help of humans through methods such as crowdsourcing, e.g., [186] or some hybrid method that mixes human and machine classification, e.g., [202], our work is solely concerned with solutions based purely on machine learning (ML). A number of notable ER and EM solutions relying solely on ML have been proposed in recent years, and they include Magellan, DeepER, DeepMatcher, and Ditto.

2.2.2 Machine learning, deep learning, word embeddings

Deep learning is a specific type of ML featuring many layers of neural networks stacked together, hence the term “deep”. Traditional ML commonly refers to learning solutions prior to the advent of deep learning that do not use these stacked neural layers.

As an example of traditional ML, there is Magellan, an ER solution that uses attribute equivalence and overlap to perform EB and uses algorithms such as random forest, Naive Bayes, and SVM for EM. Magellan also represents the upper limits in performance achievable by traditional machine learning, performing well on Structured and Textual datasets from the ER-Magellan benchmark (average F1 score of 88.8 and 83.4) but still lacking when evaluated against Dirty datasets (68.5) [347].

Deep learning led to the creation of ER solutions, e.g., DeepER [146] and Seq2SeqMatcher [361], and EM solutions, e.g., DeepMatcher [347] that typically outperform solutions using traditional ML. In the case of DeepMatcher, while it has a lower average F1 score than Magellan on Structured datasets (87.9), it performed significantly better when evaluated against the Textual (88.0) and Dirty (87.9) groups of datasets [347]. Much of deep learning’s advantage in *natural language processing* (NLP) tasks can be attributed to the use of word embeddings, e.g., GloVe [381] and fastText [63], which are low-dimensional vector representations of words that can model the relationship between words. A well-known example of the relationship that embeddings can capture is the “*king – man + woman = queen*” equation.

2.2.3 Transformers

Transformers are a further evolution of deep learning. The distinguishing feature of transformers is the presence of an attention mechanism, which allows transformers to assign greater weights to relevant context words when generating the embedding for a word as opposed to static word embeddings such as GloVe and fastText that return the same

vector for a word as they treat context equally (e.g., bank in “river bank”, “bank loan”, and “bank the plane” should be encoded differently but static embeddings do not do this).

Larger — more tunable parameters — transformers being correlated with increased performance in NLP tasks have led to the pursuit ever of larger models, leading to the development of a family of “large language models”.

Many researchers [77, 223, 282] have looked into solving EM with the aid of transformers. Owing to all the interest garnered by the application of transformers to EM, some researchers have even attempted to tease apart the factors contributing to the success of transformers on EM [373].

EMTransformer [77] did not make alterations to the transformer architecture as the authors’ intention is simply to evaluate how vanilla transformer varieties fare in EM. EMTransformer was only evaluated against datasets with long textual data and “dirty” datasets with inappropriate or missing values because the authors regard other benchmark datasets as being close to solved problems (near perfect F1 scores) when using non-transformer EM approaches.

Ditto [282], contemporaneous to the 2020 EMTransformer, emphasized the use of a number of “optimizations” to enhance the performance of the transformer that lies at the heart of Ditto. These optimizations include domain knowledge, data augmentation, and TF-IDF summarization. Domain knowledge aims to increase the amount of machine-readable informative signals by identifying important text spans and fencing them in with special tokens. Data augmentation augments the training data by adding additional training data. The additional training examples are generated by altering existing examples through operations such as deletion, shuffling, and swapping — learning from these “harder” examples may make the model more robust against noise. Summarization attempts to keep transformers’ attention from wandering through unimportant bits of text by distilling long text spans to only the most important segments.

DAEM [223] is a more recent (2022) EM solution featuring transformers and it claims better performance compared to Ditto and all other competing solutions. Unlike EMTransformer and Ditto, DAEM does not rely on pre-trained transformers; it uses the original Transformer’s encoder block sans pre-trained weights. However, in lieu of pre-trained weights, DAEM leveraged the pre-trained embeddings from Facebook’s fastText to vectorize words. The Transformer encoder is used to encode the vectorized words. Another key feature of DAEM is its use of an adversarial active learning framework, which serves similar purposes to Ditto’s data augmentation: active learning presents the model with “good” training examples while adversarial learning augments existing training data for better robustness. To deal with missing data, DAEM also trains an inter-attribute completion

algorithm.

The Hierarchical Graph Attention Transformer, or HierGAT for short, [515] is another recent (2022) EM solution that relies on transformers. The key innovation of HierGAT is its combination of the Transformer attention mechanism with the hierarchical graph attention network into a Hierarchical Heterogeneous Graph (HHG). Within the HHG there exists three layers in the hierarchy, which are, in descending order, entities, attributes, and tokens. The lowest level token nodes relies on contextual word embeddings from pre-trained transformers such as BERT. Each higher level node aggregates the embeddings of the constituent component nodes in the level below it; attribute nodes aggregate token nodes, entity nodes aggregate attribute nodes. A variant of HierGAT, HierGAT+, extends the use of graph from pairwise EM to the problem of collective ER where multiple candidates matches exist. As the scope of problem (pairwise EM vs collective ER) alters the testing basis, we restrict ourselves to examining only HierGAT in our paper.

2.2.4 State-of-the-art EM

In the sea of transformers-based EM solutions, Ditto is considered the quintessential baseline SOTA architecture. Despite its age, Ditto’s consistent presence in benchmarks of multiple EM works published within the last two years (2022–2023) — e.g., EM with hierarchical graph attention networks [515], EM with adversarial active learning [223], domain adaption within EM [467], and explainability of deep learning ER solutions [462] — is a testament to its enduring quality. Works in other fields closely related to ER that were published within the last two years have also favored including Ditto as one of the SOTA baselines for comparison. These include a non-neural entity alignment (EA) solution [274] and an unsupervised text matching (TM) solution [5].

The towering stature of Ditto is cemented by its re-publication as a Special Issue Paper in the VLDB Journal [283]. This paper has been updated with additional tables, data, and discussion giving greater insight into the inner workings of Ditto. Crucially, although separated by 3 years, the experimental results from the 2023 paper re-introducing Ditto (Table 5 in [282]) are exactly the same as the ones found in 2020 (also Table 5 in [283]). The results from papers presenting solutions that have demonstrated themselves to have outperformed Ditto such as HierGAT and DAEM are not found in the 2023 Ditto paper. The new Ditto paper did include results from EMTransformer (Table 13 in [283]) and described EMTransformer’s approach as being similar to “baseline” Ditto without the enhancements that make up the “full” Ditto; the results in the new Ditto paper still shows EMTransformer outperforming the “full” Ditto when evaluated against the Abt-Buy dataset.

In this paper, when validating *AttendEM*, we replicated as closely as possible the experimental conditions that produced the results found in the Ditto paper as we intend to compare *AttendEM* against Ditto.

We did note that the two different 2022 papers that evaluated EM solutions on much of the same ER-Magellan benchmark datasets, DAEM [223] and HierGAT [515], did not reuse the scores from Ditto but evaluated the open source Ditto code in conditions that appeared similar to the original paper’s. The reported scores for Ditto in the DAEM and HierGAT papers are significantly lower than those found in the original Ditto paper. DAEM and HierGAT, naturally, outperformed Ditto in their respective publications. For the sake of fairness and comprehensiveness, we are including in our comparison (1) DAEM’s replication of Ditto and (2) HierGAT’s replication of Ditto as well as (3) DAEM’s own scores, (4) HierGAT’s own scores, and (5) the scores of the second best model in the DAEM paper, EMTransformer [77].

2.2.5 Architectures of transformers as they pertain to *AttendEM*

This section describes the distinguishing architectural features of transformers, with particular emphasis on parts of the architecture that we either modified or repurposed, to facilitate better understanding of the enhancements found in our framework. A more complete picture of the different architectures can be found in the respective papers that introduced them.

The Transformer, first introduced by [475], transformed the NLP research landscape by achieving then state-of-the-art results purely through the use multi-head self-attention and the deliberate omission of recurrence in its architecture, although it is still auto-regressive. Copies of the self-attention mechanism are grouped into encoders and decoders, with encoders responsible for parsing inputs and decoders responsible for generating text as outputs (Section 2.2.5.2). To distinguish between the original Transformer and its many derivatives, we refer to all descendants as *transformers*, always in the lowercase though not necessarily always in plural.

The next major machine learning architecture to emerge in the NLP domain after the introduction of the Transformer is BERT (Bidirectional Encoder Representations from Transformers). By omitting the original Transformer’s decoder layer, which expects temporally masked inputs and is therefore unidirectional, and incorporating only the encoder layer which is direction-agnostic as it attends to both left and right contexts, BERT generates bidirectional encoder representations. BERT improved the model’s language representation through unsupervised pre-training, with masked language modeling (MLM) and next sen-

tence prediction (NSP) as its objectives. Weights for pre-training are used to initialize the transformer during fine-tuning, where BERT is trained to perform supervised downstream tasks such as question answering. Since there are a variety of pre-training and downstream task objectives, BERT allows the output layer on top of the encoder stacks to be swapped based on the task at hand. The swappable output layers are also known as task-specific heads (Section 2.3.4).

DeBERTa [207] improved upon BERT with disentangled attention and enhanced mask decoder. Disentangled attention refers to a mechanism where each input word is represented by two vectors, one for its content and another for its position, instead of one vector combining content and position as is standard. The enhanced mask decoder refers to the improvements DeBERTa made to BERT’s masked language modeling pre-training objective. It is important to note that the decoder in the enhanced mask decoder is not equivalent to the decoder found in the Transformer.

ConvBERT [235] substituted the regular self-attention heads in BERT with a form of mixed attention, which is self-attention with integrated convolution, that better captures local dependencies of different tokens. The convolution operation is called span-based dynamic convolution as it takes a local span of tokens as inputs instead of just a single token as in regular dynamic convolution.

Both DeBERTa and ConvBERT, as derivatives of BERT, are bidirectional.

2.2.5.1 Tokenization

A feature common to transformers is that the first step for processing text spans is to tokenize them. The tokens are then vectorized into embeddings. Different tokenization algorithms such as byte-pair encoding (BPE) and WordPiece exist to break text down into words or subwords. For instance, common words in BPE are tokenized into full words while uncommon words are turned into subwords, e.g., “defragmentation” is decomposed into “defragment” and “ation”. Subword tokenization allows the vocabulary of the tokenizer to remain small but still capable of returning tokens when it encounters out-of-vocabulary words.

In addition to word and sub-word tokens, transformers also utilize special tokens. The Transformer was built with translation in mind, where the number of input tokens often do not match the number of output tokens, hence the need to allow the decoder to dynamically determine the length of the output. The special tokens inform the decoder when to start (start-of-sentence/beginning-of-sentence token) and when to stop (end-of-sentence token) generating text.

In BERT, where text generation is not the goal and decoders are absent, the special

beginning- and end-of-sentence tokens remained in the architecture though their purposes have evolved. BERT's [CLS] serves to both mark the beginning of a text sequence and to aggregate information of the entire sequence. The aggregated information can be used in downstream tasks such as classification, where the classifier head relies only on the embeddings of the [CLS] token to make predictions. BERT's [SEP] marks the end of a text span, but if two text spans are present such as in the case of question-answer pairs, an additional [SEP] separates the first and the second spans.

The flexibility of these special tokens, as can be seen from the transition of their purposes from Transformer to BERT, motivated us to investigate if a new scheme for the insertion of the special tokens, e.g., using the special tokens in numbers greater than prescribed in BERT and Transformer, can improve classification performance (Section 2.3.4).

2.2.5.2 Self-attention, encoders, and decoders

The Transformer essentially consists of stacks of encoders feeding into stacks of decoders. An encoder layer consists of a multi-head self-attention mechanism followed by a fully connected feed-forward network. The self-attention mechanism in encoders attends to all words in a text sequence. A decoder layer differs from encoders in two ways. First, the multi-head self-attention sub-layer in the decoder takes the decoder's own masked output from the previous timestep as its input — the very first timestep uses the special beginning-of-sentence token (Section 2.2.5.1) as the seed as no previous timestep exists. Generation terminates when the decoder generates an end-of-sentence token. The second difference between encoders and decoders is that a decoder layer inserts an additional multi-head self-attention sub-layer that takes the output of the encoder layer as its input.

Having neural networks segmented into encoders and decoders is not unique to the Transformer. Various types of neural network architectures such as autoencoders and Recurrent Neural Networks (RNNs) have also used an encoder and decoder setup. The layout of transformers are not bound to an encoder-decoder arrangement though, as can be seen in the case of BERT, which relied only on encoders because BERT was never intended for text-generation purposes.

A key insight from BERT is that there are no strict rules in the use of the Transformer's constituent parts, e.g., encoders do not necessarily have to be accompanied by decoders. This was one of the motivating factors for our insertion of a decoder layer as an additional discriminator in the classifier head to process the embeddings from multiple aggregator tokens (Section 2.3.5).

Another lesson of transformers and BERT is that the context surrounding a token matters since context is taken into account by the self-attention mechanism when generating

embeddings. Our experiment with different text preprocessing methods (Section 2.3.3) attempts to determine if differing contexts that arise from rearranging spans within text sequences can result in aggregator token output embeddings that are sufficiently varied for ensemble construction to be viable.

2.2.5.3 Task-specific heads and aggregator tokens

Most BERT derivatives use an aggregator/beginning-of-sentence token (Section 2.2.5.1) like BERT. Furthermore, most transformers have adopted the format whereby the “core” self-attention layers are considered separable from the task-specific heads and text tokenization, with new transformer derivatives mostly confining their architectural improvements to the core whilst ignoring the task-specific heads and tokenization. Our framework takes advantage of these conventions: if we restrict our modifications to only altering parts of the architecture outside the core, then regardless of whatever enhancements are made to the core, e.g., the innovations introduced by ConvBERT and DeBERTa, so long as during tokenization the sequences are prepended with aggregator tokens whose output embeddings are used for downstream classification, the classification performance of these transformer variants will still be influenced by our modifications.

Increasing the clarity of aggregated data through the use of multiple aggregator tokens (Section 2.3.4) was one of the modifications we investigated since tokenization lies outside the core and aggregator tokens are a common design feature of BERT derivatives. Finding ways to better utilize the outputs from the core was the other avenue explored, which we achieved through ensembling the output embeddings from text that has undergone different preprocessing and the addition of a decoder layer to the classifier head (Sections 2.3.3 and 2.3.5).

2.3 AttendEM: a Framework for Entity Matching

This section introduces the problem setting (Section 2.3.1), gives high-level descriptions of the *AttendEM* framework (Section 2.3.2) and transformers’ architectures (Section 2.2.5) before delving deeper into the framework’s key components: the text preprocessing step (Section 2.3.3), the information aggregator token design (Section 2.3.4), and the ensembling of individual models (Section 2.3.5).

2.3.1 Problem definition

In the problem of entity resolution (ER), an *entity* refers to a unique real-world object, e.g., product, person, etc. A *record* refers to a mention of an entity found within a collection, i.e., *datasets* that records the attributes of said objects. Given two such datasets D and D' with symmetrical attributes (i.e., all the attributes found in one dataset have their respective equivalents in the other dataset) that contains values which can be represented as strings (for instance, if the values are integers or floats, they can be converted to text), the goal of ER is to find all pairs of records between D and D' that point to the same real-world entity. These record pairs that are found are referred to as *matches*. Pairs of records that point to two distinct entities are *non-matches*.

A standard ER pipeline consists of two stages, entity blocking (EB) and entity matching (EM). The *blocker* in EB aims to filter out the the non-matches found in the cross product $D \times D'$, which typically vastly outnumbers the matches, into C , a subset of $D \times D'$ containing candidate matches. Blockers are assumed to not produce any false negatives. A *matcher* in EM then determines which of the record pairs in C are matches and which are non-matches.

The focus of this paper is on the design of an effective matcher. Specifically, given labeled examples — record pairs which indicate whether the pairs are matches or non-matches — for every dataset pair split into three disjoint subsets — training, validation, and testing — we aim to train the matcher using training data, determine the optimal hyperparameter for the matcher with the validation data, and evaluate the matcher’s performance using testing data.

2.3.2 Process overview

When tasked with determining if a pair of textual entities, A from dataset D and B from dataset D' , are matches, we first obtain different representations of the entities by rearranging the text. The rearranged texts are then tokenized. For each set of tokens, we insert aggregator tokens in them. The number of aggregator tokens as well as the insertion location differs based on how the texts were rearranged. Each set of tokens is fed into the transformer, one set at a time. Once the transformer has processed the last set, the embeddings for all the aggregator tokens found in all the different sets are concatenated. The concatenated embeddings are then fed into a classifier block consisting of self-attention, linear, and softmax layers that outputs a probability of the pair being duplicates. The model is fine-tuned by learning from classification errors. Figure 2.2 shows the EM process within our framework.

2.3.3 Text preprocessing

There are three ways that the raw text can be rearranged within our framework. Two of these are text summarization. The third way is column/attribute alignment between entities of two datasets.

The first summarization method sorts the sentences found in a span of text based on the summed TF-IDF (term frequency-inverse document frequency) values for all the words found in a sentence, with high-scoring sentences taking precedence over others. This approach is known as *Summed*. The second summarization method sorts sentences based on their similarity to a key sentence, with priority given to high-similarity sentences. Similarity is determined by the cosine distance between the TF-IDF vectorized representations of a sentence and the key sentence. The key sentence used is the title/name of an entity. Sentences from the rest of the selected attributes/features are the ones that will be compared against the key sentence and sorted. This approach is known as *SimSents*.

For attribute alignment (*Aligned*), the values from equivalent columns of an entity pair are placed next to each other. For instance, given an entity A within dataset D with the columns “title” and “brand” and an entity B within dataset D' with the columns “name” and “brand”, an entity pair will be concatenated in the following order: A_{title} , B_{name} , A_{brand} , and B_{brand} .

2.3.4 Aggregator tokens: dual-CLS and NCLS

Transformers take tokenized text sequences as inputs. An input sequence of tokens consists of not just the original text but also additional tokens that perform special functions. Often, there are two special tokens, one meant to aggregate information for the text sequence, e.g., [CLS] and another meant to demarcate the end of a sequence, e.g., [SEP]. In pairwise tasks, the tokenization scheme of most transformers take the form of having two text sequences concatenated together with a single aggregator prepended to the first sequence and an end-of-sequence token appended to at the end of each sequence, i.e., [CLS] $tok_{seqA_1}, tok_{seqA_2}, \dots$ [SEP] $tok_{seqB_1}, tok_{seqB_2}, \dots$ [SEP].

Having observed the recent successes in modifying BERT to use multiple CLS tokens in abstractive and extractive text summarization tasks [290], we tested a similar modification for EM. For *Summed* and *SimSents*, we added an extra [CLS] to the start of the second sequence, making the final concatenated sequence [CLS] $tok_{seqA_1}, tok_{seqA_2}, \dots$ [SEP] [CLS] $tok_{seqB_1}, tok_{seqB_2}, \dots$ [SEP]. The maximum number of tokens for each entity in an entity pair is half of a typical transformer’s 512-token limit minus the number of special tokens. This aggregator scheme is referred to as *dual-CLS* to contrast with the original formulation

that uses only one token, single-CLS.

The *NCLS* aggregator scheme is only applicable for Aligned, i.e., entity pairs whose texts have undergone attribute alignment. This scheme places each entity-attribute pair between an aggregator token and an end-of-sequence token. The number of special tokens used depends on the number of attributes that the datasets have. For instance, datasets with three columns, “name”, “brand”, and “description”, will have three sets of special tokens and the final concatenated sequence will take the form of [CLS] $tok_{nameA_1}, tok_{nameA_2}, \dots, tok_{nameB_1}, tok_{nameB_2}, \dots$ [SEP] [CLS] $tok_{brandA_1}, tok_{brandA_2}, \dots, tok_{brandB_1}, tok_{brandB_2}, \dots$ [SEP] [CLS] $tok_{descA_1}, tok_{descA_2}, \dots, tok_{descB_1}, tok_{descB_2}, \dots$ [SEP]. The number of tokens allotted to each entity is the same as in the dual-CLS scheme.

2.3.5 Ensemble

In traditional machine learning, a process known as *ensembling* takes the intermediary predictions from multiple models to derive a final prediction, with Random Forest being a well known example of such ensemble classifiers. With neural language models, a common method for creating an ensemble is by averaging the outputs of models with different architectures. The intuition behind ensembles is that a group of classifiers can collectively compensate for the deficiencies of the individual models.

A barrier to the widespread use of ensembles of large language models is the size of the models as each of the GPUs that the models are trained on usually can only fit a single model in its memory. This memory limitation motivated us to devise a method for creating an ensemble while relying only on a single instance of a single transformer architecture. We accomplished this by varying the inputs given to the transformer so that it produces different outputs. Inputs are varied by using different methods for rearranging the text and inserting aggregator tokens as described in Sections 2.3.3 and 2.3.4.

The transformer takes as input the tokens from only one text variant at a time. From the transformer’s output, only the embeddings for the CLS tokens are retained. These embeddings are concatenated and fed into the classification block.

While the typical classification block consists of linear layers, our framework adds an additional self-attention layer, specifically the decoder block from the original Transformer paper [475], before the linear layers. Since we are not using our framework for predicting words sequentially like the Transformer, we did not need to mask the inputs for the decoder. Inputs for both self-attention sub-layers of the decoder layer we repurposed are the same, which is the output embeddings of the aggregator tokens. Adding self-attention to the classification block is driven by the use of multiple CLS tokens in our framework. The

standard classification blocks being made up of only linear layers may have been sufficient for accommodating the embeddings from a single CLS token, but in the presence of multiple CLS tokens, an attention mechanism may help each CLS token learn its relation to other CLS tokens, which in turn may improve classification performance.

2.4 Framework Assessment

2.4.1 Datasets

Our framework is benchmarked against datasets from [347], which have been used to assess the performance of other EM solutions, e.g., [223, 282]. This collection of datasets, going by the name of ER-Magellan datasets, predate the Magellan ER solution and have been slowly built up over the years by the AnHai Doan research group. Earlier iterations that contain a subset of the datasets found in [347] have also been used to assess the performance of EM solutions, e.g., [146]. The training, validation, and test splits follow a 3:1:1 ratio for all the datasets and the splits we used are the exact same splits used in [223, 282, 347]. The summary statistics for the datasets can be found in Table 2.1.

Datasets	# Candidates (Pairs)	# Matches (Pairs)	Attributes
Amazon-Google (AG)	11,460	1,167	3
Beer	450	68	4
DBLP-ACM (DA)	12,363	2,220	4
DBLP-GoogleScholar (DS)	28,707	5,347	4
Fodors-Zagats (FZ)	946	110	6
iTunes-Amazon (IA)	539	132	8
Walmart-Amazon (WA)	10,242	962	5
Abt-Buy (AB)	9,575	1,028	3
Company	112,632	28,200	3

Table 2.1: Statistics of ER-Magellan benchmark datasets

2.4.2 Experiment setup

Most recent publications have converged upon F1 as the metric of choice for evaluating EM classifiers; we will adhere to this convention as well.

As we have previously noted in Section 2.2.4, there have been conflicting scores reported for Ditto so we will be including in our comparison all versions of published pairwise EM scores [223, 282, 515] for the purposes of transparency.

For EMTransformer, the original paper [77] evaluated EMTransformer only on Dirty datasets and the Abt-Buy dataset while the DAEM paper [223] evaluated EMTransformer only on Structured datasets so there are no overlaps in the reported scores, hence no conflicts.

Like Ditto [282], the number of epochs (5, 10, or 20) used during training in our framework varied based on the dataset. The checkpoint used is the one with the lowest loss instead of the highest F1 score on the validation set. And like Ditto, we first determined which transformers performed best without the modifications found in our framework (text preprocessing, multiple aggregator tokens, and ensemble) before deciding on which architecture to use on which datasets. Note that for EMTransformer [77, 223], the best results from the four different transformer architectures tested (BERT, XLNet, RoBERTa, and DistilBERT) are the ones reported. With our framework, we tested only two architectures, ConvBERT [235] and DeBERTa [207], and settled on using ConvBERT for the Amazon-Google, iTunes-Amazon, Dirty iTunes-Amazon, and Dirty Walmart-Amazon datasets and DeBERTa for the rest. The pre-trained weights for ConvBERT are YituTech/conv-bert-base while DeBERTa used microsoft/deberta-v3-base. ConvBERT has 106M parameters. DeBERTa has 184M parameters. The mean of the F1 scores from 5 repeated runs are reported.

All experiments are conducted on a single Nvidia Titan Xp with 12GB of memory alongside an Intel Xeon E5-2697 v4 @ 2.30GHz. The *AttendEM* framework is implemented in PyTorch. Transformer models are from the HuggingFace transformers package. When using ConvBERT, the decoder block within the classifier block has 1 layer with 4 attention heads. DeBERTa uses 1 layer with 8 attention heads. The learning rates are $2e-5$ with a weight decay of $1e-2$ for the transformer and $1e-5$ with a weight decay of $1e-3$ for everything else. Learning rate decreases linearly with zero warmup steps. For the Python packages random, numpy, and pytorch, we used a random seed of 42.

2.4.3 Results

Table 2.2 compares the performance of *AttendEM* against several other SOTA EM frameworks.

Our framework is notably robust against Dirty datasets, where it outperforms all the EM classifiers that it was compared against in two instances and is on par in two other instances (after rounding based on the number of significant figures reported by the competing solution).

Results are mixed with Structured datasets, but *AttendEM* still made a strong showing.

Against Ditto, DAEM’s reported Ditto results, and HierGAT’s reported Ditto results, *AttendEM* outperformed Ditto in all instances — unsurprising since DAEM and HierGAT themselves outperformed their Ditto replication attempts. Against DAEM, *AttendEM* achieved similar or better results in four out of seven datasets. Against HierGAT, *AttendEM* also achieved similar or better results in four out of seven datasets. DBLP-GoogleScholar and Walmart-Amazon are the two datasets in common where *AttendEM* performed worse compared to DAEM and HierGAT. Against EMTR, *AttendEM* has the lead in all datasets save for DBLP-GoogleScholar. Evidently, *AttendEM* is weak against DBLP-GoogleScholar, although there are no immediately obvious reasons for its underperformance as the content and structure of DBLP-GoogleScholar is highly similar with DBLP-ACM, in which *AttendEM* was able to achieve performance comparable with other EM solutions.

With Textual data, out of the four solutions examined, our framework is outperformed by Ditto and EMTransformer, although *AttendEM* consistently placed second.

Datasets	DAEM (DAEM)	HierGAT (HierGAT)	Ditto (Ditto)	Ditto (DAEM)	Ditto (HierGAT)	EMTr (EMTr)	EMTr (DAEM)	<i>AttendEM</i>	$\Delta F1$
Structured									
Amazon-Google	73.1	76.4	75.58	72.9	74.1	-	71.4	77.67	+1.27
Beer	90.3	<u>93.3</u>	94.37	84.4	84.6	-	87.5	91.04	-3.33/+6.64/+6.44
DBLP-ACM	98.9	<u>99.1</u>	98.99	98.7	99.0	-	98.9	99.12	+0.02
DBLP-GoogleScholar	97.3	96.3	95.60	95.6	95.8	-	96.0	95.85	-1.45
Fodors-Zagats	100.0	100.0	100.00	<u>99.1</u>	98.1	-	100.0	100.00	+0.00/+0.90/+1.90
iTunes-Amazon	100.0	96.3	97.06	92.4	92.3	-	93.1	98.90	-1.10
Walmart-Amazon	<u>87.7</u>	88.2	86.76	83.1	85.8	-	85.9	<u>86.81</u>	-1.39
Dirty									
DBLP-ACM	-	99.1	99.03	-	98.9	98.9	-	<u>99.08</u>	-0.02
DBLP-GoogleScholar	-	<u>95.8</u>	95.75	-	95.4	95.6	-	95.89	+0.09
iTunes-Amazon	-	94.7	<u>95.65</u>	-	92.9	94.2	-	96.45	+0.80
Walmart-Amazon	-	86.3	85.69	-	82.6	85.5	-	<u>86.29</u>	-0.01
Textual									
Abt-Buy	-	89.8	89.33	-	88.9	90.9	-	<u>90.11</u>	-0.79
Company	-	88.2	93.85	-	87.5	-	-	<u>92.50</u>	-1.35/+4.30

Table 2.2: F1 scores on the ER-Magellan datasets. Scores for DAEM are from the DAEM paper [223]. Scores for HierGAT are from the HierGAT paper [515]. Scores for Ditto are from the Ditto [282, 283], DAEM [223], and HierGAT [515] papers. Scores for EMTransformer are from the EMTransformer [77] and DAEM [223] papers. The $\Delta F1$ is a comparison between *AttendEM* and the best (or next best) performing model. If more than one $\Delta F1$ are reported side-by-side, the leftmost $\Delta F1$ represents a comparison with results published in the original paper while the remaining $\Delta F1$ are with results from other papers’ replication attempts. The highest scores are in bold, second best are underlined.

2.4.4 Implications

DAEM and Ditto both heavily emphasized the effort spent in cleaning up the text during the text preprocessing stage. DAEM uses inter-attribute completion to fill in missing values (this

requires that the missing value for an attribute can be found within the values for other attributes). Ditto uses domain knowledge to identify text spans so that special tokens can be inserted to make it easy for the transformer to identify the span type (e.g., the phone number “(866) 246-6453” is tokenized as “(866) 246 - [LAST] 6453 [/LAST]” where [LAST] indicates the last 4 digits of the phone number). Domain knowledge is also used to normalize non-standard entries (e.g., rewriting “5 %” and “5.00 %” to “5.0 %”). Another aspect of DAEM and Ditto is their use of additional training data. DAEM uses active learning and selects examples high in informativeness and representativeness for the classifier to learn from. Furthermore, DAEM uses adversarial learning to create fake matches as extra training examples. As for Ditto, data augmentation is used to create more training examples and the augmentation operators can distort the training examples so comprehensively that the labels become incorrect, necessitating the use of interpolation to rein in the changes.

Compared against DAEM and Ditto, our approach forgoes text cleaning and data augmentation. Our approach’s text preprocessing only rearranges the text. In spite of that, our approach still manages to achieve higher F1 scores for many of the benchmark datasets.

The only requirement of our framework is that the two datasets of an EM problem have to have identical (or substitutable) attributes/columns. This is to facilitate attribute alignment where comparable columns from the two datasets are grouped together (see Section 2.3.3). All the benchmark dataset-pairs featured in Ditto, DAEM, HierGAT, and EMTransformer meet this criteria as the datasets in each pair have an identical number of attributes that mirror each other.

HierGAT is similar to our approach based on the absence of text cleaning and the use of additional training data in their approach. Another similarity between HierGAT and our approach is their use of multiple [CLS] tokens. Significant differences remain. HierGAT used multiple CLS tokens because each attribute embedding is represented as a CLS token and the concatenated attribute embeddings go on to form entity embeddings. Our use of multiple CLS tokens comes from text preprocessing (Section 2.3.3) and also from building ensembles (Section 2.3.5), with only the Aligned text preprocessing scheme bearing some resemblance to HierGAT’s attribute embeddings. Another difference is the core novel feature underlying each framework. In HierGAT, this is the hierarchical heterogeneous graph, which needs to be constructed and trained to obtain attribute and entity embeddings. Our approach invested in building intra-model (i.e., using only a single transformer architecture) ensembles.

2.4.5 Ablation study

Here, we investigate the influence each of our enhancements on the base transformer architecture exert on EM classification results as well as runtimes. There are too many possible configurations for ensembles (e.g., Summed and Simsents can be used with single-CLS instead of dual-CLS, default with single-CLS can be ensembled together with Summed with dual-CLS etc.) so we only presented a small informative selection of configurations. All ensembles use ConvBERT and are run for 10 epochs and the F1 scores are means from five runs. The training runtimes are means from the five runs as well. The text preprocessing runtimes are averages from the five-run averages of all ensembles.

The baseline is “default, sans attention”, which uses the transformer as originally designed with single-CLS and is comparable to the EMTransformer approach. *AttendEM* is represented by “Aligned + SimSents + Summed”. To investigate the impact of the additional self-attention layer in the classification block on ensembles that contain multiple models, we also present results for *AttendEM* with the attention in the classifier head removed.

Runtimes, which can be found in Table 2.3, appear fairly stable across different datasets. Each additional model in the ensembles increases the training runtime almost linearly by $0.7\times$, i.e., two-model ensembles takes $1.7\times$ the default’s time to run while three-model is roughly $1.7\times + 0.7\times = 2.4\times$ the default’s. During evaluation/inference (no backpropagation), the burden imposed by additional models in an ensemble is greater than during training, with each extra model adding $1.0\times$ to the runtime, thus making two-model ensembles’ runtimes around $2.0\times$ the default’s and three-model ensembles at about $3.0\times$ the default’s.

From the F1 scores obtained by the ensembles/models in Table 2.4, we can see that there is no single text preprocessing method that works well across different datasets when using additional aggregator tokens and self-attention during classification. For instance, Aligned (97.06) performs better than the default (95.42) in Dirty iTunes-Amazon but falters when facing against the Dirty Walmart-Amazon dataset (83.27, worse than the default’s 85.48). Meanwhile, the reverse applies to Summed, where it outperforms the default on Dirty Walmart-Amazon (85.52 vs 85.48) but not on Dirty iTunes-Amazon (95.06 vs 95.42).

Ensembles with two models feature noticeable variability across datasets as well. The “Aligned + Summed” ensemble performed better than the default on Dirty iTunes-Amazon (95.74 vs 95.42) but is worse when it comes to Dirty Walmart-Amazon (85.05 vs 85.48).

The performance of two-model ensembles is not necessarily the average of their constituents. For Dirty iTunes-Amazon, the SimSents (94.37) and the Summed (95.06) models individually are worse than the default (95.42) but perform almost as good as the default

in the “SimSents + Summed” ensemble (95.41). A similar situation can be found in Dirty Walmart-Amazon when comparing the individual Aligned (83.27) and SimSents (84.19) models against the ensemble (85.79) and default (85.48).

The full three-model ensemble we used to represent *AttendEM* seems to avoid wild swings in F1 scores when moving between datasets. While the full *AttendEM* may not necessarily be the best performing ensemble (e.g., the individual Aligned model with F1 of 97.06 outscores the full *AttendEM*’s 96.45 in Dirty iTunes-Amazon), it manages to maintain its performance lead over the default across different datasets unlike the smaller ensembles.

Self-attention in the classifier head appears necessary for models with multiple aggregator CLS tokens and for large ensembles featuring such models. For both Dirty iTunes-Amazon and Dirty Walmart-Amazon, the use of dual-CLS tokens resulted in worse performance (94.75 Dirty IA, 85.39 Dirty WA) than the default sans self-attention model (95.10, 85.85); the introduction of self-attention in the classifier helped regain the lost performance (96.10, 85.61). In the case of ensembles, without self-attention, *AttendEM* (96.45) would have scored lower (94.76) than the default Dirty iTunes-Amazon (95.42). When there is only a single model, such as the default, self-attention can potentially degrade performance, as can be seen in the case of Dirty iTunes-Amazon (95.10 vs 95.42).

2.5 Conclusion and Future Work

In this paper, we introduced *AttendEM*, a transformer-based framework for entity matching. Our evaluation showed that ensembled models built under this framework achieve state-of-the-art results in a majority of EM benchmark datasets. For the datasets where *AttendEM* performed better, the observed improvement in F1 ranges from 0.13 to 2.09 (the Beer and Company datasets being the two exceptions). When compared against Ditto’s own results over all 13 datasets, i.e., including datasets where *AttendEM* performed worse, the average improvement was +0.16. If we use DAEM’s reported results for Ditto, the average is +3.31 for all seven evaluated datasets. And if we use HierGAT’s reported results for Ditto, the average is +2.60 for all 13 evaluated datasets. When compared against DAEM, the range of improvement is from 0.22 to 4.57 (the DBLP-GoogleScholar, iTunes-Amazon, and Walmart-Amazon datasets being the three exceptions) with a mean of +0.30 over all seven datasets that DAEM was tested against. And when compared against HierGAT, the range is from 0.02 to 4.30 (the five exceptions being Beer, DBLP-GoogleScholar, Walmart-Amazon, Dirty DBLP-ACM, and Dirty Walmart-Amazon) with a mean of +0.48 over all 13 datasets that HierGAT was tested against.

Although the improvement may seem slight, *AttendEM*’s improvements are comparable

to two other solutions that claimed to have outperformed Ditto, HierGAT [515] and DAEM [223]. The mean of their percentage improvements over Ditto, calculated using their respective replications of Ditto results, were 2.46% for HierGAT (compared to *AttendEM*'s 2.99%) and 3.42% for DAEM (compared to *AttendEM*'s 3.93%).

AttendEM demonstrates that ensembles combined with additional aggregator tokens and self-attention can improve the performance of a transformer in the domain of EM, albeit at the cost of greater runtimes as the transformer needs to cycle through each model's preprocessed text sequentially. To alleviate the issue with runtime, future work can explore the use of processing units, e.g., GPUs and TPUs, with sufficient memory capable of fitting multiple models at once so that running the ensembles can be parallelized.

Pre-training transformers used in *AttendEM* using text that more closely matches the problem domain of the downstream classification task (e.g., pre-training on scholarly publication dataset prior to classification fine-tuning for DBLP-ACM EM task) is a potential avenue for improving EM performance, as the transformers used in our work and in previous works were pre-trained on "generic" text corpora. There are known examples of transformers benefiting from unsupervised pre-training on datasets that bears greater resemblance to the datasets found in downstream tasks. SciBERT outperformed BERT on NLP tasks in scientific domains such as biomedicine and computer science simply through pre-training on a large multi-domain corpus of scientific texts as opposed to the Wikipedia and literary books corpora used by BERT [53]. Filling in missing values and utilizing domain knowledge to normalize text spans and identify attribute types may improve *AttendEM* performance as well. After all, both DAEM and Ditto credited their careful engineering of text cleaning solutions for their frameworks' improved performance over existing solutions.

Despite the great variety of unique sources in the datasets found within the ER-Magellan benchmark, the two domains of commerce (product listings) and bibliography make up the entirety of the benchmark. While *AttendEM* has proven itself against the ER-Magellan benchmark, the generalizability of *AttendEM* to other, more specialized, problems such as the merging/de-duplicating of entries from different crime databases [7] should be studied.

Some of the modifications introduced by *AttendEM* is not specific to the task of EM, namely the use of ensembles and the addition of an attention block in the classification head. Extending the use of these modifications to problems outside of EM warrants investigation.

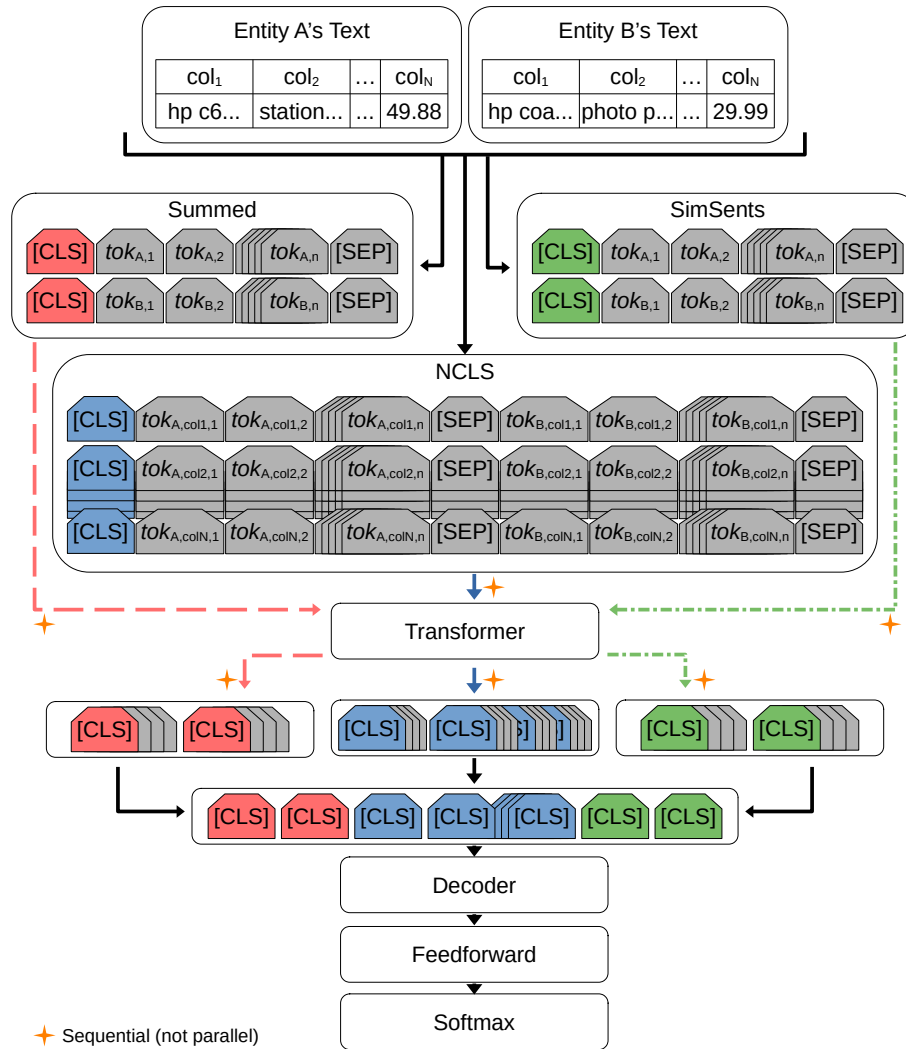


Figure 2.2: *AttendEM*'s workflow showing text preprocessing (NCLS/Attribute Alignment, Summed, SimSents, detailed in Section 2.3.3), aggregator token insertion (dual-CLS, NCLS, detailed in Section 2.3.4), ensembling the outputs from the different representations of the text span as described in Section 2.3.5, and the classifier head containing a self-attention layer (decoder) along with feedforward layers processing the concatenated [CLS] embeddings as described in Sections 2.2.5.2 and 2.3.5. Transformer stands in for a substitutable architecture, e.g., DeBERTa or ConvBERT. As stated in the bottom left corner of the figure, a four-pointed star denotes that the process is sequential as opposed to parallel.

Table 2.3: Runtimes for different ensembles and text preprocessing methods. Default uses single-CLS, SimSents and Summed use dual-CLS while Aligned uses NCLS. *AttendEM* is “Aligned + SimSents + Summed”.

Ensemble	Dirty IA	Dirty WA
Training (per epoch)		
Default	31s	9m52s
Default, w/o attention	0.969×	0.971×
Default, 2CLS	1.02×	0.998×
Default, 2CLS, w/o attention	0.969×	0.958×
Aligned	1.00×	1.04×
SimSents	1.00×	1.03×
Summed	1.00×	1.04×
Aligned + SimSents	1.64×	1.66×
Aligned + Summed	1.65×	1.65×
SimSents + Summed	1.59×	1.69×
<i>AttendEM</i>	2.34×	2.41×
<i>AttendEM</i> , w/o attention	2.57×	2.43×
Evaluation (per epoch)		
Default	3s	56s
Default, w/o attention	1.05×	0.987×
Default, 2CLS	1.03×	0.978×
Default, 2CLS, w/o attention	1.05×	0.984×
Aligned	1.02×	0.982×
SimSents	0.979×	0.977×
Summed	1.02×	0.991×
Aligned + SimSents	1.95×	1.93×
Aligned + Summed	1.96×	1.85×
SimSents + Summed	1.94×	1.95×
<i>AttendEM</i>	2.92×	2.98×
<i>AttendEM</i> , w/o attention	3.40×	2.96×
Text preprocessing		
Default	0.414s	0.162s
Aligned	0.875×	0.814×
SimSents	378×	416×
Summed	257×	261×

Table 2.4: F1 scores for different ensembles. Default uses single-CLS, SimSents and Summed use dual-CLS while Aligned uses NCLS. *AttendEM* is “Aligned + SimSents + Summed”.

Ensembles/models	Dirty IA	Dirty WA
Default	95.42	85.48
Default, w/o attention	95.10	85.85
Default, 2CLS	96.10	85.61
Default, 2CLS, w/o attention	94.75	85.39
Aligned	97.06	83.27
SimSents	94.37	84.19
Summed	95.06	85.52
Aligned + SimSents	96.08	85.79
Aligned + Summed	95.74	85.05
SimSents + Summed	95.41	85.13
<i>AttendEM</i>	96.45	86.29
<i>AttendEM</i> , w/o attention	94.76	86.23

Chapter 3

Distinguishing between Fake News and Satire with Transformers

Identifying entities that appear different but are in essence the same (Chapter 2) represents only one dimension of addressing saturation attacks. The obverse — differentiating between entities exhibiting high degrees of similarity but containing subtle but crucial differences — needs to be addressed as well, counterintuitive though it may seem at first glance. We shall illustrate the need for this capability with the case of highly similar fake news and satirical news being disseminated online.

Research has found that repetition of highly similar messages (i.e., not exact duplicates) can cause a positive attitudinal change from the audience towards the core message [323]. And, a meta-analysis of studies on the truth effect showed that simple repetition of a message can also make that message more credible [131].

Satirical news present a saturation attack threat on both dimensions. Satire is often highly similar yet distinct from fake news — sufficiently distinct that classifiers can separate the two, as we will demonstrate in this chapter. But, satire can also be literal duplicates of fake news because there have been documented instances of fake news site operators copying articles from satirical news sites [36]. There are also known instances of rumors slowly losing hyperlinks attributing the rumors to satirical sites through repeated re-shares despite being initially shared with hyperlinks [436], essentially depriving readers the ability to easily identify the content as satire and potentially leading readers to treat them as credible news. Satirical news fake news can therefore inadvertently boost the credibility of fake news simply by co-existing alongside fake news.

Despite being a threat, removing satirical news may not be the optimal solution. In the previous chapter (Chapter 2), the saturation attack involved information that were duplicated or counterfeited, which are unambiguously faulty and can be excised without

ethical concerns. Satirical news occupy a more tenuous position. Even though some researchers choose to label satire as disinformation, satire has often been championed as part of the fifth estate keeping powers like the fourth estate, i.e., mainstream news, in check.

By improving our capabilities to distinguish between fake news and satire without obvious markers, e.g., hyperlinks to a satirical news site, decision makers are given the opportunity establish separate policies for managing fake news and satire. A less restrictive policy towards satire compared to fake news can not only preserve healthy political engagement in a society, but it can also garner goodwill. Reputation management is an important aspect of information warfare after all [64, 375].

3.1 Introduction

Fake news spreads quickly online, more so than verified news [483]. The uncontrolled spread of fake news can have damaging consequences, as in the case of telecommunication towers being burned due to a conspiracy theory linking COVID-19 with 5G technology [9]. A growing awareness of the danger posed by fake news has been accompanied by calls to keep its spread in check [510]. However, when combating fake news, satirical news risks being lumped in with fake news as they exhibit highly similar textual cues, especially when contrasted against real news [220]. So great is their similarity that researchers have had success in classifying fake and legitimate news using a classifier trained on a dataset where fake news has been substituted with satirical news [233].

We must avoid grouping satirical news together with fake news as they serve very different purposes. *Fake news* is “news articles that are intentionally and verifiably false, and could mislead readers” [20]. *Satirical news* is “factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise ‘bad’ ” [187].

Satire serves as a part of the Fifth Estate along with news parody and non-mainstream media sources like columnists and bloggers in critiquing both the people in power and the mainstream news media. This form of criticism wrapped in humor indirectly educates people on political issues and can motivate their participation in the political process, which is crucial in the functioning of a healthy civil society.

Developing the capability to distinguish between fake and satirical news can aid social media companies in crafting more nuanced policies on managing fake news. Such policies are not only helpful in preserving the Fifth Estate but are also more likely to be amenable to government officials concerned over corporate overreach, e.g., European leaders’ dis-

pleasure over Twitter’s banning of Trump [231]. To contribute to tackling the problem of separating fakes and satires, we present a transformer-based classifier framework. Comprehensive benchmarking on a meticulously constructed dataset of real-world fake and satirical news [187] demonstrates the viability of our approach and provides information on optimal model configuration. Further benchmarking on two more datasets demonstrates that the performance advantage held by our approach is generalizable to different datasets and data availability conditions.

The contribution we present in this paper can be summarized as follows:

- Development of a novel framework for classifying satire and fake news that achieves state-of-the-art classification results. Our framework is based on DistilBERT [420], which also makes it the first ever application of a transformer architecture of any kind to the problem of distinguishing between fake and satirical news.
- Pre-training transformer. Differing from many previous works that focus on the fine-tuning stage of transformers, we investigate how different pre-training choices, in terms of the sample sizes of the pre-training process and the dataset used — target dataset, target dataset with sorted sentences, and in-domain external dataset — can ultimately impact the classifier performance.
- Information aggregation method. We examined how two non-standard tokenization schemes using multiple aggregator tokens, one based on article section segmentation and the other based on sentence segmentation, fare against the use of a single aggregator token that is standard to vanilla DistilBERT, vanilla BERT, and BERT variants.
- Text truncation strategy. A text truncation strategy where sentences are sorted according to their informativeness is compared against the naive baseline of using the text as is to address the issue of the entire news article text not fitting within transformers’ token limits.
- Extensive empirical evaluation. We conducted extensive testing of our framework on a carefully constructed dataset consisting of real-world fake news articles from a balanced and varied number of sources. The influence that each framework component has on final classification quality is investigated.
- Generalizability evaluation. Additional testing of our framework on two other datasets featuring fake and satirical news was conducted to observe if our approach can be applied to new datasets without degrading its state-of-the-art performance, especially

when encountering non-ideal conditions, such as a dataset with more than two classes and a highly imbalanced class distribution.

3.2 Preliminaries and related work

3.2.1 Defining satire and fake news

As defined earlier in the introduction, this paper considers *fake news* to be “news articles that are intentionally and verifiably false, and could mislead readers” [20]. As for *satirical news*, it is news that is “factually incorrect, but the intent is not to deceive but rather to call out, ridicule, or expose behavior that is shameful, corrupt, or otherwise ‘bad’ ” [187].

Establishing what constitutes fake and satirical news is crucial as there is a lack of consensus on what each of these terms mean even among academic researchers [187, 460]. In this section, we will discuss the importance of distinguishing satire from fake news, the criteria that can be used to distinguish them, and the reasons for adopting these criteria.

Satire needs to be considered as distinct from fake news because satire is a form of social good. There are those who disagree. First Draft News, a non-profit addressing “challenges relating to trust and truth in the digital age” that counts Google News Initiative as its founding member [359] and has received funding from the Facebook Journalism Project and Twitter, considers satire a form of misinformation that does not intend to cause harm but has the potential to mislead [494]. First Draft News’s opposition to satire is countered by much existing evidence on the benefits brought about by satire. Beyond their entertainment value, researchers have found satirical news stories to serve critical functions that include raising awareness among its audience [51], building greater understanding and sharpening opinions on political issues among its audience [71], galvanizing its audience into political action [93], and telling truth to power [237, 349]. We believe these benefits outweigh whatever potential harm satirical news may do.

We are not alone in our desire to treat satirical news and fake news as two separate categories. For instance, Golbeck et al. [187] created a fake and satirical news dataset mostly because they wanted to determine if linguistic and thematic features are sufficient to differentiate satire and fake news. Tandoc et al. [460] personally disagree with satire being lumped into the category of fake news even though their literature survey found that there are many researchers who consider satire as fake news.

If one accepts our premise that satire should be considered as distinct from fake news, the question now turns to finding the criteria that will allow us to label articles as either satire or fake news. These criteria are not only for separating satire and fake news but also

for determining whether an article is fit for inclusion in either category. Two major criteria — factuality and intent — that have been used in existing attempts at providing a typology for fake news [460, 521] serve equally well as criteria for identifying satire and fake news.

Factuality is an important criterion in determining if a news article counts as fake news. In the absence of this criterion, fake news as a term risks being co-opted by “some circles as an attack on legitimate, factually correct stories when people in power simply dislike what they have to say” [187]. Some researchers consider propaganda as a form of fake news and define propaganda simply as news stories created to influence public perceptions for political aims [460, 521], without a stringent requirement for the presence of factual inaccuracies within the news stories. Our requirement would exclude factually accurate propaganda from being considered as fake news. The requirement for the presence of factual inaccuracy extends to satirical news as well; factually accurate satire cannot exist as it would simply be legitimate news.

Intent is an important criterion for separating satire and fake news. Without intent as a criterion, satire and fake news would be indistinguishable from each other as both types of news share the commonality of being factually inaccurate. With intent as a criterion, satire stands apart from fake news because satire is an attempt to mock bad behavior while fake news intends to mislead its readers. For some researchers such as Zannettou et al. [521], satirical news articles being oft-disseminated on social media to audiences unaware of their provenance is sufficient basis to condemn them to the category of fake news. But since we consider intent when categorizing news in our paper, satirical news does not count as fake news as people mistaking it for real news is an *unintended* consequence. Similarly, legitimate articles with factual inaccuracies that appear to be products of honest mistakes or unreliable information sources are not considered fake news.

3.2.2 Automatically classifying satire and fake news

Distinguishing between satire and fake news can be difficult for algorithms as fake news bears a lot of textual cues that make it resemble satirical news more than real news [220, 411], and often satirical news is reprinted in trustworthy outlets as if the reporting were true [411]. Researchers have even found success in using satirical news to stand in for fake news when training a classifier for legitimate and fake news [233].

Much research on automatic classification has been done in areas that are adjacent but not directly relevant to ours. For instance, there have been numerous attempts to identify fake news, such as [179, 396] and as evinced by various literature reviews, e.g., [482, 521]. There is research on detecting news occupying the gray area between real and

fake — news of unknown veracity, i.e. rumors [18]. There is also some research interest in telling satirical news apart from real news, e.g., [6, 79, 128, 411]. Work also has been done on classifying real, fake, and satirical news at the same time, e.g., [127, 472]. However, we are aware of only three pioneering works that address the more challenging task of distinguishing *purely* between satire and fake news: [187, 220, 482].

Volkova et al. [482] explored two problems, classifying verified and suspicious news and classifying different types of suspicious news. Their attempt at classifying suspicious news into four types of suspicious news categories — propaganda, hoaxes, satire, and clickbait — is the only portion of their work that is relevant to ours. Classifying different types of suspicious news is not strictly the same as separating fake news from satire; their pernicious inclusion of propaganda under the banner of fake news is a major methodological difference from our work of separating fake news and satire. To illustrate why we believe so, we will use the New York Times as an example. In Figure 1 of their paper, the New York Times (NYTimes) is classified as a “verified” news source. But according to Noam Chomsky, who along with Edward Herman put forward the well-cited propaganda model in *Manufacturing Consent* [210], the New York Times falls under the category of pure propaganda [102]. Another methodological difference with the dataset used in [482] is the highly imbalanced class distribution, with the number of propaganda posts being an order of magnitude greater than other categories of fake news. While Volkova et al. [482] stated that the source labels were manually verified for quality assurance, they did not state that they manually verified the labels for each and every news story. This is tantamount to judging the contents of a book (all news stories) by its cover (the source).

Horne and Adali [220] sought to identify which news stories are real, fake, or satire. Of the three datasets included in their study, only one dataset is relevant to our work, which is their political news data set. The dataset consists of real, fake, and satire categories and has a collection of 75 stories for each category. While the class distribution is balanced for this dataset, the sources are not, as the fake news stories are sourced from only eight outlets and, for satirical news stories, only six outlets. Like Volkova et al. [482], they also did not hand code the stories individually and instead labeled them according to the trustworthiness of their respective sources.

In the work by Golbeck et al. [187], fake news and satire are given definitions that align closely with the definitions used in this paper. Additionally, a well-considered set of data collection guidelines are used to not only clearly delineate the boundaries between fake and satirical news but also to ensure diversity and reliability in the collected data (Section 3.4.1 details those guidelines). As Golbeck et al. [187] are the only ones we are aware of thus far to have released such a rigorously vetted fake and satirical news dataset, later works

dealing with the automatic identification of fake and satirical news such as [121, 427] have relied solely on the dataset shared by Golbeck et al. [187] and did not use the dataset from either Horne and Adali [220] or Volkova et al. [482].

Based on the quality of the datasets and methodological concerns, we had intended to rely solely on the Golbeck et al. [187] dataset. After all, claims of generalizability made in two other papers [121, 427] rested entirely upon results obtained from just the single high quality Golbeck et al. [187] dataset. However, to avoid accusations that we are cherry-picking datasets and to assuage concerns over our approach basing its generalizability claims on evaluations from a single dataset, we reported our framework’s performance on Horne and Adali [220] and Volkova et al. [482] datasets as well, even though we are opposed to the inclusion of these datasets in our paper on methodological grounds.

3.2.3 Transformers

Transformers are a class of neural network architecture that has proven itself superior in many natural language processing (NLP) tasks, such as news rumor detection [19], question answering, and language inference, compared to recurrent neural networks as well as traditional machine learning [475]. The success of transformers is due to the attention mechanism being able to better model dependencies between tokens even when the tokens are far apart from each other compared to models that rely on recurrence. Some of the more successful and widely used transformer variants are BERT (Bidirectional Encoder Representations from Transformers) [137], OpenAI’s GPT (Generative Pre-trained Transformer) [395], and XLNet [514].

3.2.3.1 BERT and DistilBERT

Of the existing transformer architectures, we chose to build our framework around DistilBERT [420], a BERT derivative. BERT is a language model that incorporates only the encoder section of the original transformer model out of the decoder and encoder sections [475], allowing each token in BERT to attend to context from both directions instead of only from the left as in the original transformer. We chose DistilBERT because BERT is considered a baseline architecture and DistilBERT is simply a more efficient derivative of BERT.

BERT’s status as a baseline architecture can be seen by the many fake news detection studies that uses it unmodified (see Section 3.2.3.2). The large number of BERT-derived architectures, such as RoBERTa [291] and FlauBERT [271], is further evidence of BERT being a baseline architecture.

DistilBERT’s development was spurred by BERT’s computationally demanding nature as even BERT-Base, which requires significantly less memory than BERT-Large to run, can prove too taxing for many systems. Through the use of the compression technique known as knowledge distillation, DistilBERT’s makers, huggingface, claim that despite DistilBERT being 40% of BERT’s size and 60% faster, it retains up to 99% of BERT’s language understanding capabilities.

3.2.3.2 Identifying fake news with transformers

The classification of fake news using transformers have almost always involved leveraging vanilla BERT, e.g., [4, 238, 242, 288], instead of other transformer architectures, with few exceptions, e.g., [204]. Of those who used BERT, all of them are concerned with fine-tuning the transformer except for Jwa et al. [238] who did pay attention to further pre-training the transformer, albeit only with an *external* in-domain dataset and not with the *target* dataset itself (Section 3.4.1 of our paper discusses the distinction between *external* and *target* datasets).

The core of the BERT architecture is usually left un-tampered with. Where researchers tend to focus their creativity is on modifying the inputs received by BERT and finding new ways to process the outputs from BERT. The two-stage model from Liu et al. [288] feeds a coarse-grained label obtained from the first-stage BERT classifier as input to the second-stage BERT classifier. FakeBERT from Kaliyar et al. [242] uses layers of convolutional neural networks to pool the outputs from BERT instead of more conventional linear layers.

There are researchers who do not even attempt to make significant alterations to the inputs or outputs, which can be taken as a testament to the robustness of the BERT architecture. The BAKE model by Jwa et al. [238], in the authors’ own words, is “essentially a BERT model, [and they] call it BAKE because [they] first applied it to the task of fake news detection in [their] case”. The exBAKE variant does feature a significant change as it uses weights altered by additional pre-training on an external in-domain dataset. Aggarwal et al. [4] also used the BERT model as is, with the only notable change being their dataset cleaning step. This consists of the removal of outlier data, which are texts that are either too short or too long, and the removal of what they termed “noise”, consisting of characters such as punctuation marks, numbers, and newlines.

The way in which Hassan and Lee [204] relied on transformers was not to fine-tune one to a chosen target task but to obtain sentence-level embeddings without further fine-tuning. They used the Universal Sentence Encoder (USE), specifically the `universal-sentence-`

encoder-large model that is trained with a Transformer [475] encoder¹.

For the validation of their classifiers, each paper relied on just one dataset. Liu et al. [288] used the LIAR dataset, which is a collection of short statements from Politifact, Aggarwal et al. [4] used the NewFN dataset, Kaliyar et al. [242] used a Kaggle dataset of fake and real news covering the 2016 US presidential election, and Hassan and Lee [204], Jwa et al. [238] used the Fake News Challenge FNC-1 dataset. For pre-training, the External dataset that Jwa et al. [238] used was a corpus of CNN and Daily Mail articles originally used to test text summarization algorithms.

Jwa et al. [238] fed BERT with both the headline and body text of an article and no indication was given on how text sequences that exceeded BERT's token length limit were handled. The work by Liu et al. [288] dealt with short statements and their dataset features 17.9 tokens on average, which is well under the 768-token length limit of the BERT-Base variant they used. In the work by Aggarwal et al. [4] the text sequences in their dataset has a median length of 597 tokens. Aggarwal et al. [4] opted for a naive strategy of limiting sequence length to just 256 tokens, half that of the 512-token length limit of their particular variant of BERT-Base. Going well below the limit was done in the interest of reducing runtime and computing resource usage. Kaliyar et al. [242] also used naive truncation with the token length limit of 512 matching that of their BERT-Base model. The USE found in the classifier built by Hassan and Lee [204] does not have a hard length limit.

Vanilla BERT has also made appearances in multimodal fake news classifiers. Multimodal classification problems are where different modalities (text, images, etc.) are considered before arriving at a labeling decision. In the circulation of fake news, doctored or misleading images can often be found alongside untruthful text articles. Two teams, one composed of researchers affiliated with institutions from India and Japan [442] and another featuring members hailing from Spain and China [180] arrived at very similar approaches for making full use of available image and text data in detecting fake news. The graphical data found in a news article is handled by convolutional neural networks, with Singhal et al. [442] using VGG-19 and Giachanou et al. [180] using VGG-16. The original BERT architecture with BERT-Base weights is leveraged in both works to deal with the textual component of news articles. One of the contributions made by the later work of Giachanou et al. [180] is the use of text-image similarity as a classification feature. This similarity is computed between word embeddings of image tags extracted by VGG and the title text's word embeddings. Singhal et al. [442] used Twitter and Weibo data while Giachanou et al. [180] used the GossipCop subset of the FakeNewsNet collection. BERT is capable of accommodating lengthier texts, but both teams elected to limit the length of text sequences processed by BERT. In [442],

¹<https://aihub.cloud.google.com/p/products%2F42c1bfd4-8104-450c-a348-29b047d3691c>

BERT has access to up to 23 words in Twitter tweets and 200 characters of Weibo posts. In [180], BERT is fed the text sequences derived only from an article’s text and each sequence is length-limited to 64 tokens.

Although the vast improvements brought about by transformers has led to their widespread application across a variety of domains, we are only aware of transformers being used for the detection of fake news among real news, and we have not seen transformers applied in any capacity to the specific task of telling apart fake news and satire.

3.3 Identifying satire and fake news with transformers

This section presents the three major components of our proposed fake and satirical news classifier framework: the information aggregator token design, transformer pre-training choice, and the text preprocessing method for both the pre-training and fine-tuning stages. Each unique combination will be referred to as a model. DistilBERT coupled with pre-trained `distilbert-base-uncased` weights serves as the starting point for all our experiments.

Figure 3.1 illustrates one model from our framework and shows how the components interact with each other.

3.3.1 Single-CLS, TB-CLS and multi-CLS for information aggregation

BERT and DistilBERT take sequences of tokenized texts as input. A sequence can be any arbitrary span of text, be it a single linguistic sentence or multiple sentences. Tokenization relies on the WordPiece Model (WPM) vocabulary, built on the basis of statistical NLP to address the issue of keeping fixed vocabulary sizes while still being able to process rare words [509]. In BERT and DistilBERT, additional tokens that perform special functions are inserted among the tokenized texts. One such special token is the classifier token, [CLS], used to aggregate information for an entire sequence of tokens, with that aggregated information typically being used for classification tasks as the name implies. The separator token, [SEP], demarcates the end of a sequence. CLS tokens are placed at the start of sequences and SEP at the end.

In the original formulation of BERT and DistilBERT, only one CLS token is used per sequence of tokens. Recently, researchers found that state-of-the-art results for abstractive and extractive text summarization can be achieved [290] when using multiple CLS tokens instead of one, in addition to other modifications to standard BERT architecture. Their results prompted us to investigate if the use of multiple CLS tokens would also help in our fake news and satire classification task. We experimented with two new schemes: (1)

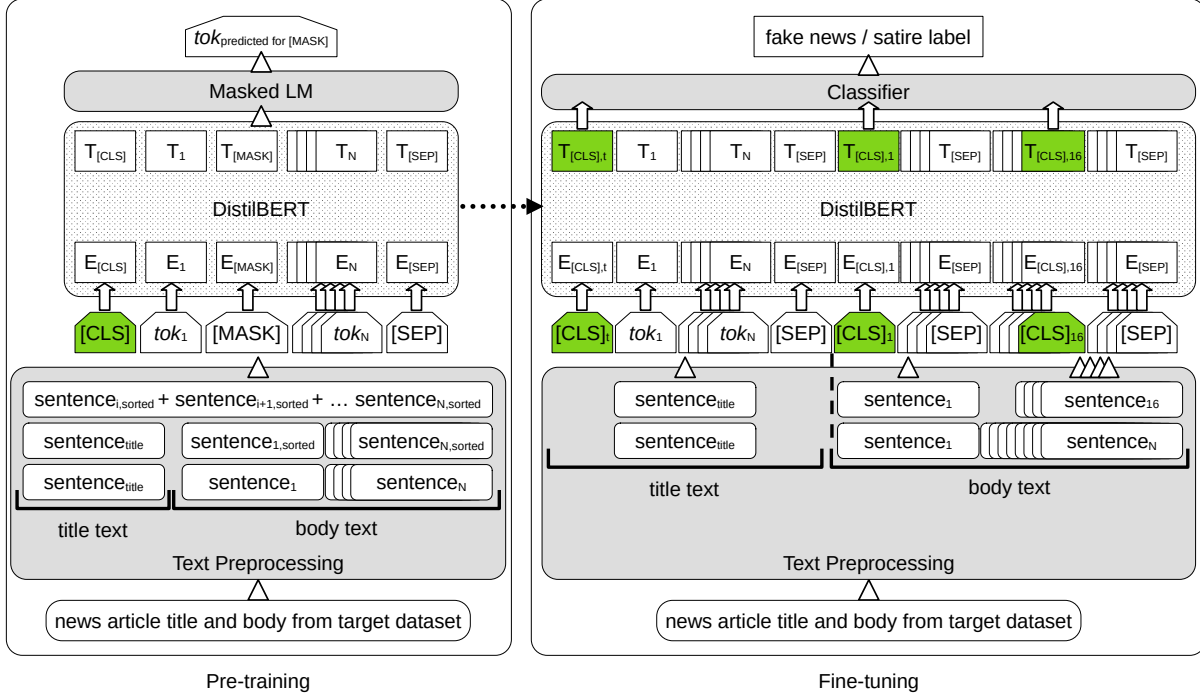


Figure 3.1: Fake news and satire classification framework. The particular configuration presented in this figure, MultiCLS-SimSents-Title+Body, is pre-trained using the masked language modeling task with the same dataset as the dataset used for the target/downstream fake news and satire classification task. During pre-training, sentences from an article’s body are first re-ordered by similarity to its title (SimSents). Then, a random sentence i is picked as the starting point, concatenated with sentences following it, and tokenized. During fine-tuning, DistilBERT parameters are initialized using values obtained at the end of pre-training. 17 special aggregator tokens, [CLS], are used during fine-tuning (MultiCLS), with 1 for the article title and the remaining 16 for the first 16 sentences from the article’s body with their original order preserved (Title+Body).

TB-CLS, short for Title-Body-CLS, where the article title is represented by one CLS token, and the entire body text is represented by another CLS token; and (2) multi-CLS, where every linguistic sentence gets one CLS token that serves to represent information from that sentence only. In essence, TB-CLS uses article section segmentation while multi-CLS uses sentence segmentation when placing CLS tokens. The differences in the tokenization schemes can be observed in Table 3.1.

We limited ourselves to having only a maximum of 16 sentences and 1 title sentence for a total of 17. In cases where we have less than 17 sentences and consequently less than 17 CLS tokens, the values for the missing CLS tokens are substituted with zeros.

While transformers output contextual vectors for all tokens, we used only CLS token vectors during classification. Accordingly, the number of neurons for each of the two hidden feedforward neural layers that together output the fake/satire class probabilities

Scheme	Tokens
SingleCLS	[CLS] animal noises . cats me ##ow . dogs bark . [SEP] [PAD] [PAD] [PAD] [PAD]
TBCLS	[CLS] animal noises . [SEP] [CLS] cats me ##ow . dogs bark . [SEP] [PAD] [PAD]
MultiCLS	[CLS] animal noises . [SEP] [CLS] cats me ##ow . [SEP] [CLS] dogs bark . [SEP]

Table 3.1: Tokenization under single-CLS, TB-CLS, and multi-CLS schemes given an article titled “Animal noises.” with the body text “Cats meow. Dogs bark.”, assuming a 16-token limit.

corresponds to the number of CLS tokens multiplied by the dimensionality of the token embedding, with the single-CLS, TB-CLS, and multi-CLS models having 768, 1,536, and 13,056 units per layer, respectively. The final classification layer applies logarithmic Softmax to the outputs, while the preceding layer uses ReLU (rectified linear units).

3.3.2 Unsupervised pre-training

Although the default `distilbert-base-uncased` weights for DistilBERT has already undergone extensive unsupervised pre-training on massive general text datasets, namely the English Wikipedia and the Toronto Book Corpus, its performance can still be improved upon with further unsupervised pre-training on datasets that are closer to the problem domain of interest [452]. This can be accomplished by either using the same dataset as the task of interest (denoted as Target) or a separate dataset from the same domain as the task of interest (denoted as External). We selected a significantly larger within-domain external dataset (NELA-GT is 2GB, while the data by Golbeck et al. [187] used for our target task is only 1.44MB) to observe if the performance gain typically associated with using a larger dataset for pre-training [291] can overcome the fact that the pre-training dataset differs from the target task dataset, even if the external dataset is within-domain.

The original BERT uses MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) as pre-training objectives. NSP and MLM are unsupervised objectives, allowing a transformer’s language modeling capability to be improved with unlabeled text data, which is vastly more common than labeled data. In MLM, the transformer is tasked with predicting tokens that have been randomly masked. In NSP, the transformer is tasked with determining if a sentence B is actually the next sentence that follows a sentence A. As RoBERTa (a Robustly Optimized BERT Pre-training Approach) [291] demonstrated that downstream task performance is retained or even improved when excluding the NSP objective during pre-training in certain settings, we chose to pre-train with MLM as the only objective. DistilBERT developers themselves pre-trained with MLM as the only objective as well.

To generate the samples for pre-training, we follow the set of guidelines outlined in

RoBERTa. A random document is first selected. Then a random linguistic sentence is selected as the starting point within the document. This sentence and the sentences following it are tokenized, then truncated to the first 512 tokens (DistilBERT’s token length limit). 15% of the tokens in the sequence are randomly selected for possible replacement. 80% of these randomly selected tokens are replaced with the [MASK] token, 10% are replaced with random tokens in the vocabulary, and 10% are left unchanged. Only dynamic masking was used, i.e., the masking pattern changes every time we retrieve a sequence. Unlike RoBERTa, if there is space for additional tokens (e.g., when a very short sentence at the end of a document was selected), we filled the additional spaces with padding tokens, [PAD]; in RoBERTa, a special separator token is inserted at the end of the first document, and the additional space is filled by another randomly masked sequence of sentence(s) from another randomly selected document. RoBERTa performed slightly better when sequences were restricted to come from a single document instead of multiple documents, but its developers ultimately chose to use sentences from multiple documents to avoid variable batch sizes. There are no explicit guidelines for handling documents with distinct title and body sections in RoBERTa, so we simply treated the title and body texts for each document as one contiguous sequence of text.

In addition to using the datasets as is, we also investigated the effect of training on text with sentences sorted by informativeness, reasoning that models that rely on texts sorted using the same method during the fine-tuning stage would benefit from it. Specifically, the sentences in each article’s body text are sorted in an ascending order based on TF-IDF (term frequency-inverse document frequency) cosine distance from the article’s title (see Section 3.3.3 for details). Since the fine-tuning stage is performed only on the target dataset, we do not believe that pre-training with sorted sentences for the external dataset would be beneficial. Therefore, we tested the effect of text sorting during pre-training only on the target dataset and excluded the external dataset. Weights obtained via pre-training on articles with sorted sentences are referred to as *SimSents*.

Pre-training is usually a time-consuming process. In the interest of making our transformer-based approach more attractive than existing options, we also investigated whether light pre-training with a small number of samples could give the same benefits as more extensive pre-training using a large number of samples. We ended up with six different sets of pre-trained weights for our comparisons (Table 3.2).

Although there have been BERT-based architectures that attempted to adapt to a domain by pre-training from scratch, such as SciBERT (pre-trained on scientific papers [53]), the pre-training in this work was not begun from scratch with randomly initialized weights but was instead started from *distilbert-base-uncased* weights. This decision was made out

Name	Dataset	# steps	Batch size	Total samples
SimSents(7.5M)	Golbeck et al. [187] with TF-IDF sorted body text	750K	10	7.5M
SimSents(30K)	Golbeck et al. [187] with TF-IDF sorted body text	15K	2	30K
Target(7.5M)	Golbeck et al. [187]	750K	10	7.5M
Target(30K)	Golbeck et al. [187]	15K	2	30K
External(7.5M)	NELA-GT	750K	10	7.5M
External(30K)	NELA-GT	15K	2	30K

Table 3.2: Pre-trained weights.

of concern for DistilBERT’s language understanding capabilities, which might not be fully developed if pre-training from scratch using an extremely small dataset (1.44MB Golbeck et al. [187] dataset), especially when the total number of training samples is low. To put the sample size into perspective, our most extensively trained model at 7.5M samples is only 0.9% of the lightest RoBERTa model, which was pre-trained over 800M samples (100K steps with a batch size of 8K). DistilBERT’s developers mentioned they used “up to 4K examples per batch” but did not mention the number of steps.

3.3.3 Text preprocessing

During pre-training and fine-tuning we experimented with two text preprocessing strategies. The naive strategy uses the provided text as is and cuts off tokens that exceed DistilBERT’s 512-token limit, even though we might potentially be excluding text that could help us discriminate between fake and satirical news. The other strategy we looked into is to sort sentences in an article’s body text based on their informativeness, reasoning that there are sentences that we can safely exclude when they exceed the token limit, as they are filled with extraneous details that will not be conducive in helping us determine whether an article is fake or satire. This second strategy is essentially text summarization.

Rather than immediately going for resource-intensive state-of-the-art unsupervised text summarization methods, we opted to start with a very simple baseline method. That method is to pick sentences from the body text that most closely resemble the title of the news article because the words in a title often capture the gist of a story. Our implementation involves first using the pre-trained Punkt English sentence tokenizer from the NLTK Python package to get a list of sentences for each article, then creating a TF-IDF values table for all the words of all the sentences found in an article, and finally sorting the sentences so that those with the smallest TF-IDF cosine distances from the title are at the top. During the pre-training phase, the above method was used to sort each article’s body’s sentences in the

target dataset so that the dataset could be used for MLM pre-training to obtain the SimSent set of weights.

The tokenization of a text sequence is achieved through splitting by punctuation and WordPiece [509] prior to truncating them to fit DistilBERT’s 512-token limit. If a model calls for sentence sorting, then sentence tokenization occurs before this word tokenization step.

In the fine-tuning stage for multi-CLS models, regardless of whether TF-IDF sorted or unsorted sentences are used, a limit N must be imposed on the number of body text sentences selected for consideration, as the classifier cannot account for a variable number of CLS token vectors. We let $N = 16$ for all multi-CLS models. Both TB-CLS and single-CLS models do not have a limit on the number of sentences but are limited by the space remaining on the per-sequence 512-token limit after accounting for the title text tokens and the special CLS and SEP tokens. As TB-CLS and single-CLS models do not need a CLS token inserted before every sentence, the sentences are concatenated together with one whitespace acting as a separator between them prior to word tokenization.

In multi-CLS models, each sentence (including the title) is limited to a maximum of 62 tokens, not including the special CLS and SEP tokens. In the hypothetical case where we did not use the title text and have 16 sentences with ≥ 62 tokens prior to prepending and appending the 2 special tokens, the transformer would only be able to process the first 8 sentences because $8 \times 64 = 512$. No per-sentence length limits are imposed on single-CLS and TB-CLS models.

When the body of a news article is used unaltered, we denote it as `Body` in the model name. When sentences from the body bearing the highest TF-IDF cosine similarities with the title are used before other sentences, we denote it as `SimSents`. If the title text is used, it will be indicated by `Title`. In single-CLS models, `Title` is always prepended to either `Body` or `SimSents` using a period combined with a whitespace as the separator. For all models, `Title` is never used by itself, as our initial exploratory tests showed that the `Title` alone does not give satisfactory results.

Parameters

For the Python packages `random`, `numpy`, and `pytorch`, we used a random seed of 42. The DistilBERT implementation we used is from the `transformers` package released by `huggingface`. Source code for our classifier framework has been made available publicly².

For pre-training and fine-tuning, AdamW is used as the optimizer with a learning rate of $2e-5$ and a weight decay of $1e-2$ for the core DistilBERT model. The same values for learning rate and weight decay are used for MLM during pre-training and for the classifier

²<https://github.com/jwenfai/fakenews-clean>

Dataset	Class	Text	# Characters				# Words				# Sentences				# DistilBERT Tokens			
			Min	Max	Med	90%	Min	Max	Med	90%	Min	Max	Med	90%	Min	Max	Med	90%
Golbeck et al. [187]	Fake	Title	14	182	73	99	2	34	13	19	1	3	1	1	3	46	15	21
		Body	86	25439	1897	5663	19	4155	350	1074	1	191	13	35	19	5304	392	1174
	Satire	Title	31	147	69	97	5	27	11	19	1	2	1	1	5	33	13	21
		Body	404	30490	1722	3423	80	6023	336	713	2	297	12	30	91	6678	366	747
Horne and Adali [220]	Fake	Title	46	146	75	94	6	27	14	18	1	3	1	2	6	31	16	20
		Body	590	10052	2241	5934	114	1985	432	1134	4	61	17	37	122	2192	474	1228
	Satire	Title	23	112	61	86	4	20	11	14	1	2	1	1	4	20	12	17
		Body	190	8368	1355	3637	37	1653	270	728	1	127	7	31	39	1755	298	779
Volkova et al. [482]	Propaganda	Body	7	152	122	140	2	48	21	28	1	12	1	2	4	120	39	57
	Clickbait	Body	17	148	120	140	3	33	21	28	1	5	1	2	6	103	41	53
	Satire	Body	9	148	117	140	2	46	20	27	1	5	1	2	4	63	37	50
	Hoax	Body	19	148	118	140	3	39	21	28	1	6	1	3	6	71	36	50

Table 3.3: Summary statistics for text features found in target and generalizability datasets. Med stands for median and 90% stands for the 90th-percentile.

layers during fine-tuning. These values are within the generally recommended range for pre-training and fine-tuning. The number of steps and batch sizes for pre-training are detailed earlier in Section 3.3.2. During fine-tuning, the models are trained for 10 epochs with a batch size of 8 for each cross validation split.

3.4 Evaluation and discussion

3.4.1 Datasets

We used four datasets in our experiments. The first dataset is used for pre-training and for evaluating the performance of our classifier. This dataset is known as the *target* dataset. The second dataset is used exclusively for pre-training only. This dataset is known as the *external* dataset. The third and fourth datasets are used to allay concerns over our classifier’s generalizability. Table 3.3 contains the descriptive statistics, separated by class and text feature, for each of the three datasets that were used for evaluations, i.e. the target and generalizability datasets.

Fake and Satirical News (Target)³. The dataset provided by Golbeck et al. [187] is used to benchmark our DistilBERT-based classifier framework on the target/downstream task — distinguishing between fake news and satire — hence its designation as the *target* dataset. This dataset is also used for pre-training DistilBERT. It contains a total of 283 fake and 203 satirical news articles. One sample from each category from the dataset can be found in Table 3.4.

³<https://github.com/jgolbeck/fakenews.git>

Fake News		
Title	URL	Body
Trump Has Fired Mus- lim Sharia Judge Ar- rested And Charged	https://worldpoliticsnow.com/2017/08/bringing-trump-fired-muslim-sharia-judge-arrested-charged/	As it turns out, in Donald Trump’s America, you can’t just go declaring religious laws from pagan countries that hate us to be legal if you sit on the federal judiciary. 22nd Circuit Court of Appeals Judge Hassan al-Hallamallala-Smith — an Obama nominee in 2009 — was taken into federal custody by the FBI for breaking the oath he took to uphold the Constitution. Under title 18 US Code Subsection 1209.3 ...
Satirical News		
Title	URL	Body
President Trump Fires All 14 Mus- lim Federal Judges	http://asamericanasap.plepie.org/2017/08/16/of-course-he-cant-do-that/	Barack Obama managed to pack the federal judiciary with bleeding heart liberals, professional lobbyists for climate change and other science-fiction scams and 15 Moslems, none of whom were born in this country. President Trump had no choice but to fire one of them last month when he ruled that Sharia Law was OK if the defendant is Moslem and reasonably expected to follow his faith or face eternity in hell. You can force a Christian to pay for birth control ...

Table 3.4: Examples from the “Fake News vs Satire” dataset released by Golbeck et al. [187]

A carefully thought-out set of guidelines was used in building this dataset. News defined as fake must not include opinion pieces, satire, and legitimate news with factual inaccuracy. To ensure articles are similar so as to diminish the chances of an article’s topic influencing its fake/satire classification, (1) the stories are all exclusively centered around American politics and (2) the articles have to be posted after January 2016 (this resulted in dates ranging from January 2016 to October 2017). There must be no borderline cases; fake news stories are incontrovertibly “factually incorrect and deceptive”, and satirical stories are “obviously satirical”. Another requirement is that only five articles/samples at most from each website/source are included in the dataset — a balanced number of samples reduces the risk of training a classifier that overfits on some specific writing style/tone/markers/indicators that might be commonly found within all articles from a particular source. The authors hand coded each individual story instead of relying on a proxy indicator like the reputations of source websites. Crucially, the dataset authors recognize that there are fake news websites using disclaimers of being satirical as a way to deflect from accusations of publishing fake news, so the authors have ruled out grey area cases from their dataset. For fake news stories, the dataset authors have also included a link rebutting the claims made by the stories.

There are three features for each news article: title, URL, and body. We did not use the

URL because we aim to distinguish satire from fake news purely through processing natural language. Many of the URLs in the dataset also no longer function.

NELA-GT (External)⁴. The NELA-GT dataset, provided by Nørregaard et al. [364], is *never used for the evaluation* of our classifier framework’s effectiveness and is used solely to pre-train DistilBERT, hence its designation as the *external* dataset. The motivating factor in our use of this dataset is the desire to observe if pre-training with a dataset from the same domain as the dataset used in the downstream task (news stories) can improve performance and if a larger and more diverse dataset is better for pre-training.

NELA-GT is a corpus of 713,534 English news articles pulled from a wide variety of popular and less well-known news sources (totaling 194) such as Al Jazeera, Infowars, Shadow Proof, Spiegel, and The Michelle Malkin Blog. As the dataset was built for the study of misinformation, the veracity of the sources is as diverse as the sources themselves (e.g., the dataset included articles from satirical news sites such as The Beaverton and The Onion). The dates of the articles in NELA-GT range from February 1st, 2018, to November 30, 2018, which do not overlap with the target dataset’s range of dates.

Horne and Adali [220] (Generalizability dataset 1)⁵. This dataset contains three categories of political news — real, fake, and satire — with 75 items for each category. As our work is concerned only with distinguishing between fakes and satires, we only used the 75 fake and 75 satirical articles. Each article is composed of two sections of text, title and body. The categorization of news articles in this dataset is wholly dependent upon the articles’ respective sources. An article from a source deemed to be satirical is considered satirical irrespective of its content. The same applies to fake news.

There are a total of eight sources used for fake articles and they are derived from a list of fake and misleading news websites from Zimdars [532], an assistant professor of communication and media at Merrimack College at the time the list was first published [533]. To be included, the fake news sources must have had at least one story proven as false on a fact-checking website. There are six sources for satirical news and the criterion for inclusion is that they must explicitly state their satirical nature on their website’s front page.

Volkova et al. [482] (Generalizability dataset 2)⁶. This dataset consists of 30,894 propaganda, 836 clickbait, 2,083 satire, and 1,341 hoax tweets. The number of tweets we managed to retrieve from the “rehydration” process ⁷ is significantly lower than the

⁴<https://doi.org/10.7910/DVN/ULHLCB>

⁵<https://github.com/rpitrust/fakenewsdata1>

⁶https://github.com/jwenfai/fakenews-clean/blob/master/volkova/volkova_rehydrated_id.csv

⁷In accordance to Twitter data sharing policy, many researchers only share tweet IDs publicly and other

number of tweets found in the original paper⁸ (56,271 propaganda, 1,366 clickbait, 3,156 satire, and 4,549 hoaxes) likely due to many tweets being deleted over the intervening time period. The class distribution of the dataset remains highly imbalanced in favor of propaganda tweets, although hoax tweets now constitutes a smaller portion of the dataset than in the original.

Unlike the other three datasets used in our work, the news items are tweets, meaning that they are severely limited in length compared to regular news articles due to the character limit imposed by Twitter. They do not have clearly delineated title and body sections. Many tweets are simply the titles and/or subtitles of complete articles followed by a hyperlink to said articles.

The categorization of the tweets are wholly dependent upon the reputation of the tweets' respective sources (the Twitter account where the tweet originated from and that account's corresponding website if it has one). The sources themselves are classified based on information from "several public resources that annotate suspicious Twitter accounts" and these resources were not identified by Volkova et al. [482] except for PropOrNot⁹ and Fake News Watch¹⁰. A total of 174 suspicious news accounts were identified, although we do not know how many of these accounts remain among the tweets we succeeded in rehydrating. All tweets were collected from the suspicious accounts one week before and after the Brussels bombing on March 22, 2016. Retweets from other users that mention the suspicious accounts were also collected and given the same label as the originals'.

3.4.2 Metrics and comparison basis

We adhere to the standard of reporting means of scores from 10-fold cross validation used by Golbeck et al. [187] as well as other publications [121, 427] that tested their classifier models against the Golbeck et al. [187] dataset. The stratified k-fold implementation from `scikit-learn` was used for this purpose.

We compare models from our framework against other approaches that are purely machine-based, as in methods that have no human involvement in determining if a news story is fake or satire. In our comparison against existing state-of-the-art approaches, we

researchers intending to access data associated with the IDs such as the text itself must reconstitute the tweet using Twitter APIs in a process known as rehydration, i.e., request the full Tweet, user, or Direct Message content. Twitter's content redistribution policy can be found at <https://developer.twitter.com/en/developer-terms/agreement-and-policy>.

⁸<http://www.cs.jhu.edu/~svitlana/TwitterList>

⁹<http://www.propornot.com/p/the-list.html>

¹⁰The original website at <http://www.fakenewswatch.com/> is inaccessible, US Library of Congress archival snapshot of the site on March 2017 can be found at <https://webarchive.loc.gov/all/20170301201107/http://www.fakenewswatch.com/>.

report F1 and accuracy scores. F1 scores were reported in the original paper by Golbeck et al. [187] and a subsequent paper by Das and Clark [121]. Accuracy is reported only because it was the sole metric reported in the Shabani and Sokhn [427] paper. Class imbalance biasing accuracy scores is not an issue with the Golbeck et al. [187] dataset as both classes have fairly equal numbers of instances (283 fake and 203 satirical articles).

For comparisons between different model configurations (pre-training, CLS type, text preprocessing) within our framework, we report the support¹¹-weighted F1.

3.4.3 Metrics and comparison basis for generalizability datasets

We reserved the taxing comprehensive evaluations of our satire versus fake news classifier to be performed on just the Golbeck et al. [187] *target* dataset; the testing of our framework against the Horne and Adali [220] and Volkova et al. [482] *generalizability* datasets are limited but rigorous. The tests are rigorous in that they employed the same test conditions and metrics as those used by the dataset creators and they are limited in that not every model configuration within our framework was tested, partly due to limitations imposed by the nature of the dataset themselves and the aforementioned test conditions.

While both the Volkova et al. [482] and Horne and Adali [220] papers have received large numbers of citations, our literature review failed to reveal any subsequent work that tested classifiers against the datasets found in these papers. Therefore, we report only the metrics used by the original authors of the datasets to enable comparison. For the Horne and Adali [220] dataset, the original metric is accuracy. For the Volkova et al. [482] dataset, the original metrics are micro-averaged F1 and macro-averaged F1. Horne and Adali [220] tested their classifier using 5-fold cross validation while Volkova et al. [482] used 10-fold cross validation; we tested on the same numbers of folds for the respective datasets.

The reported scores for these datasets should not be directly compared against those obtained from the Golbeck et al. [187] dataset due to different evaluation designs. The main difference is that our evaluations with the Golbeck et al. [187] dataset relied on using title and body texts together, which could not be done on either generalizability datasets for different reasons.

In their evaluations, Horne and Adali [220] did not use features generated from both the title and body texts simultaneously, instead opting to classify an article either solely with features from an article’s title or solely with features from an article’s body. A fair comparison required us to limit our classifier to using either title or body text. This excludes evaluating the TBCLS aggregator token design. We also excluded the SimSents

¹¹Number of true instances for each label.

text summarization technique from evaluations as it summarizes an article’s body text based on the article’s title. When using title text as the feature, sentence tokenization was performed on the title as otherwise the MultiCLS models would be nigh indistinguishable from SingleCLS models.

The Volkova et al. [482] dataset consists of tweets without distinct title and body sections. In the absence of separate title and body sections, SimSents and TBCLS were excluded from evaluations. Rehydration has managed to recover only about half of all the tweets used in the original evaluations performed by Volkova et al. [482] in addition to changing the class distribution slightly, so our scores are only good proxy indicators of how our framework would perform if the complete Twitter data were made available to us.

For the evaluations involving both generalizability datasets, we elected not to pre-train the classifier on the respective target texts as pre-training can be time-consuming, choosing instead to use the set of weights pre-trained extensively on the NELA-GT corpus of English news articles (External). This still allows us to evaluate if using weights that were further pre-trained to fit the news domain can boost performance over using weights that have received no further pre-training.

3.4.4 Results discussion

Models from our framework succeeded in attaining state-of-the-art performance. Overall, the ideal configuration for our framework is to first pre-train the transformer on millions of masked examples of sorted sentences from the texts of the target task and then using that set of weights when fine-tuning with multiple CLS aggregator tokens, one for each sentence, on unsorted sentences.

Performance comparison of existing approaches for classifying fake news and satire against the three best models built within our framework can be found in Table 3.5. A breakdown on how choices in pre-training sample sizes, pre-training dataset, aggregator token architecture, and text preprocessing affect the performance of a model built within our framework can be found in Table 3.6.

To fully reap the rewards of pre-training, it needs to be extensive (e.g., 7.5M samples) and not light (e.g., 30K samples). This is evinced by models using extensively pre-trained weights topping the performance table. Light pre-training not only results in performance that is at best mildly better than no pre-training but also in some cases can be even worse than no pre-training. The most illustrative example is the group of SingleCLS models using Title+Body text, which all performed worse when using lightly pre-trained weights than models with no pre-training, regardless of the pre-training dataset used. Many lightly

Model	F1	Acc.
TBCLS-SimSents(7.5M)-Title+Body	.8679	.8686
MultiCLS-Target(7.5M)-Title+Body	.8671	.8685
MultiCLS-SimSents(7.5M)-Title+Body	.8630	.8644
Das and Clark [121]	.825	-
Shabani and Sokhn [427] (machine-only)	-	.8164
Golbeck et al. [187]	.791	-

Table 3.5: Performance comparison of models from our framework against existing state-of-the-art on the Golbeck et al. [187] dataset. Accuracy is the sole metric reported in Shabani and Sokhn [427].

pre-trained models also failed to outperform existing approaches, just like their counterparts without pre-training.

The size and diversity of the dataset used during pre-training do not appear to affect the efficacy of the pre-training process, at least not as much as the resemblance of the pre-training dataset to the downstream/target task dataset does. Across all CLS schemes, text preprocessing methods, and pre-training sample sizes, models using weights pre-trained on the target dataset or sorted sentences from the target dataset (SimSents) always perform better than those using weights pre-trained on an external dataset. Still, using weights from extensive pre-training with an external dataset is beneficial, as doing so always results in better performance than comparable models with no pre-training. A number of models from our framework that did not use pre-trained weights fared worse than existing state-of-the-art approaches but managed to outperform the existing approaches when using pre-trained weights extensively pre-trained on the external dataset.

As we expected, fine-tuning with sorted sentences (Title+SimSents) performs better when paired with weights pre-trained on sorted sentences (SimSents) than when paired with unsorted sentences (Target), especially when the pre-training process is extensive. While we also expected that this would translate into SimSents-Title+SimSents models performing better than SimSents-Title+Body, the reverse actually holds true. Furthermore, models using SimSents weights that place themselves among the best performing ones (see Table 3.5) did so when paired with unsorted sentences (Title+Body).

In general, fine-tuning with sorted sentences does not confer the same benefit as pre-training with sorted sentences. Based on the averaged scores, we can observe that models using sorted sentences performed similarly and are always worse than models using unsorted sentences when extensively pre-trained. This is also borne out in a comparison of the models on an individual basis. Among extensively pre-trained models, models that fine-tune with unsorted sentences almost invariably perform better than equivalent ones using

			Pre-training dataset				Pre-trained (mean)
			SimSents	Target	External	NoPretrain	
CLS type	Text	# pre-train samples	F1	F1	F1	F1	
MultiCLS	Title+SimSents	7.5M	.8613	.8478	.8248	.7800	.8446
		30K	.8233	.8187	.7938		.8119
	Title+Body	7.5M	.8630	.8671	.8371	.7913	.8557
		30K	.8077	.8251	.8037		.8122
TBCLS	Title+SimSents	7.5M	.8483	.8564	.8311	.8091	.8453
		30K	.8256	.8295	.7977		.8176
	Title+Body	7.5M	.8679	.8568	.8251	.7872	.8499
		30K	.7983	.8242	.7875		.8033
SingleCLS	Title+SimSents	7.5M	.8627	.8415	.8275	.8005	.8439
		30K	.8247	.8239	.8160		.8215
	Title+Body	7.5M	.8542	.8606	.8471	.8294	.8540
		30K	.8150	.8283	.8071		.8168
Title+SimSents (mean)		7.5M	.8574	.8486	.8278	.7965	.8446
		30K	.8245	.8240	.8025		.8170
Title+Body (mean)		7.5M	.8617	.8615	.8364	.8026	.8532
		30K	.8070	.8259	.7995		.8108

Table 3.6: Distinguishing between fake and satirical news with DistilBERT-based models. Scores under the Pre-trained (mean) column are row-wise means of scores that used pre-trained weights, SimSents, Target, and External. The Title+SimSents (mean) and Title+Body (mean) rows are column-wise means.

sorted sentences, with two exceptions being SingleCLS-SimSents and TBCLS-External.

Our observations led us to conclude that for the task of distinguishing fake news and satire, the best approach is to pre-train extensively on sorted sentences and use unsorted sentences when fine-tuning.

Among all extensively pre-trained models that fine-tune on unsorted sentences, the best models on average are those that use multi-CLS, followed by single-CLS, and finally TB-CLS. Coupled with the fact that two multi-CLS models occupy the top three spots in terms of F1 and AUC-ROC, we believe the multi-CLS class of models to be best suited for the task for classifying fake news and satire, even though the top scoring model uses TB-CLS.

Pre-training is the most important factor in realizing the full potential of transformers as the top models all use extensively pre-trained weights. Without pre-training, the majority of models do not outperform existing state-of-the-art approaches, with SingleCLS-NoPretrain-Title+Body being the sole exception. As the pre-training step is in essence masked language modelling, this implies that developing a language model that is better

Model/ CLS type	Text	# pre-train samples	Pre-training-	
			External	NoPretrain
			Acc.	Acc.
MultiCLS	Title	7.5M	.8133	.8000
	Body	7.5M	.8733	.8533
SingleCLS	Title	7.5M	.8400	.8267
	Body	7.5M	.8733	.8467
Horne and Adali [220]	Title	-	-	.55
	Body	-	-	.67

Table 3.7: Distinguishing between fake and satirical news with DistilBERT-based models on the Horne and Adali [220] dataset. Accuracy is the sole metric reported in Horne and Adali [220].

fitted to the texts found in the target task results in better classification results. As for models using more than one aggregator CLS token occupying the top spots in performance, being allowed to spread out information aggregation over multiple CLS tokens instead of squeezing all the aggregated information into one token may have allowed for more nuanced classification decisions to be made.

3.4.5 Results discussion for generalizability datasets

Our approach generalizes well beyond the Golbeck et al. [187] dataset as it outperforms by a considerable margin the best classifiers that can be found in [220] and [482], which are the only other papers featuring publicly available datasets that distinguish between satire and fake news that we are aware of. Results from the Volkova et al. [482] dataset also shows that our approach is not just applicable to a multiclass scenario but is also capable of correctly classifying minority classes even when a severe imbalance exists in class sizes.

Table 3.7 shows our DistilBERT-based classifier framework’s performance when tested against the Horne and Adali [220] dataset. Results from testing against the Volkova et al. [482] dataset are shown in Table 3.8.

Findings from our evaluations using the Golbeck et al. [187] dataset are mostly equally valid for both generalizability datasets. Models using weights that have been further pre-trained on an in-domain dataset definitively perform better than those using weights that were not pre-trained further. The case for using multiple CLS aggregator tokens is slightly weaker here as SingleCLS models outperform MultiCLS models in the Horne and Adali [220] dataset when classifying an article based on its Title text. This can perhaps be attributed to many article titles being one sentence long, preventing the advantages of using

Model/ CLS type	Text	# pre-train samples	Pre-training-dataset			
			External		NoPretrain	
			F1 Micro	F1 Macro	F1 Micro	F1 Macro
MultiCLS	Body	7.5M	.9984	.9938	.9981	.9923
SingleCLS	Body	7.5M	.9981	.9923	.9979	.9918
Volkova et al. [482]	Body	-	-	-	.92	.71

Table 3.8: Distinguishing between propaganda, hoaxes, clickbait, and satire with DistilBERT-based models on the Volkova et al. [482] dataset. Micro-F1 and macro-F1 are the only metrics reported for multi-class classification in Volkova et al. [482].

multiple CLS tokens from being fully expressed. Both single- and multi-CLS configurations performed equally well in all other scenarios for both datasets. Regardless of the number of CLS tokens used, our models still obtained better scores than previous approaches.

Using the body text instead of the title text when classifying an article in the Horne and Adali [220] dataset always improves classification performance. This concurs with our experience of poor performance when classifying articles in the Golbeck et al. [187] dataset using title text alone during early trials that ultimately led us to exclude this framework configuration choice from in-depth evaluations. Unlike the classifier used in Horne and Adali [220], our approach exhibited a less severe performance gap between using title text alone and using body text alone, with a maximum difference of 0.06 in accuracy score instead of 0.12.

The Volkova et al. [482] dataset is dominated by the propaganda class (30,894 instances), which is several times the size of the satire, hoax, and clickbait classes (4,260 instances combined). Recall that a macro-F1 score is obtained by calculating the F1 on a per class basis and taking the unweighted mean of those scores, while a micro-F1 score is calculated by counting the number of true positives, false negatives, and false positives globally. A macro-F1 score significantly lower than a micro-F1 score in a class-imbalanced scenario indicates that a classifier is worse at correctly labeling the minority classes than it is at labeling the majority class, which is the case with the best classifier presented in the original paper by Volkova et al. [482]. Since our classifier achieved near parity on both micro- and macro-F1 metrics, our classifier is equally adept at correctly labeling the minority classes as it is at labeling the majority propaganda class.

3.4.6 Potential improvements

There are a number of unexplored opportunities for improving upon the existing framework.

3.4.6.1 Different transformer architectures

In this work, the particular transformer variant we are using, DistilBERT, can be fairly described as lightweight. There are many more complex and computationally taxing transformer variants that are not BERT-derived that we have yet to test, e.g., CTRL, GPT2, and XLNet, that have higher token limits or other ways to treat long text sequences that exceed their token limits.

3.4.6.2 Different text summarization techniques

Similarly, there are better text summarization techniques in the literature that we have not explored, such as hierarchical attention networks, or even transformers-based text summarization, e.g., [290]. Summarization techniques that do not rely on using pre-existing title text can also be explored. An example is one that sorts sentences based on the sum of their TF-IDF values. Another example is the use of Code Quality Principle (CQP) to automatically identify sentences highly relevant to a text's topic, which is also automatically identified and is composed of words with high Term Frequencies (TF) [294].

3.4.6.3 News in other languages

So far, American political news written in English has received the lion's share of research attention when it comes to the problem of distinguishing between real, fake, and satirical news, as reflected by the datasets found in wide usage in the literature. Applying transformers to datasets in other languages, such as this French fake news and satire dataset [293] or these Brazilian Portuguese datasets that contain legitimate, fake, and satirical news [127, 233], would allow us to ascertain if transformers can hold their performance advantage across different languages.

3.4.6.4 Outperforming classifier models that involve humans

Crowdsourcing the task of discerning fake news from satire achieved an accuracy of 0.84 [427], a result bested by almost all of our framework's models that have been extensively pre-trained on the target dataset with either sorted or unsorted sentences. Unfortunately, even our best-scoring model attained only 0.8686 in accuracy, which, while close, is still lower than the 0.872 achieved by their hybrid machine-crowd ensembling approach [427].

Developing a machine-only approach that does better than methods involving humans offers several advantages. In a media environment where a high volume of news is being pumped out at every waking hour, human labor can be a bottleneck for these labor-intensive

methods. Methods that depend on humans can also be finicky to tune to achieve the best result. The aforementioned hybrid machine-crowd ensembling approach has to find the correct threshold to determine when the classification label by the ensemble of machine-learning models is insufficient and a news story needs to be evaluated by crowd workers; a threshold value that works well on one dataset often does not work well in another.

3.4.6.5 Improved interpretability

Feature engineering and feature importance are discussions that are very prominent in fake news detection works that rely on traditional machine learning, e.g., [121, 411, 427], but are noticeably absent in works that rely on transformers, e.g., [238, 288], because neural network language models are notoriously difficult to dissect. There is space for future work that focuses solely on the mechanisms by which transformers distinguish between fake and satirical news, perhaps drawing upon existing work on visualizing the innards of transformers, e.g., [112].

3.5 Conclusion

In this paper, we presented a novel DistilBERT-based framework for separating fake news and satire. We demonstrated that with sufficiently extensive MLM pre-training, be it using the target dataset or an external in-domain dataset, a number of our models using different text preprocessing strategies in the fine-tuning stage can achieve results better than existing state-of-the-art methods, with the best models using multiple CLS tokens instead of the standard singular CLS token. Our work shows that when using weights that have previously been pre-trained on large text corpora (Wikipedia etc.), further pre-training with the downstream task's dataset can drastically improve results even when the downstream dataset is relatively minuscule. Successful validation of our approach over several datasets demonstrates that the observed performance improvement over existing approaches is unlikely to be a localized phenomenon but is instead general. Additionally, our approach for telling apart fakes and satires offers an advantage over existing approaches by using open source code throughout the entire process instead of generating features using proprietary code such as IBM tone analyzer in the case of [121] and Linguistic Inquiry and Word Count (LIWC) in the case of [427]. We hope that our work will contribute towards keeping humor (and, to a lesser extent but of no less importance, political awareness and political engagement) from being unintentionally suppressed in our relentless pursuit to eradicate fake news.

Chapter 4

Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets

In the two preceding chapters (Chapters 2 and 3), we focused on the minutiae. That is, we focused on singular solutions for particular scenarios of saturation attacks. This chapter, though restricted in scope to just social media, takes a broader high-level view of informational saturation attacks, where the “solutions” are varied (e.g., hijacking conversations and/or concentrating attention to a single conversational thread) and the textual information are often treated in the abstract.

Specifically, in this chapter, we studied the shifting opinions of a contested issue among the masses on social media. Although we assign a stance to each piece of information (tweet) based on the information’s position with regards to the contested issue, information in this study is treated impartially unlike the preceding two chapters where value judgments were imposed upon the information (e.g., malicious or benign). The goal of the study is to observe how a swarm of presumably human agents belonging to a stance attempt to achieve and/or maintain dominance against other swarms of agents with different stances. This study is a necessary step towards developing an understanding of opinion dynamics that is required for creating an accurate model of contested information diffusion on social media, a model which we will describe in Chapter 5.

4.1 Introduction

In the ongoing COVID-19 pandemic, vaccines are a *necessary* but insufficient method of limiting transmission as vaccines prove less effective against new SARS-CoV-2 variants such as Delta [297, 312, 528] while vaccination campaigns are stymied by vaccine hesitancy

[416] and insufficient global supplies [506]. Masks are important too, with some healthcare experts venturing so far as to call masks comparable or even better than vaccines at stopping the coronavirus [171, 211, 356]. To influence the public to accept protective measures to combat the the COVID-19 pandemic, capturing hearts and minds on social media [89, 331] is essential due to the declining trust in mainstream media [260] and the growing share of social media users, which consists of 55.1% of the global population [249] and seven-in-ten Americans [31].

Most research have focused on charting vaccine discourses on social media; we opt to study the oft-overlooked but equally crucial debates on masking to chart the evolution of the position adopted and the issues raised by social media users. We assigned stances and themes to a large corpus of tweets by training a modified DistilBERT classifier. Stance and theme information are then used in conjunction with the counts of likes, retweets, and replies as well as the ratios of the counts to study temporal trends. Further probing was done by examining TF-IDF word importance and summaries generated by BART.

Our analysis showed that the largest share of tweets are those with a stance supportive of universal masking and featuring a theme of encouraging mask-wearing at the start of the pandemic. As the pandemic progressed, tweets opposing universal masking and a theme of criticizing others' masking behavior slowly gained ground at the expense of supportive tweets although they never became the majority. This can be observed on the right of Figure 4.5, where the share of *encourage* theme shrunk over time while the *appraisal-criticize* expanded. The decline in popularity, noncontroversiality, and agreeableness of supportive tweets abruptly and briefly reversed their trend during an anomalous event between late May and early June. Closer examination of this disruptive anomalous event and cross-referencing with information from another study on the masking conversation on Twitter strongly indicates an attempt to censor tweets of particular themes and stance during that period. Other notable findings include (1) user attention concentrated on a small number of conversations being more effective in altering overall stance than dispersed attention, (2) explicitly politicized debates garnering more attention than non-politicized ones, and (3) narrative from tweets being capable of supplanting narrative of other tweets with the same theme but opposing stance. Our investigation into the anomalous event suggests that future studies involving social media data should not presume platform neutrality and actively work to account for platform owners taking actions that would alter the observed data distribution, e.g., the dearth of a particular viewpoint on a platform may not just be a product of the platform's demographics but also the platform owners working to suppress that viewpoint.

4.2 Background

4.2.1 COVID-19 conversation on social media

COVID-19 struck the global population in an age of widespread social media usage. Researchers have leveraged the unprecedented access to social media data [86] to study the thoughts and concerns voiced by the social media-using crowd, e.g., [48] and studying how information from social media influence their thought processes, e.g., [481]. Massive curated and processed datasets, e.g., [140], have been released to the public to facilitate detailed investigations. Of particular interest to many researchers is the issue of an “info-demic” occurring in parallel with the real-life pandemic, where the indiscriminate sharing of dubious information or outright misinformation risks drowning out sound public health advice, with most studying misinformation in general [15, 106, 326, 382] while [230] focused on vaccine hesitancy misinformation and [153] focused on profiling users who share information on controversial treatments such as hydroxychloroquine. Researchers also inferred the sentiments expressed within the text of social media conversations [432, 518], with [302] taking the unique route of interviewing users about their sentiment *towards* information on COVID-19 found on social media. Researchers have also looked into debates (explicitly oppositional unlike conversations) occurring on social media, such as those concerning vaccination [315]. While stances are touched upon in [315], they are not studied in-depth, with the authors themselves remarking that “[d]eeper analyses on text content should be performed to identify users who are either criticizing the anti-vax movements or supporting them”. All in all, stances on various debatable issues surrounding COVID-19 appear to have received comparatively little research attention.

4.2.2 Mask-wearing conversation on social media

The closest related work we have encountered in our literature review is [418]. Like us, they studied the Twitter discourse on mask-wearing in the context of COVID-19. Their analysis focuses on the automatic labeling of sentiments using lexicon-based VADER and automatic discovery of topics through clustering of embedded representations of tweet texts. In contrast, our work begins from a foundation of manual annotation of stances (which differs from sentiments) and themes (which differs from topics) of tweets.

Sentiments are emotive. Positive and negative sentiments on tweets are not equal to supportive and oppositional stances on mask-wearing. A simple example can help illustrate the distinction between stance and sentiment: if person A is angry at Trump’s cowardly behavior of wearing a mask and person B is angry at the high prices of masks, they both

share the same *sentiment* which is negative (anger) but they do not share the same *stance*, as A is against mask-wearing while B is for it.

Topics in [418] are derived from k -means clustering of tweets's Google's Universal Sentence Encoder embeddings instead of being identified through a manual coding process. As a result, the topics are only mostly reflective of the proximity of words in the embedding space and are decontextualized from the subject of interest, the COVID-19 mask-wearing conversation. This can be observed in [418] having what appears to be overlapping topics, e.g., Cluster 8 with keywords of *wearamask* / *maskwearing* / *masking*, Cluster 14 with *face* / *facemask* / *coronavirus*, and Cluster 13 with *coronavirus* / *face* / *make*. Also, the researchers did not appear to intend to let tweets be associated with more than one topic. In our research, a tweet can be composed of multiple themes.

While there have been studies of an earlier major outbreak of a respiratory disease — the H1N1 pandemic — using social media data, face masks have not been an exclusive concern of these studies. In fact, a 2019 thematic analysis of H1N1-related tweets [8] contained nary a mention of face masks, demonstrating the unimportance of face mask in public conversation on pandemics until COVID-19. A 2010 study [98] has a single finding related to face masks, which was that a campaign comparing the need for masks during the H1N1 pandemic with condoms for AIDS have led to the pandemic's risk being downplayed. A 2011 study focusing on H1N1 disease activity and public concern [439] similarly had little to say about masks, only remarking that spikes in the number of mask-related tweets appear to be timed with public health messages from the CDC and exhibited a downward trend over time.

As for even earlier respiratory disease outbreaks such as the 2004 SARS outbreak, the interaction between them and social media have not been studied. This is very likely due to the fact that these earlier outbreaks have not been truly global in scope and major social media platforms have yet to be established (Facebook was initially a university-only social network that only opened itself up to everyone aged 13 and above in 2006; Twitter was founded in 2006). For seasonal flu endemic to temperate climate zones, they have not generated much social media activity on masking as a topic likely due to the absence of risk messaging in mass media [337] and also the lack of encouragement for mask-wearing from public health authorities.

As far as we know, there are no existing literature on social media data used to study the evolution of stances expressed towards mask-wearing over the early months of the COVID-19 pandemic.

4.2.3 Competing views on social media

There has been much research on identifying the stances expressed in texts or the stances held by people on contentious topics such as climate change, abortion, gay marriage, and gun control, not just on social media but also on in other text repositories such as `idebate.org` and `debatewise.org` [119, 344, 368]. Closer to our research domain is the research on the ongoing vaccination debate on social media, which is an area that has received intense scrutiny, as can be seen by the many published papers that focus on it exclusively as a topic, e.g., [119, 134, 230, 236, 519].

Masks are only ever thought of as being useful against diseases that spread through airborne aerosol and/or droplet routes while vaccines are used to fight all manners of transmissible illnesses. As a result, there is considerably less interest and debate over universal masking compared to universal vaccination in the public sphere, and this extends to social media as well. As even the H1N1 pandemic failed to arouse public interest on debating the role of masks on social media (Section 4.2.2), research on masking debate stances on social media could have only been conducted with the emergence of the COVID-19 pandemic.

4.3 Methods

4.3.1 Data collection

To collect tweets discussing mask-wearing, we used *Twint*, a tool that scrapes Twitter profiles through Twitter’s search operators but does not use the Twitter API, thus bypassing rate limits. *Twint* has been used in academic research dealing in a broad range of topics including COVID-19 discourse [317, 376], disinformation campaigns [165], and citizen engagement [66]. For the data collection’s starting point, we selected January 1, 2020, the date on which China alerted WHO of pneumonia cases of indeterminate cause in Wuhan and the knowledge of a respiratory disease outbreak began entering public consciousness in the Anglosphere.

The search keywords used are “wear” and “mask”. The two search keywords were chosen after performing extensive exploratory searches on Twitter which demonstrated the non-viability of using other keywords. “Face”, “put”, “on”, “don”, “filter”, “cover”, “covering”, “facemask”, “faceshield”, “facepiece”, “respirator”, “scarf”, “bandanna”, “cloth”, “elastomeric”, “N95”, “KN95”, “FFP”, “FFR”, “protection”, “shield”, and their various combinations were excluded either due to their non-specificity or their relative unpopularity. In

the end, only “wear” and “mask” proved sufficiently unambiguous and popular to retrieve tweets where the vast majority are talking about donning protective face covering to combat COVID-19.

The search engine is case-insensitive and even includes tweets with #wearmask hashtag in its results. However, different forms of the keywords “wear” and “mask”, e.g., “wearing”, “worn”, “masked”, “masks”, were not retrieved by Twitter’s search engine. Search results included tweets which have neither “wear” nor “mask” in them; those keywords were instead found in the tweets’ authors’ names, as it was in vogue for Twitter users to include variations of “wear a mask” in their display names. Tweets without keywords in them were excluded from the dataset. The final dataset consists of 2,100,932 tweets with dates ranging from January 1, 2020 to June 21, 2020. We considered refining the results by requiring the tweets to include a term associated with COVID-19 such as “coronavirus”, “pandemic”, or #stayhome, but we decided against it as our exploratory searches found that most tweets did not mention COVID-19 at all and assumed that the context in which mask-wearing is being discussed would be self-evident, e.g., “Pence refused to wear a mask” is far more common than “Pence did not wear a mask, which is in violation of coronavirus health regulations”.

Based on Twitter’s guidelines on data sharing¹, we make available the tweet IDs of our dataset in the publisher’s data repository. Note that tweets are removed rapidly from Twitter either because the tweets themselves are deleted or the accounts associated with them are gone (some possible reasons are going private, being excluded from search, and being deleted), and these actions can either be initiated by the users themselves or by Twitter enforcing moderation policies (e.g., Trump being banned from Twitter) so a portion of our dataset may not be successfully rehydrated.

4.3.2 Classification

We adopted the strategy of first manually labeling a subset of all the collected tweets, then training a machine learning classifier on this human-labeled set, and finally labeling the remaining unlabeled entries in the dataset using the trained classifier. We are not the first to use such a strategy; an earlier paper studying vaccination stances of Twitter users [519] hand-labeled 2% of tweets and machine-labeled the rest.

Each tweet can only be associated with one stance. Stances are mutually exclusive, and the stance assigned to a tweet is representative of the entire tweet. A tweet, however, can have one or more themes. Themes can encompass the entire tweet or sections of it,

¹<https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases>

with themes overlapping the same sections of text possible. Our work is not the same as aspect-based sentiment analysis where each aspect is coupled with its own sentiment. Stance and themes are separate attributes, and while we will explore potential linkages between these attributes in our analysis as in Section 4.4.2, each tweet’s themes are not meant to be coupled with the tweet’s stance from the outset. Sentiment is also not equal to stance as explained in detail earlier in Section 4.2.2.

4.3.2.1 Stance standard

For every perspective (tweet), we test it against the following claim (standard)² to determine the tweet’s stance: *“everyone wearing masks in public will help end the COVID-19 pandemic”*. “Masks” include makeshift fabric face coverings such as bandannas and scarves to professionally manufactured filtering facepiece respirators but excludes costume mask with visible gaps in the mouth and nose areas. “Public” covers all areas outside a person’s place of residence without any exception, be it engaging in vigorous activities, the presence of social distancing, or being outdoors. “Everyone” is literal and does not brook exceptions either, not even people suffering from mental health issues such as anxiety attacks and PTSD or health/respiratory issues such as asthma and COPD.

The standard that we test against may appear lenient in what passes for masks and excessive in what constitutes “public” and “everyone”, but the wording of this standard is an attempt to harmonize technical terms and specific interests of healthcare experts with the non-specialized terms and broad concerns found among Twitter users.

Our standard is based upon a survey of peer-reviewed scientific literature on mask usage against respiratory diseases. The prevailing consensus among the systematic reviews and meta-analyses [74, 105, 284, 308, 309] is that masks could be beneficial in limiting the transmission of respiratory viral infections. Evidence from ecological studies [74, 97, 307] also favors community-wide mask-wearing as an effective measure in controlling the spread of COVID-19. There is a lack of agreement on how different *types* of masks perform as source control and protection in community settings [39, 74, 97, 105, 261, 284, 308, 309, 311], hence the broad definition used in the standard.

We consciously avoided basing the standard upon advisories issued by public authorities, both health (e.g., WHO and CDC) and non-health, as they have changed over time and, in the case of non-health authorities, they may not be based entirely on sound evidence and may be influenced by practical (e.g., material supply) and/or political considerations. The WHO issued an interim guidance on April 6, 2020 [371] recommending against mask

²The “perspective” and “claim” terminology is from [96]

usage in community settings except for those who are symptomatic and people in contact with them, but on June 5, 2020 [372], this was changed to encourage more widespread mask usage among healthy individuals. Cities and counties in California and Colorado have banned valved facemasks out of fear that they function poorly at source control [248], but in December 2020 the CDC published a report showing that masks with exhalation valves are no worse at source control than surgical and cloth masks [387].

4.3.2.2 Stance labeling procedure

There are four stance labels: negative (oppose), neutral, positive (support), and out-of-topic. Their corresponding numeric codes are 1, 3, 5, and 9. Out-of-topic is for tweets that discuss mask-wearing outside the context of COVID-19, such as masking to avoid breathing in volcanic ashfall, and for tweets with insufficient English content to accurately ascertain their stance towards universal masking. Each tweet can only be assigned one out of the four labels.

A number of factors influence the stance assigned to a tweet, and these are the explicit and implicit statements found within the tweet, non-body-text resources (images, videos, and hyperlinks), the conversation which the tweet is a part of (if any), the history of its author (timeline and profile page), and sincerity (humor and sarcasm). An explicit statement is one that states its position on universal public masking in a fairly clear manner, e.g., “please wear a mask”. An implicit statement requires the coder to make a conjecture on its stance, e.g., expressing anger at people in shops not wearing masks typically implies that the tweet is supportive of universal public masking.

The primary goal of the labeling process is to obtain the most accurate stance for the tweets, so we leveraged resources that might not be accessible to a purely text-based machine learning classifier, e.g., images and videos which cannot be parsed, contents of hyperlinks’ destinations which cannot be accessed. Another reason for using resources beyond a tweet and its immediate context is because those alone can be misleading or insufficient to determine the tweet’s stance. An example of extra resources being helpful would be a tweet that appears supportive of mask-wearing, “don’t forget to wear you mask when going out”, that turns out to be anti-mask when one looks at the accompanying picture showing a person wearing a pair of panties on their face.

For tweets advocating for conditional masking, we evaluate their stances based on the tweets’ intended impact on universal masking. Here are some examples. One commonly encountered type of tweet encourages anti-maskers to not wear masks and attend large gatherings, with the anti-maskers falling ill from COVID-19 often being the implied outcome. As these tweets’ intended effect is likely to be scaring everyone into wearing masks by

sincerely advocating for a segment of the population to not wear them, these tweets are classified as pro-mask. There are also tweets advocating for people to wear masks forever so that their ugly faces can be covered up. As these tweets intend to shame people into not wearing masks, they are classified as anti-mask. When a tweet advocates for public figures to wear/not wear masks or criticize public figures for wearing/not wearing masks, we examine the public figures' stances on wearing masks and the tweet author's own historical stance to determine the tweet's true stance. When criticism for not wearing a mask during some public event comes from a tweet author who has a history of anti-mask tweets and the criticism is directed towards a politician who has been an outspoken advocate for universal masking, we classify the criticism to be anti-mask.

A total of 3,010 tweets were manually assigned stance labels. This is the same set of randomly selected tweets whose theme labels were manually coded (see Section 4.3.2.3).

4.3.2.3 Theme identification and labeling procedure

Unlike stance classification where a standard is first defined before labels are assigned to tweets, defining and labeling themes are concurrent processes. What remains the same as with the stance labeling process is the thorough use of all publicly available information to ensure the accuracy of the theme labels assigned to tweets.

We relied on the grounded theory/inductive coding process as described in [286]. In the first round/cycle of coding, concepts were identified through the textual data of the tweets as well as accompanying media (images, videos) from a sample of 3,000 randomly selected tweets from the dataset. As ten total duplicate instances originating from three unique tweets were found during the coding process, we sampled an additional ten random new and unique (non-intersecting) tweets. In the subsequent round of coding, concepts from the 3,010 tweets that demonstrated significant overlap with each other were merged into categories/themes. The final tally is 15 themes. Each tweet can have more than one theme with the exception of tweets that fall under either one of the two out-of-topic themes. Additionally, we picked one theme to be the "main" theme of a tweet. This main theme is *never* used in our analysis due to its low F1 scores but it is used to rectify the very rare issue of the machine learning classifier not assigning any theme at all to a tweet (Section 4.3.2.7).

The themes we have identified are listed below and their list numbering here corresponds to the numerical codes we assigned to them during coding:

1. **Declarative-personal:** Tweet author stating their stance on masking as it applies to themselves and/or people under their care (children), including policies they intend

to enforce upon visiting guests. Tweets recounting personal observations, such as a Twitter user's description of the degree of compliance towards mask mandates at their neighborhood store, also falls under this theme.

2. **Authority-medical:** Tweets where medical authorities directly or indirectly exercise their authority. A medical authority may be an official Twitter account representing a healthcare agency or a scientific institution. The authority can also be an individual emphasizing their healthcare or scientific credentials when making claims about masking, including the deliberate use of technical jargon to appear more knowledgeable or credible. The veracity of the claims — whether they are scientific or pseudo-scientific — is immaterial. The focus of this theme is on tweets leveraging the public's trust in and deference to science (not as a process of knowledge generation but as a source of knowledge) and scientific authorities to compel others to wear or not wear masks.
3. **Encouragement:** Tweets actively encouraging others to wear masks or discouraging others from doing so. Encouragement must be explicit and not inferred. For instance, criticism of people not wearing masks alone is insufficient to count as encouragement.
4. **Appraisal-criticism:** Tweets criticizing or, very rarely, complimenting masking behaviors. This includes jokes, sarcastic remarks, and insults, e.g., mocking mask-wearers for being fearful of COVID-19 and lacking masculinity or mocking non-mask-wearing Trump for lacking masculinity and being afraid of masks ruining the orange cosmetic foundation he wears on his face.
5. **Concerns-effects:** Tweets discussing challenges, concerns, and side effects related to universal masking. The concerns can be about physiological (e.g., breathing difficulty) and psychological (e.g., PTSD) issues brought about by the act of wearing the mask itself. The concerns can also be about masking compliance and issues in procuring masks (cost, lack of supplies etc.). A tweet cannot be assigned this theme based solely on the presence of observations (e.g., tweet author noting that people do not wear masks in stores) within it. Tweet author must state their opinions on their observations (e.g., tweet author expressing fear of being infected by unmasked people in stores).
6. **Individualism-liberty-collectivism:** Tweets where the impact mask ordinances have over freedom, specifically personal freedom and individual liberty, is mentioned. Tweets assigned this theme most commonly emphasize either the need to put collective, communal, and group interests ahead of individual and personal interests

or vice versa. Tweets discussing the freedom to wear masks (e.g., hospitals barring their workers from wearing masks) or not wear masks (e.g., non-masked individuals prohibited from entering stores) also fall under this theme. Another common manifestation of this theme are tweets talking about restrictions on freedom. This can be as simple as a tweet where the author celebrates not *having* to wear masks.

7. **Conspiracy:** Tweets mentioning conspiracies (actions hidden from the general public), such as masks harboring tracking chips or masks being secret signs of subservience to a new world order.
8. **Conditional:** Tweets discussing mask-wearing only in certain situations/circumstances or if the masks meet certain criteria. Examples are masking only when social distancing cannot be observed, masking only if the government allows for shops to reopen and matches in stadiums to be held, and masking only if the masks are of a better grade than normal surgical masks.
9. **Authority-non-medical:** Tweets leveraging non-medical authority to support their stances. The support requested can go beyond making an argument more persuasive, it can also be asking for the authorities to enact or enforce policies, such as asking store owners to ensure customers in stores wear masks. Some examples of the forms which this type of authority can take include political leaders, celebrities, organizations (businesses, countries), law enforcement (agencies, legal rights, legal documents such as a nation's constitution), and religion. The commonality shared by all these authorities is the absence of any ostensible scientific or healthcare credentials. The authority is assumed to have a receptive audience outside of Twitter and/or some means to make others comply with it/its edicts.
10. **News:** Tweets sharing news articles, including tweets made from a media company's official account and journalists reporting news in the form of tweets.
11. **Mask-type-style:** Tweets advocating for certain types of masks. This could be people advocating for masks which are more effective, more comfortable, more stylish etc. It can also be commercial ads for masks and mask-related products.
12. **Questions-jokes-ambiguous-neutrality:** Tweets where only an ambivalent attitude towards masking can be detected. Personal observations and commentaries that do not lean either way in the masking debate count as having ambiguous neutrality. People asking similarly neutral questions, such as those concerning the safety of wearing masks long-term and the technicalities of being in full compliance with mask

mandates, are also ambiguously neutral. Another subcategory are jokes that cannot be construed as being either in favor or against masking; jokes that can be will fall under the “appraisal-criticism” theme instead.

13. **Principled-neutrality:** Tweets that clearly state their neutrality on the mask-wearing debate, e.g., declaring that they do not know enough to advocate for or against masking.
14. **Out-of-topic (mask):** Out-of-topic tweets discussing mask-wearing in non-COVID-19 contexts, such as air pollution from volcanic eruption, seasonal allergies, anonymity, movies, video games, and sexual purposes.
15. **Out-of-topic (language):** A number of tweets mix English with a non-English language, most commonly Tagalog, Urdu, Hindi, or Malay, when discussing mask-wearing. Within the context of this research, Nigerian pidgin does not count as English. If the English portions of the tweets are insufficient for determining the stance and theme of the tweet, then they will be assigned this theme, even if they are clearly discussing masking as it relates to COVID-19.

4.3.2.4 Distribution of metric values over tweets with multiple themes

Since our tweets can have more than one theme, the problem of splitting the share of the values of tweet metrics — likes, retweets, replies — between the themes naturally arises. For instance, if a tweet with *encourage* and *appraisal-criticize* themes received 20 likes, what percentage of those likes is attributable to each theme? There is no information that could guide us in giving each theme its correct share of the values. Equal distribution is our only option. As we felt that equally dividing the values excessively penalizes themes that are commonly found in the company of other themes instead of being alone, we opted to let all themes receive the full share of values. If a tweet received 20 likes, each theme associated with the theme is assumed to be responsible for all 20 likes. This attribution method inflates the metrics of themes but not tweets. In other words, tweets with multiple themes still have the same values for their metrics, but it does result in commonly occurring themes having larger values.

4.3.2.5 Automatic labeling with DistilBERT

BERT (Bidirectional Encoder Representations from Transformers) belongs to the transformers class of neural network architecture which has proven adept at natural language

processing tasks. Unsurprisingly, BERT has been used with great success in determining the stances of text sequences [96, 368, 386]. However, the formulation of stance classification tasks in these existing works are different from ours. In [368], the task is to determine if pairs of text sequences have the same stance towards an issue. In [96] and [386], the goal is to determine if a perspective (e.g., “global warming is a natural cyclical process”) supports or opposes a given claim (e.g., “human-driven climate change is real”). Our work covers only one topic, which is fighting the COVID-19 pandemic through widespread mask-wearing, so there is effectively only one claim. Conditioning the transformer to determine stances based on a claim is unnecessary within our specific application.

The features used for classification are the username, name, date, mentions, and body text of a tweet. A username is unique while a name is a non-unique display name. Mentions are a list of usernames whom a tweet is being directed at. Names may contain text that hints at a person’s stance (e.g., users in favor of masking have often appended “wear masks” to their display names) hence using them as a feature.

We used DistilBERT instead of BERT. DistilBERT runs 60% faster than BERT while still retaining 97% of its performance, which greatly speeds up the labeling of two million tweets. We modified DistilBERT so that each feature text sequence is enclosed with a [CLS] aggregator token in front and a [SEP] separator token at the end. Using the embedded representation from multiple CLS tokens gave better performance than using just one. We settled on this configuration after having tested other configurations, which were (1) concatenating all feature text sequences and enclosing them with only one pair of CLS and SEP and (2) using only a single CLS token at the start but appending a SEP token at the end of each feature text sequence.

Machine learning classifiers are essentially statistical pattern recognition algorithms and are not true artificial general intelligence (AGI). Therefore, they cannot actually understand the guidelines for labeling stances and themes that we outlined in earlier sections. The algorithms instead attempt to seek out patterns within the data that allow them to reliably reproduce the labeling results from human coders, regardless of whether humans relied on the pattern/information during the coding process.

4.3.2.6 Labeling performance

Theme-identification is a series of binary classification tasks as we allow for each tweet to have multiple themes assigned to it. Stance identification is a multiclass classification problem with four classes.

To determine the optimal number of training epochs for maximizing classification performance and to assess the generalizability of the model, we used a nested cross

validation approach that used iterative stratification and relied on micro-F1 and macro-F1 scores for both loops. The scores of the validation sets of the 3-fold inner loop is used to determine the optimal number of epochs. Means of the scores for the test sets of the 10-fold outer loop reported in Table 4.1 estimate generalization error.

Note that prior to classifying the full 2M set of tweets, we retrained the classifier with all 3,010 hand-labeled tweets.

Different problems have a different optimal number of training epochs so we settled on 5 epochs as the best compromise. Our classifier’s performance is comparable to the micro-F1 scores (average of 0.583 for [344]) and macro-F1 scores (average of 0.691 for [344], maximum of 0.7313 for [519] compared to our maximum of 0.7309) for tweet stance classification tasks reported in earlier works. [519] is a single-claim three-class problem (pro-vaccine, anti-vaccine, and neutral) while [344] is a multiple-claim (multiple-target) three-class problem (support, against, and neither). Note that the computation of F1 in [344] is non-standard in that it “does not give any credit for correctly classifying ‘neither’ instances [but] the system has to correctly predict all three classes to avoid being penalized for misclassifying ‘neither’ instances as ‘favor’ or ‘against’.”

# classes	Label	Micro-F1	Macro-F1
4	Stance	0.5451	0.6980
15	Main theme	0.3078	0.5433
2	Theme 1 <i>declarative-personal</i>	0.8692	0.9123
2	Theme 2 <i>authority-medic</i>	0.7849	0.9003
2	Theme 3 <i>encourage</i>	0.8067	0.8266
2	Theme 4 <i>appraisal-criticize</i>	0.7917	0.7930
2	Theme 5 <i>concerns-effects</i>	0.6790	0.9030
2	Theme 6 <i>indiv-lib-collective</i>	0.7430	0.8319
2	Theme 7 <i>conspiracy</i>	0.4910	0.9648
2	Theme 8 <i>conditional</i>	0.5905	0.9037
2	Theme 9 <i>authority-non-medic</i>	0.7838	0.8691
2	Theme 10 <i>news</i>	0.8238	0.9671
2	Theme 11 <i>mask-type-style</i>	0.6358	0.9379
2	Theme 12 <i>question-joke-ambi</i>	0.5031	0.9322
2	Theme 13 <i>neutral-prinp</i>	0.4964	0.9857
2	Theme 14 <i>OOT-mask</i>	0.6297	0.9671
2	Theme 15 <i>OOT-language</i>	0.4987	0.9947

Table 4.1: DistilBERT classification performance for stances, themes as a single label 15-class problem, and themes as multilabel binary class problem. Scores are means from the test splits of 10-fold cross validation after training for 5 epochs.

4.3.2.7 Cleaning labels

A very small number of machine predicted labels have issues. A tweet that has an out-of-topic stance cannot have a non-out-of-topic theme and vice versa. However, there are 5,407 tweets with OOT stances and non-OOT multilabel themes and there is also 1 tweet with a non-OOT stance and an OOT multilabel theme. Additionally, 616 tweets are assigned both OOT and non-OOT multilabel themes. 65,339 tweets were not given a single multilabel theme by the classifier. In total, these tweets with problematic labels constitute 3.40% of the dataset.

To remedy the issue of tweets not having any multilabel themes, we took the predicted main theme as the multilabel theme provided that doing so does not cause an OOT theme to be assigned to a tweet with a non-OOT stance or vice versa. This reduced the number of tweets without any multilabel themes down to 5,789. All 616 tweets with both OOT and non-OOT themes had their OOT-themes removed. We could not address the other issues as we are uncertain if the prediction error lies with the stance labels or the theme labels. As the remaining tweets with problematic labels total 11,197 or just 0.53% of the dataset, their exclusion should not affect our analyses significantly.

4.3.2.8 Examples

To give readers an insight into the stance and theme labeling process, we provide some examples here. The tweets found in these examples are in no way representative of the diversity of tweets that can be associated with these stances and themes.

“The airport is a private place of business that is glad to receive City taxpaying funds. The requirement to have to wear a mask is outrageous. Maybe @MayorGallego should stop giving millions to help them and see if they continue requiring people to wear them.” This tweet has a negative stance and features the *appraisal-criticize* and *authority-non-medic* themes.

“When did I dispute mask wearing? I simply said people should be free to choose their own level of risk tolerance and decide where they want to go. If you have to go into a business and they require masks, you should wear it. But you should be allowed to take any risks you want”. This tweet has a neutral stance and features the *indiv-lib-collective*, *authority-non-medic*, and *neutral-prinp* themes.

“Humans are strange. We all do basic things to survive, like eat and sleep, but those human who feel their rights are restricted or that they won’t wear a mask because “I don’t want to” - they are the reason our new normal is going to last so much longer. Thanks a lot, shitbags.” This tweet has a positive stance and features the *appraisal-criticize* and

indiv-lib-collective themes.

4.3.3 Automated summaries with BART

Taking inspiration from [418], where DistilBART³ was used to summarize tweet clusters, we also used transformers to summarize groups of tweets for analysis. Key differences in our method are that the groups we are summarizing are manually defined and we did not use DistilBART trained on the XSum extreme summarization task. We instead used Facebook’s original BART [276] trained on the CNN summarization task. A desire for more accurate summarization motivated this choice. As we did not need to generate a large number of summaries, the speed advantage offered by DistilBART over BART did not matter to us. We avoided using the weights from the XSum task because XSum summaries are highly abstractive, meaning that the summaries tend not to resemble the source sentences. In practice, this meant that there is a strong tendency for a transformer trained on XSum to “hallucinate” words and narratives that were absent in the seed tweets, giving us misleading summaries. In comparison, summaries in the CNN summarization task tend to resemble source sentences so a transformer trained on this task is much less inclined towards inventing new details.

To examine some phenomenon that interests us through BART summaries, we first filter tweets in our dataset so that only tweets representative of the phenomenon of interest remain. This filter can either be a specific time window, stances, themes, conversation IDs, or some combination of all of them. Usually the number of tweets that remain after filtering are too large to be accommodated by the transformer, which has a length limit on the seed text used to generate summaries. Therefore, a second round of filtering is almost always required to select a smaller number of tweets with even greater representativeness, with the centroid proximity acting as the criterion this time. In the second round of filtering, BART is used to compute the sentence representation (encoded in the special <s> token) for all the tweets that remain after the first round of filtering. The top 30 tweets closest to the centroid of the sentence representations are selected to be the seed text for generating summaries.

Tweets selected as seed text are cleaned by removing hyperlinks, concatenated, and tokenized. If the number of tokens exceeded BART’s 1024-input token limit, they are truncated to fit the limit. BART generates a summary with a maximum output length of 100 tokens from the input tokens.

The method of obtaining the top 30 tweets closest to a centroid was also used by us to

³BERT, BART, DistilBERT, and DistilBART are different but related types of transformer.

study tweets representative of a conversation or of a particular day without generating a summary through BART.

4.3.4 Likes and retweets as proxies for the silent majority

Studying the mask-wearing conversation purely through tweet text alone may give an inaccurate impression on the interests and leanings of Twitter users. This is because many people who register for social media platforms are often not prolific users or are users who engage with the platform in a low-visibility manner; these *lurkers* form a silent majority. Lurkers are not to be confused with churners, who register for a social media service but never use it. Research has indicated that, while not quite nine-tenths⁴, there are still three-quarters of Twitter users who are lurkers [25]. Not accounting for lurkers have been given as a reason for the failure in predicting election outcomes through polling Twitter sentiment [177].

Likes and retweets can offer a glimpse into where the silent majority stand on issues such as masking as both are low effort engagement activities, requiring but a tap or a click from the user. While not as clear-cut as explicit statements, liking a tweet is highly likely an unironic indicator of supporting the tweet's stance. Retweeting indicates that an idea expressed in a tweet is attention-worthy and/or cannot be ignored. If one agrees with a tweet, then positive attention is garnered through exposing the tweet to a sympathetic audience. If one disagrees with a tweet, then negative attention is gained through mockery or refutation by an unsympathetic audience. And if one is ambivalent, increasing visibility of a tweet can help foster debate.

4.3.4.1 Ratiometrics

Ardent users of Twitter have long noticed that disagreeable tweets tend to exhibit a certain tendency on the three metrics — replies, retweets, and likes — visible to the end user, which is that the more disliked by an audience a tweet is, the more the replies outnumber the likes or the retweets that the tweet receives [138, 334, 351]. This phenomenon is known colloquially among Twitter users as “being ratioed” and it has garnered sufficient attention that Merriam-Webster has put “ratio”, “ratioed”, and “ratioing” on the list of words that they are watching [334]. The idea of “ratio” being capable of revealing how a tweet is judged by its audience is gaining traction in scholarly circles as well. There has been a recent pioneering work that focuses solely on examining how ratios of Twitter metrics can be used

⁴Nielsen's rule

to characterize tweets, using a case study of tweets from two highly polarizing political figures, Donald Trump and Barack Obama [341].

In our work we focus on three types of ratios (ratiometrics) — $N_{\text{retweets}} : N_{\text{likes}}$, $N_{\text{replies}} : N_{\text{likes}}$, and $N_{\text{replies}} : N_{\text{retweets}}$. We chose not to divide one raw number by another, e.g., $N_{\text{replies}}/N_{\text{likes}}$, to obtain the ratios because many tweets in our dataset have not received either a single like, retweet, or reply. An excessive number of tweets would have to be excluded due to the denominator being zero if we simply divided two metrics. In the alternate scenario where the numerator is zero, division between two raw numbers also risks destroying information on the differences between tweets (e.g., a tweet with 0 replies and 1 likes is much less popular than another tweet with 0 replies and 100 likes but both would have ratios equal to 0). Using raw numbers could also lead to outliers dominating the results. To overcome those problems, all ratios are calculated using the following equation:

$$\begin{aligned} R_{\text{metric1-metric2}} &= N_{\text{metric1}} : N_{\text{metric2}} \\ &= \frac{N_{\text{metric1}} - N_{\text{metric2}}}{N_{\text{metric1}} + N_{\text{metric2}}} \end{aligned} \quad (4.1)$$

Even with this formulation, there are still many tweets whose denominator terms sum to zero that need to be excluded from analysis. Only 1,247,174 tweets in total were used in our ratiometrics analysis, with retweets-likes, replies-likes, and replies-retweets ratios each having 1,058,102, 1,226,976, and 801,960 tweets respectively.

4.3.5 Word importance

To discover which words were most important to different mask-wearing themes and stances, for each theme (or stance), we divided the sum over all tweets of each word's TF-IDF values by the number of times each word has appeared in the tweet corpus. The count for a word's appearance is boolean per tweet, e.g., a word appearing thrice in one tweet still counts as one. This translates to the following equation:

$$\frac{(\text{theme}_{a \times b})^T \cdot \text{TF-IDF}_{a \times c}}{(\text{theme}_{a \times b})^T \cdot A_{a \times c}} \text{ where } A = \begin{cases} 1 & \text{if term freq.} > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.2)$$

for a tweets, b themes, and c words

The equation above is the same for stances.

We selected only the top 60 words for analysis, with the ranking based on the sum of each word's TF-IDF values across all tweets. English stop words from NLTK and uninformative

words, namely “wear”, “wearing”, “mask”, “masks”, “https”, “www”, “twitter”, “com”, “pic”, “status”, were excluded from our analysis. The uninformative words are either different ways of expressing the key words we used to build our dataset (“mask” and “wear”) or they are words that make up the hyperlinks commonly found in tweets, including links to pictures.

4.4 Analysis

4.4.1 Co-occurrence of themes

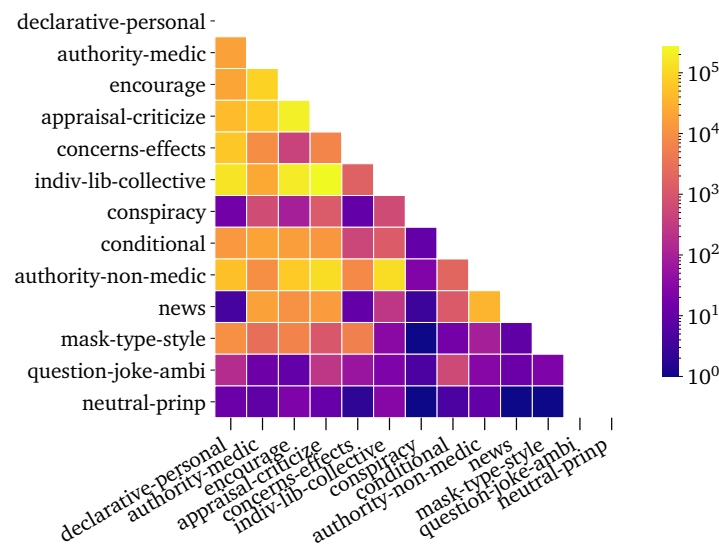


Figure 4.1: Counts of co-occurring themes. Color scale is on a log axis.

Figure 4.1 shows how frequently different themes are found within the same tweet.

When a tweet is appraising the masking behavior of others (*appraisal-criticize*), two other themes commonly appear alongside it, which are those that raise the issue of individual freedoms or collective interests (*indiv-lib-collective*) and those that encourage others to wear or not wear masks (*encourage*). A synthetic example of a tweet embodying all three themes, based on our experience coding tweets: “People refusing to mask up want to kill grandmas. Wear your mask!”

Tweets invoking medical authorities (*authority-medic*) are strongly associated with almost all themes except for those asking questions, joking, or expressing neutrality (*question-joke-ambi* and *neutral-prinp*). Tweets concerned with non-medical authorities that have the ability to influence universal masking compliance such as celebrities, politicians, law

enforcement agencies, and department stores (*authority-non-medic*) also have a strong association with many other themes. Tweets with a *conspiracy* theme is a notable exception, as they are a little less likely to bring up *authority-non-medic* than *authority-medic*.

News coverage being shared in tweets are rarely those that relate to the side effects of universal masking (*concerns-effects*) such as breathing difficulties. News regarding situations where masks can selectively be not worn (*conditional*) are more likely to be found than *concerns-effects*.

4.4.2 Co-occurrence of themes and stances

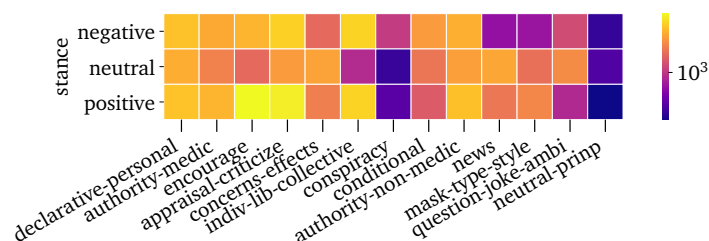


Figure 4.2: The number of tweets belonging to all possible combinations of stances and themes, excluding the out-of-topic stance and themes. Color scale is on a log axis.

Figure 4.2 shows the counts of tweets for different combinations of stances and themes. The most commonly occurring types of tweet in our dataset are those encouraging people to wear masks (positive and *encourage*) and those criticizing, praising, or giving appraisal to the mask-wearing habits of other people in a way that advances an overall pro-mask agenda (positive and *appraisal-criticize*).

Certain themes occur far less frequently among tweets bearing a particular stance towards universal masking. The most prominent example is the *indiv-lib-collective* theme that is commonly found among tweets that raise the issue of personal freedom in choosing to either wear or not wear masks. Few neutral tweets are associated with *indiv-lib-collective*; the bulk of the tweets with the *indiv-lib-collective* theme are evenly split between negative and neutral stances. Discussion or promotion of mask types and styles (*mask-type-style*) are also much more common among tweets which either support or are neutral towards universal masking. Presumably, those who are against masking would not bother themselves with the issue of picking the right mask to wear. Tweets featuring *news* are much less frequently associated with a negative stance towards masking, perhaps reflecting a distrust or disagreement towards universal masking position reported or promoted by mass media sources.

Differences in the distribution of stances are less stark with other themes. The number of tweets discussing mask-wearing on a conditional basis (*conditional*) becomes increasingly common the less positive a tweet is about universal masking. Tweets expressing concerns over issues related to universal masking (*concerns-effects*) are less likely to express either support or opposition to universal masking at the same time.

The findings in this section apply to the distribution of the number of likes, retweets, and replies received by tweets of a particular theme and stance combination as well because of the high Pearson’s correlation coefficients between the counts of tweets, likes, retweets, and replies, which range from 0.9626 to 0.9958 (p -values range from $4.7e-40$ to $1.4e-22$) for all 6 unique combinations of the different counts. As an example, the two most commonly posted types of tweets, positive & *encourage* and positive & *appraisal-criticize*, are also the types receiving the most likes, retweets, and replies.

4.4.3 Word importance

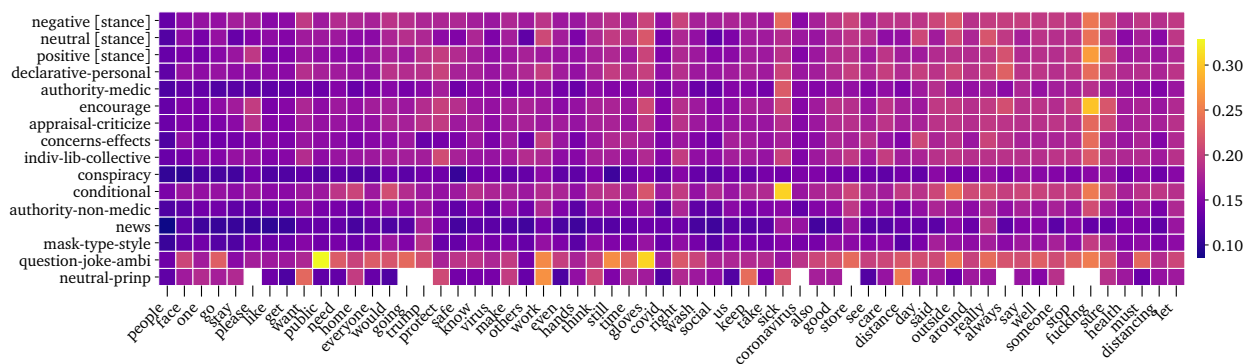


Figure 4.3: Word importance to different stances and themes of mask-wearing tweets.

Figure 4.3 shows which words are important to each theme of mask-wearing tweets based on aggregated TF-IDF scores.

The expletive “fucking” is important to tweets of all themes but especially with the *encourage* theme, which accords with our experience during the coding process where variations of “wear a fucking mask” are commonly encountered. The word “sick” is important to tweets with the *conditional* and *authority-medic* theme, reflecting the many tweets using scientific/medical authorities to bolster the claim that people should only wear masks if they are sick or caring for someone who is sick and not wear them if they are healthy. A slightly less important word for the *conditional* theme is “outside”, which is often used by people proclaiming that masks are for indoors use only and should not be worn in open-air areas. The *question-joke-ambi* theme have multiple words that are important to it,

likely due to the tweets bearing this theme being highly diverse. From examining a random sampling of tweets featuring the theme, the importance of the words “public” and “work” is due to neutral questions or pronouncements about having to wear masks in public and/or at work. The word “glove” is likely due to people uncertain about the necessity of wearing gloves alongside masks.

When comparing word importance scores purely across stances, we observe that the expletive “fucking” is important for all stances but most of all the positive stance. The word “sick” is more important to tweets with a negative stance than other stances and based on our examination of the corpus, the rhetoric of such tweets focus on masks not being able to prevent the getting sick or spreading sickness, reserving mask-wearing only for the sick, and the sick staying home negating the need for masks. The word “please” and “always” is important to positive tweets relative to other stances due to tweets encouraging others to mask up forming a large part of the positive-stance corpus. For neutral tweets, the relatively important words are “really”, “work”, “day”, and “gloves”, which are similar to those for *question-joke-ambi*, reflecting neutral tweets’ tendency to feature people describing experiences wearing masks and sometimes gloves all day on their job and/or questioning the necessity of doing so.

A number of other keywords have low importance scores across all themes but still scored high enough to place among the top 60, which shows that they are important overall to the tweet collection even though they are not important to any particular theme within the context of the top 60 words, and these words are: “coronavirus”, “virus”, “social”, “distancing”, “please”, “stay”, “home”, “keep”, “safe”, “take”, “care”, “protect”, “health”, “wash”, and “hands”. Their presence validate our data collection strategy, as they turned up even though we did not use any COVID-19 or healthcare related keywords. An additional word identified as important to many tweets is “trump”, demonstrating both the US-centric nature of our dataset and the centrality of President Trump to the mask-wearing discourse.

4.4.4 Temporal trends

4.4.4.1 June anomaly

Plotting the counts of tweets, N_{tweets} , over time (Figure 4.6) revealed that a massive drop in volume occurred in a short period of time spanning the end of May (approximately May 26, 2020) and the start of June (approximately June 8, 2020). For convenience’s sake, we shall refer to this event and this time period as the *June anomaly*. We call attention to the June anomaly here because we will be constantly referencing to it due to it altering many aspects of the prevailing conversation dynamic.

The possibility that the June anomaly is a result of some data collection error can be ruled out due to the existence of another, earlier paper studying COVID-19 mask-wearing conversation [419] which showed a similar drop in their plot of tweet count. Crucially, this drop in tweet volume was recorded in spite of a different data collection strategy (coronavirus-related tweets were first collected before further filtering through mask-related keyphrases), a different API (Streaming API), a different set of keyphrases, and covering a different time period (March 17, 2020 to July 27, 2020).

4.4.4.2 Proportions of stances and themes

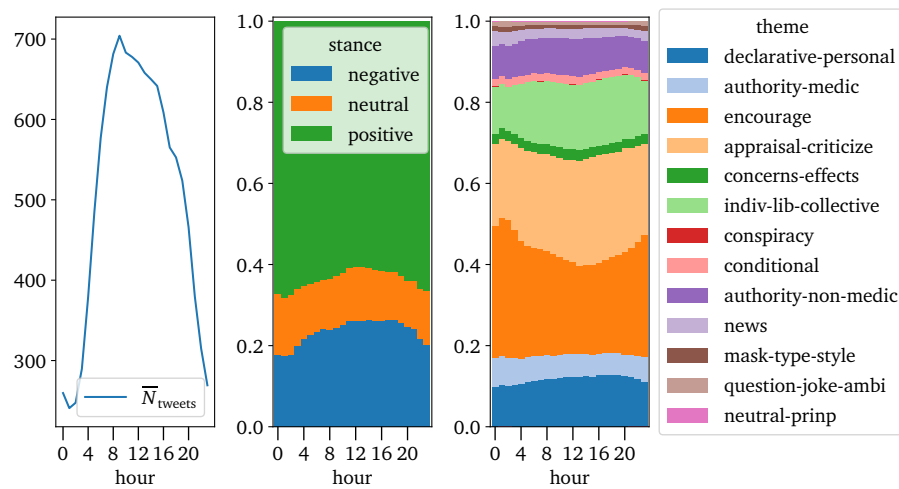


Figure 4.4: Left: mean tweet count per hour. Middle: per hour proportions of stances towards universal masking averaged (mean) over all days. Right: per hour proportions of mask-related tweets’ themes averaged (mean) over all days. Timezone used is Pacific Daylight Time.

Tweets with a negative stance and the *appraisal-criticize* theme exhibit a diurnal pattern, as can be seen in the midday bulges on both stacked bar plots in Figure 4.4. Late in the night and early in the morning, the proportions of tweets with the aforementioned stance and theme shrink, with the lost shares taken up by tweets with positive stance and *encourage* theme. The proportions for the remaining stance and themes remain mostly unchanged. Based on these patterns and the assumption that most tweets during the pandemic are posted during each region’s non-nocturnal hours⁵, the opposition towards masking as well as the harshest criticism for masking behavior appears to be stronger in the Western

⁵Plots of Twitter activity for cities around the world, including two American ones [296, 429], show that activity peaks twice in a day, once in the morning and once more in the evening. Activity is the lowest between midnight and dawn.

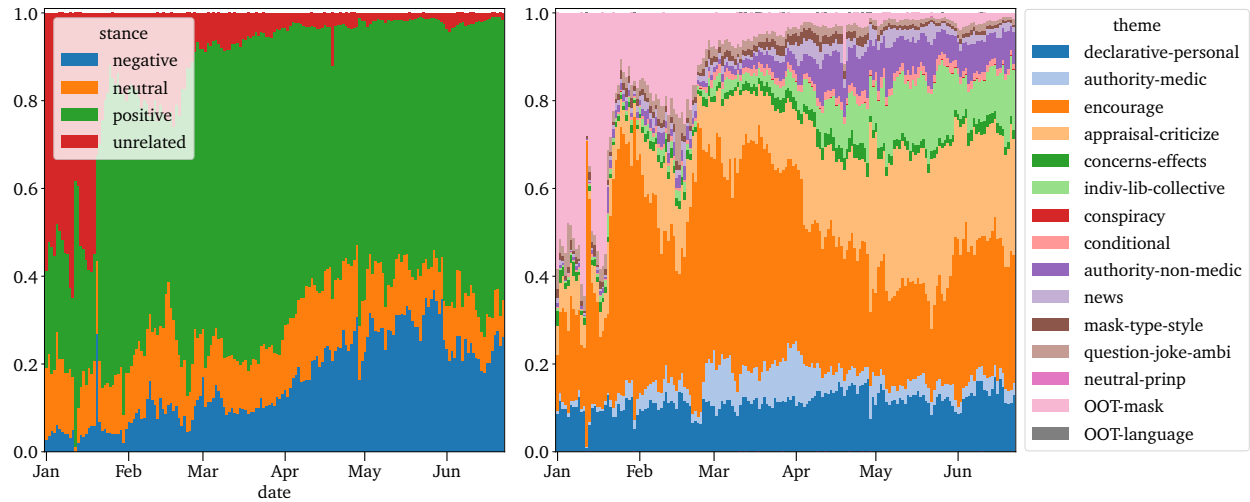


Figure 4.5: Left: temporal evolution of the proportions of stances towards universal masking. Right: temporal evolution of the proportions of mask-related tweets’ themes.

Hemisphere — most likely the US based on the language of our dataset as well as the celebrities and politicians regularly mentioned in tweets — than other parts of the world.

Figure 4.5 portrays the changes in proportions of stances and themes at a much longer time scale of months. Unlike Figure 4.4, Figure 4.5 includes the out-of-topic stance and themes to demonstrate that (1) tweets using a mix of English and non-English language (*OOT-language*) were never a significant minority over the entire dataset’s timespan and (2) the share of tweets that discuss masking in a context other than COVID-19 (*OOT-mask*) and out-of-topic tweets in general (*stance-OOT*, *OOT-language*, *OOT-mask*) virtually disappeared after March.

The proportions plotted in Figure 4.5 allow us to see sudden growth in tweet volume that might be less obvious on raw count plots plotted in Figure 4.6. Take for instance the period between January 12 and January 14, which showed ephemeral spikes in positive stance and *encourage* theme in Figure 4.5. We shall discuss the January 12–14 period in greater detail because it reveals a weakness in our data filtering and classification strategy. Examining the four conversations with the highest tweet counts in that period, we found that they were made up of three standout events, none of which were related to COVID-19. Two of the events could easily be mistaken to be COVID-19 related: fans asking Trump supporter Scott Presler to wear a mask to shield himself from being infected by airborne diseases carried by homeless people and actress Jameela Jamil expressing concern over catching the seasonal flu from recycled air on planes. The third event is the January 12 volcano eruption in the Philippines, which led to many tweets calling for masks to be worn

that did not contain words to indicate their relationship to the eruption, e.g, “ashfall”. Out of the 73 tweets that make up the four conversations, only 20 were assigned an out-of-topic label.

There was another peak of tweets with positive stance and *encourage* theme from the end of January to early February. This time, based on our examination of news articles and the six largest Twitter conversations from this time period, there is much less uncertainty on whether there are misclassifications of out-of-topic tweets; they are not misclassifications. One conversation consists of fans of Filipino boy band SB19 wishing JoshTin well and asking him to wear a mask after that person has fallen ill. Another three conversations were asking for Korean pop stars Cho Seungyon, Rocky, and Jaehyun to wear masks. The fifth conversation along the same vein featured a Chinese pop star named Jackson Wang. Although it is possible that these requests could be in response to seasonal flu, Korean news articles⁶ from this time period spoke of a new trend in the entertainment industry encouraging mask-wearing in response to the coronavirus, ruling out the seasonal flu possibility. Efforts to evacuate citizens of other countries from Wuhan, China also began in earnest near the end of January⁷ and the pandemic was sufficiently threatening that news outlets have begun publishing special features on coronavirus⁸, further lessening the probability that the calls to wear masks was due to seasonal flu instead of COVID-19. The sixth and final conversation, which consisted of people defending their decision to wear a mask due to being perceived as sick and thus ostracized, was unambiguously related to the coronavirus because it featured keywords such as “coronavirus”, “corona virus”, “coronovirus”, and “the virus”.

Our decision to concentrate on analyzing tweets after March 1, 2020 is partially motivated by the potential for misclassification of tweets in earlier time periods as being COVID-19-related; the issue is not a concern post-March 1 based on our analysis of a random sampling of conversations after that date.

Early March saw a peak in proportions of positive stance and *encourage*. The conversations responsible for the peak will be elaborated upon in Section 4.4.4.3 so we will withhold from discussing them here.

We now turn to the overall prevailing trends in proportions of stances and themes from early March all the way until the advent of the June anomaly. In this time period, the proportion of tweets ambivalent towards universal masking have stayed mostly the same. The share of tweets supportive of masking, while remaining a majority, withered slightly

⁶e.g., <https://n.news.naver.com/article/088/0000630884>

⁷<https://www.cnn.com/2020/01/29/world/wuhan-coronavirus-evacuations-intl/index.html>

⁸e.g., <https://www.cnn.com/asia/live-news/coronavirus-outbreak-01-29-20-intl-hnk/index.html>

over time. Tweets against masking though has seen near-uninterrupted growth. In terms of themes, this period saw the rapid growth of *appraisal-criticize*, *indiv-lib-collective*, and *authority-non-medic* themes. Their growth came at the expense of the share of *encourage* and *authority-medic* themes, whose mostly stable proportions in March entered a constant state of decline in April and May. The change in proportions of the themes can perhaps be explained as people growing increasingly tired of admonitions to wear masks, a disinterest or perhaps distrust in advice from medical authorities, a growing interest in criticizing others for their stance on masking, and a growing concern over the enforcement of masking policies and what was believed to be the violation of personal freedom with the introduction of public masking requirements.

The June anomaly saw a decline in the shares of tweets with negative and neutral stances towards masking, with those shares taken up by tweets with positive stance. The proportions of themes found among tweets were also altered. Themes that were previously growing in size, *appraisal-criticize* and *indiv-lib-collective*, now entered into a state of decline. Meanwhile, *encourage* tweets grew its share, reversing its previous trend of decline.

While Twitter may not be the most accurate proxy for the American public due to its global nature, potential censorship (Section 4.5.2), and self-selecting membership/participation, the proportions of stances actually concur quite well with public polls on mask-wearing stances. For instance, a survey conducted by the survey and market research firm SSRS for the Commonwealth Fund from May 13 to June 2, 2020, and involving 2,271 respondents found that “85% of adults believe that it is very or somewhat important to require everyone to wear a face mask ‘at work, when shopping, and on public transportation’.”

9

4.4.4.3 Counts, stance means, stance standard deviations

The left half of Figure 4.6 shows the temporal evolution of the counts and proportions of tweets, conversations, and users. Based on the N_{tweets} line, we see that discussing mask-wearing was not much of a popular topic on Twitter in the early weeks of January, averaging less than 1,000 tweets on most days. In response to growing awareness of a coronavirus outbreak, as discussed earlier in Section 4.4.4.2, there was a rash of tweets urging celebrities to wear mask in late January and early February before interest level dropped off again. Late February was when mask-wearing tweet count began growing almost daily, and late February coincided with the time period when COVID-19 cases were first detected in North America and Europe. At the beginning of May, there was a spike

⁹<https://www.medscape.com/viewarticle/933313>

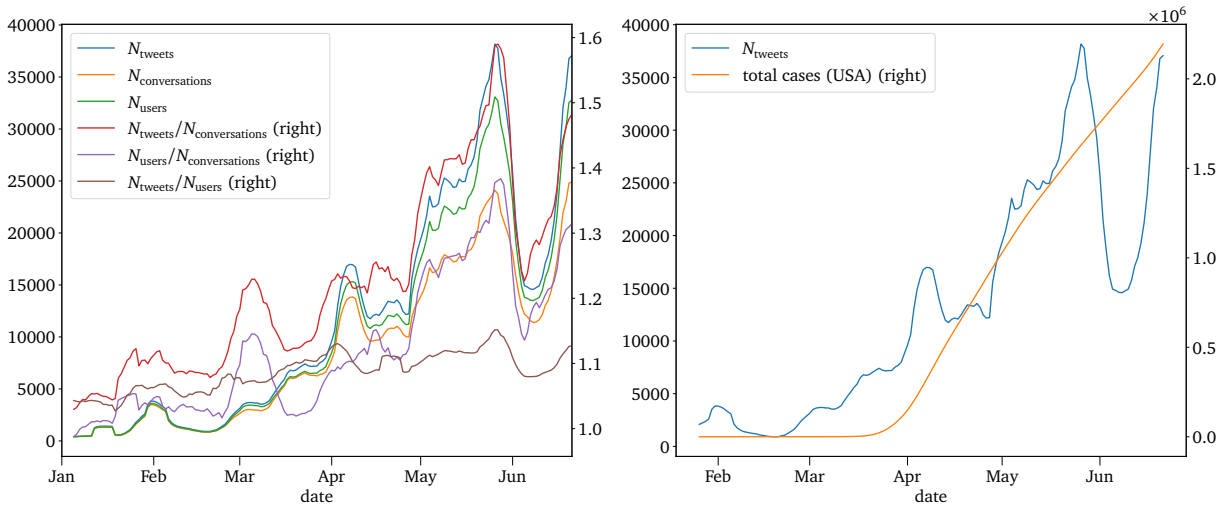


Figure 4.6: Left: seven-day rolling average of per day counts of tweets, counts of conversations (tweet threads), counts of tweets divided by counts of conversations, counts of users divided by counts of conversations, and counts of tweets divided by counts of users. Right: seven-day rolling average of per day counts of tweets with the cumulative total number of coronavirus cases in the US overlaid for comparison.

in the count of daily tweets discussing mask-wearing. The middle of May saw another sharp increase in tweet count. These two volume spikes happened despite declining daily (not cumulative) case count for the two major Anglophone countries, the US and the UK, throughout the entire month of May. The daily tweet count reached a peak of more than 350,000 at the end of May then abruptly dropped off over the next few days to around 150,000 — this is the event which we have been referring to as the June anomaly. So severe is this decline that all the growth in tweet count for the entire month of May was undone, returning to the level found at the end April. The conclusion of the June anomaly is followed by a quick recovery where the number of tweets returned to the same figure as the peak achieved at the very end of May.

Throughout the entire time range of our dataset, the shape of the $N_{\text{users}}/N_{\text{conversations}}$ line generally follows that of $N_{\text{tweets}}/N_{\text{conversations}}$, indicating that the entry or exit of unique users every day is the primary driver behind the changes in the number of tweets found within a single conversation, not the same set of users engaging in prolonged conversation.

When there are peaks in $N_{\text{users}}/N_{\text{conversations}}$ and $N_{\text{tweets}}/N_{\text{conversations}}$ but N_{users} , $N_{\text{conversations}}$, and N_{tweets} stay level, it indicates the presence of “lightning rod” conversations that attracts and accumulates attention. Early March serves as an illustrative example of such a situation. Examining the six conversations with the highest tweet counts around early March revealed that the gradual rise and drop off in interest on masking in this period can be attributed

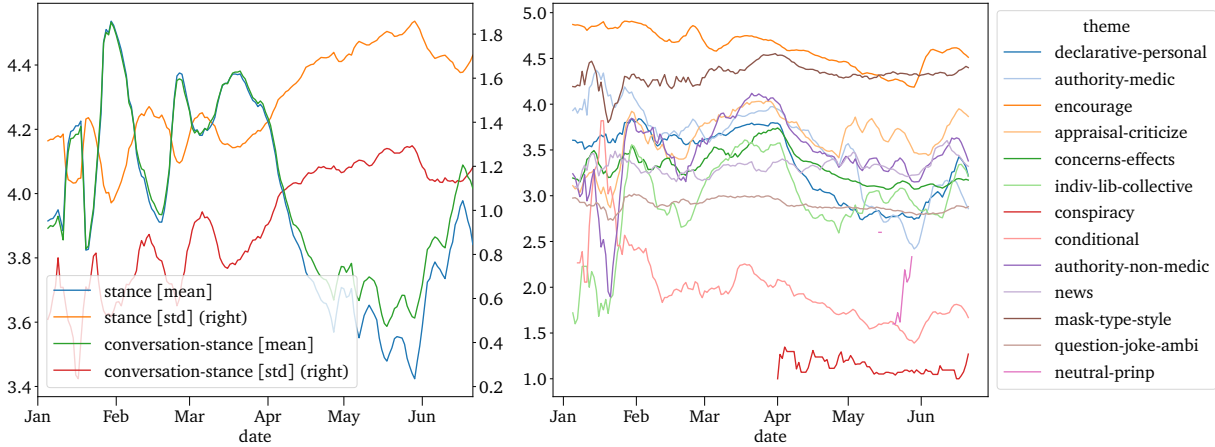


Figure 4.7: Left: seven-day rolling average of per day averages of tweet stances is represented by stance [mean]. Out-of-topic tweets were excluded. Numerical values of 1, 3, 5 stand for negative, neutral, and positive stances. Tweet stances averaged on a per conversation basis, then on a per day basis, is represented by conversation-stance [mean]. Standard deviations of the averages are included in the plot as well. Right: seven-day rolling average of per day averages of tweet stances grouped by themes.

to a confluence of three events: Korean pop fans asking their idols to take better care of themselves by wearing masks (Korean pop fans’ presence on English Twitter is significant enough to engage and succeed in vigilante acts such as taking over racist hashtags [365]), Republican Matt Gaetz’s decision to wear a gas mask on the House floor drawing criticism, and debates over the ability of masks to prevent the coronavirus infection. If we look at the left half of Figure 4.7, we can see that as more participants joined these conversations, it had an overall effect of lowering the mean (less support for masking) and raising the standard deviation (making masking more contentious) of stances over all tweets and also when taking the average of the averages in conversations.

Around early April, a situation that is the inverse of early March’s occurred. $N_{\text{tweets}}/N_{\text{conversations}}$ and $N_{\text{users}}/N_{\text{conversations}}$ declined slightly but N_{users} , $N_{\text{conversations}}$, and N_{tweets} peaked. Users are dispersed into multiple conversations instead of being concentrated into a few large ones. Examining the raw data instead of the running averages showed that the peaks in tweet volume are on April 3 and 4. The largest conversation on April 3 was criticizing someone because their husband refuses to wear a mask. Looking at all tweets in April 3 and 4 without delving into any specific conversations showed a lot of generic tweets calling for people to wear mask, but we also noticed criticisms directed at a certain nebulous “he”, whom we suspect to be President Trump. Our suspicion was confirmed when examining The Guardian’s coronavirus news summary for April 4, which mentioned Trump ignoring US health officials’ advice to people to cover their faces when outside. The dispersed influx of

users and tweets however did not change the stance means and stance standard deviations, which had been trending downwards and upwards respectively since the end of March.

To summarize what we have learned thus far: concentrated attention without increased numbers of users and tweets can cause a shift in mean stance while dispersed attention with increased numbers of users and tweets appears incapable of changing mean stance.

We now explore an additional scenario: an abrupt and simultaneous dispersal of attention and reduction in user and tweet counts. We speak of the June anomaly, which stretches from the end of May to early June. The impact on stance mean is immediately observable; its slow decline was reversed immediately and returned to a local maximum that is comparable to the value at mid-April (left half of Figure 4.7). A reduction in the diversity of opinions is evinced by a drop in the stance standard deviation, which had been trending upwards since mid-March. The June anomaly also caused a number of themes whose mean stances were declining to reverse their trends, most notably *encourage*, *authority-non-medic*, and *authority-medic* (right half of Figure 4.7). The inflection points for these themes were before WHO changed their guidelines in June 5 so it cannot be attributed to the changed guideline.

Although we have previously relied upon the method of examining conversations and tweets from a particular time period to find the potential causes for increases in tweet counts, the same method cannot be used to satisfactorily explain decreases. The presence of tweets is insufficient for explaining the absence of user participation. So, at this point, we can only offer some speculations. One possibility is that COVID-19 is no longer deemed to be a threat and/or people have made peace with their position on universal masking. As a result, people no longer felt the need to proselytize, engage in vigorous apologia for their own position, and/or attack others who did not adhere to a similar position. This was ruled out because after the June anomaly ended, N_{users} , $N_{\text{conversations}}$, and N_{tweets} rebounded swiftly to pre-anomaly levels. We did not collect data beyond June 21, but the plots in [418] showed that counts for mask-related tweets increased up until the end of their data collection period, which was July 27.

The stance means for the *indiv-lib-collective*, *authority-non-medic*, and *appraisal-criticize* themes exhibit highly similar patterns (right half of Figure 4.7). Their mean stances all peaked in late March. Examining the tweets bearing all three themes in that time period showed that this was a result of a rise in complaints about store employees not being given the freedom to wear masks to protect themselves in stores. A month later in late April, their mean stances have all dropped. When we examine the tweets in late April, we saw that another narrative featuring all three themes, one that perhaps has greater appeal than store employee not being able to wear masks, has taken hold. Interestingly, this new narrative

still revolves around stores, but this time around, it is the customers who are criticizing stores for infringing upon their personal rights by requiring them to wear masks.

4.4.4.4 COVID-19 statistics

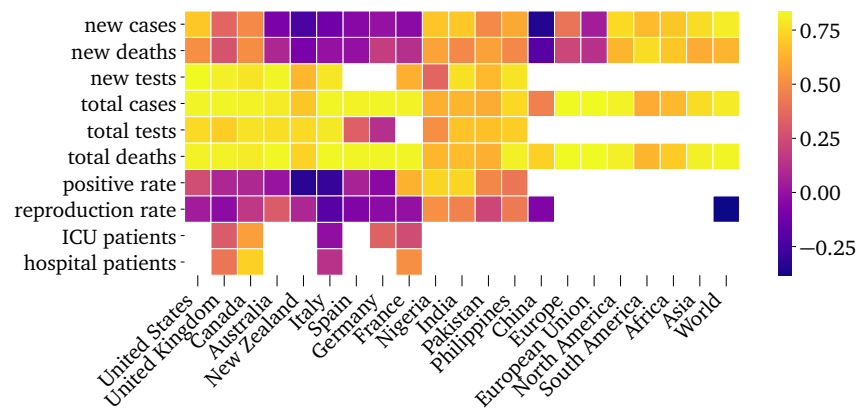


Figure 4.8: Pearson's correlation coefficients between the daily count of mask-related tweets and different daily coronavirus pandemic statistics for various countries and regions.

Case counts, deaths, and other COVID-19 statistics are often reported in news media and used by proponents of masking to encourage masking. The objective of this section is to ascertain if the mask-wearing discussion on Twitter mirrors the “local” development of the pandemic, if they are driven more by the global pandemic situation, or neither. “Local” in this section refers to the US. The prevalence of US-centric entities in the tweet conversations we have examined and the cumulative per hour tweet count seen in Figure 4.4 strongly suggests that Americans, or at least people residing in the Western Hemisphere, are responsible for the vast majority of tweets in our dataset. To compare “local” with non-local, we also looked at the statistics for core Anglosphere countries, countries with large populations of English speakers (whose native languages are coincidentally the ones most commonly encountered for the *OOT-language* theme), non-Anglophone countries, and regions of the world. Our COVID-19 statistics came from Our World in Data (OWID)^{10,11}, and OWID in turn obtained their data from various sources such as official reports released by countries and the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).

Figure 4.8 shows the Pearson's correlation coefficients between the daily tweet count and COVID-19 statistics for various countries and regions. For all countries and regions,

¹⁰<https://github.com/owid/covid-19-data/tree/master/public/data>

¹¹<https://covid.ourworldindata.org/data/owid-covid-data.csv>

total cases and total deaths are the statistics that consistently have high correlations with the tweet count. More obscure statistics, i.e. those that are not regularly reported in news media or reported by countries as evinced by the blanks space in Figure 4.8, tend not to correlate well with the tweet count. While many of the US's COVID-19 statistics for the US have good correlations with the tweet count, they are not exceptionally good compared to other Anglophone countries. Italy, a non-Anglophone country, have stronger correlations than the US for total cases, total tests, and total deaths. New case counts for Nigeria, Asia, and the world have stronger correlations with the daily tweet count than the US's daily new case count.

The June anomaly is not responsible for the low correlations. If we move back the time window to May 21, 2020 when calculating correlation coefficients, the correlations do increase for many countries and regions, but the patterns we have discussed earlier still hold.

The right half of Figure 4.6 depicts the counts of tweets with different stances alongside the US's total COVID-19 case counts. Tweet count and case count lines both follow a general upward trend but they are not close matches.

All in all, the evidence is weak for the tweet count to be directly influenced by the local, i.e. the US, pandemic situation as represented by statistics such as case and death counts. Furthermore, in our examinations of various conversations in Sections 4.4.4.2 and 4.4.4.3, case counts were never explicitly mentioned in them. Therefore, we believe that the likelier driver for new tweets discussing mask-wearing are the actions taken in response to the threat of the pandemic and the chain of reactions towards those actions (e.g., mask mandates, debate about mask mandates, people flouting mask mandates, debates about rule breakers).

4.4.4.5 Ratiometrics and counts of likes, retweets, and replies

Figures 4.9 and 4.11 shows change over time of tweet metrics' ratios, with the former figure broken down by stances and the latter by themes. Figure 4.10 shows change over time of tweet metrics for different stances while Figure 4.12 does the same for different themes. Figure 4.13 shows the daily totals and averages of tweet metrics. As daily means of the counts for likes, retweets, and replies are prone to being influenced by outliers, even after smoothing with seven-day rolling averages, their plots tend to be bursty, making it difficult to extract meaningful patterns from them. Our discussion in this section will therefore focus more on the ratio plots, which feature normalized values.

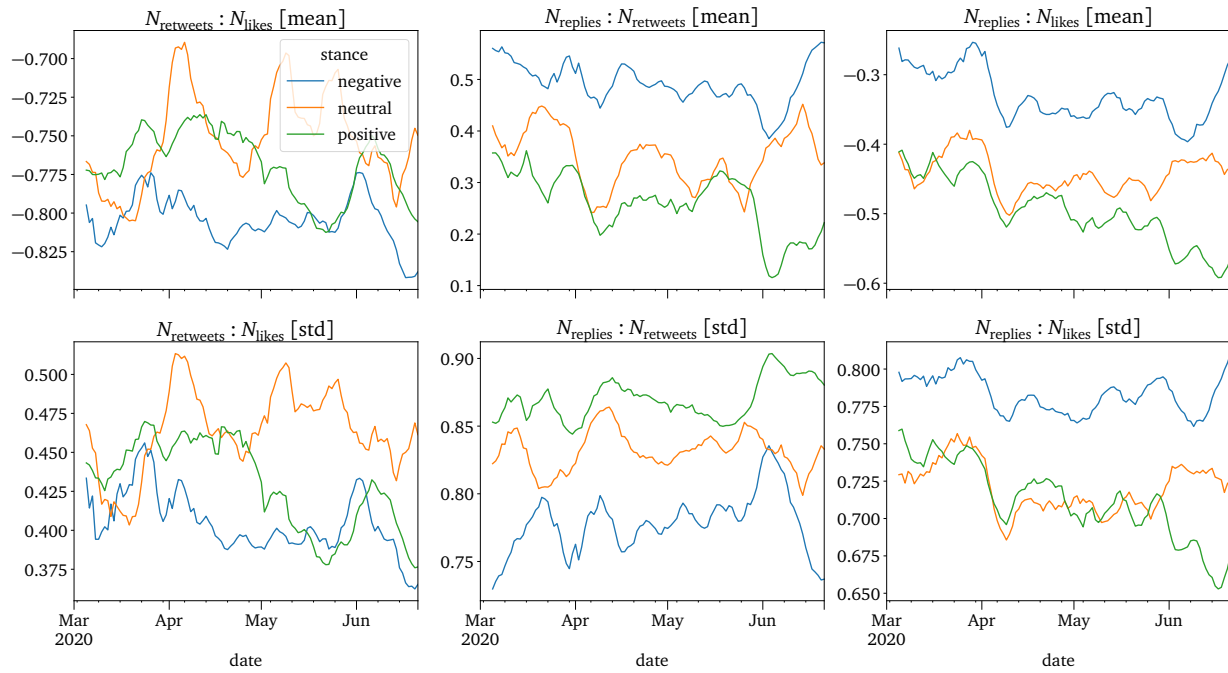


Figure 4.9: Upper row: seven-day rolling averages of various ratios between likes, retweets, and replies received by tweets, grouped by the tweets’ stances. Lower row: standard deviations of the averages.

Interpreting ratios In short: controversial to mildly offensive tweets have positive $R_{retweets-likes}$, mildly offensive to very disagreeable tweets have positive $R_{replies-retweets}$, and very disagreeable tweets have positive $R_{replies-likes}$.

We will work out how we arrived at these interpretations below.

In general, most tweets exhibit the following pattern for their metrics: $N_{likes} > N_{retweets} > N_{replies}$. In our dataset, based on the means, the ratio is approximately 55 : 13 : 4. Ratiometrics are meant to highlight tweets deviating from this pattern. $N_{replies} : N_{likes}$ is the most commonly used ratiometric. If $N_{replies} > N_{likes}$ for a tweet, then that tweet is likely to be considered by its audience to be disagreeable or bad. This is a widely accepted definition of being “ratioed” on Twitter (Section 4.3.4.1). Only 221,214 tweets (10.53%) in our dataset meet the criteria of being ratioed, with the difference in magnitude between $N_{replies}$ and N_{likes} within that small selection being just 1 at the 75th percentile. Using our ratio equation, $\frac{(N_{replies} - N_{likes})}{(N_{replies} + N_{likes})}$, increasingly positive values for the ratio $R_{replies-likes}$ corresponds to a tweet being increasingly disagreeable.

We offer two possible explanations for why replies outnumbering likes makes a tweet disagreeable. First, replying to a tweet is the only way to clearly express disagreement with a tweet because the only other interaction options available to a user are to like or

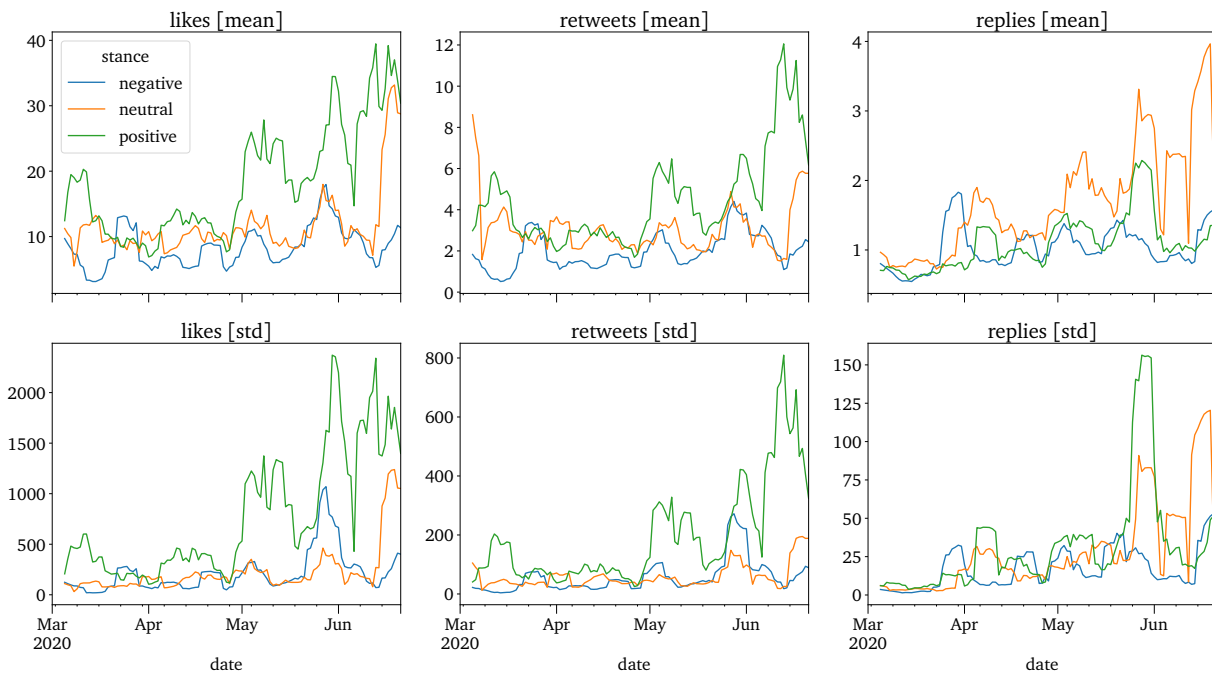


Figure 4.10: Upper row: seven-day rolling averages of the counts for likes, retweets, and replies received by tweets, grouped by the tweets’ stances. Lower row: standard deviations of the averages.

to retweet the tweet. Second, typing out replies is a high-effort method of engaging with a tweet compared to simply ignoring a tweet, making replying a stronger expression of disapproval than ignoring a tweet. Not all replies are antagonistic towards the original tweet, but non-antagonistic replies tend to be accompanied with likes, thus keeping replies from outnumbering likes.

The other two ratios, $R_{\text{retweets-likes}}$ and $R_{\text{replies-retweets}}$, do not have commonly accepted interpretations. Describing these ratios in terms of their relationship with $R_{\text{replies-likes}}$ and with each other can provide some illumination on what they could potentially represent. $R_{\text{replies-likes}}$ has a positive correlation (0.7093) with $R_{\text{replies-retweets}}$ and non-existent correlation (0.05) with $R_{\text{retweets-likes}}$. A negative correlation (-0.7067) exist between $R_{\text{replies-retweets}}$ and $R_{\text{retweets-likes}}$. There are 494,964 tweets with positive $R_{\text{replies-retweets}}$ and 41,851 tweets with positive $R_{\text{retweets-likes}}$. Of the 221,214 tweets with positive $R_{\text{replies-likes}}$, 216,366 (97.80%) tweets also have positive $R_{\text{replies-retweets}}$ while only 4,827 (2.18%) have positive $R_{\text{retweets-likes}}$.

Based on these intersections and correlations, our point of view is that $R_{\text{replies-retweets}}$ serves to identify disagreeable tweets in the same manner as $R_{\text{replies-likes}}$ where positive values indicate disagreeableness but it is a laxer criterion that also ends up capturing tweets that are not as offensive as those that have positive $R_{\text{replies-likes}}$. Positive $R_{\text{retweets-likes}}$ identify tweets with debatable to mildly disagreeable positions — the positions are not so egregious

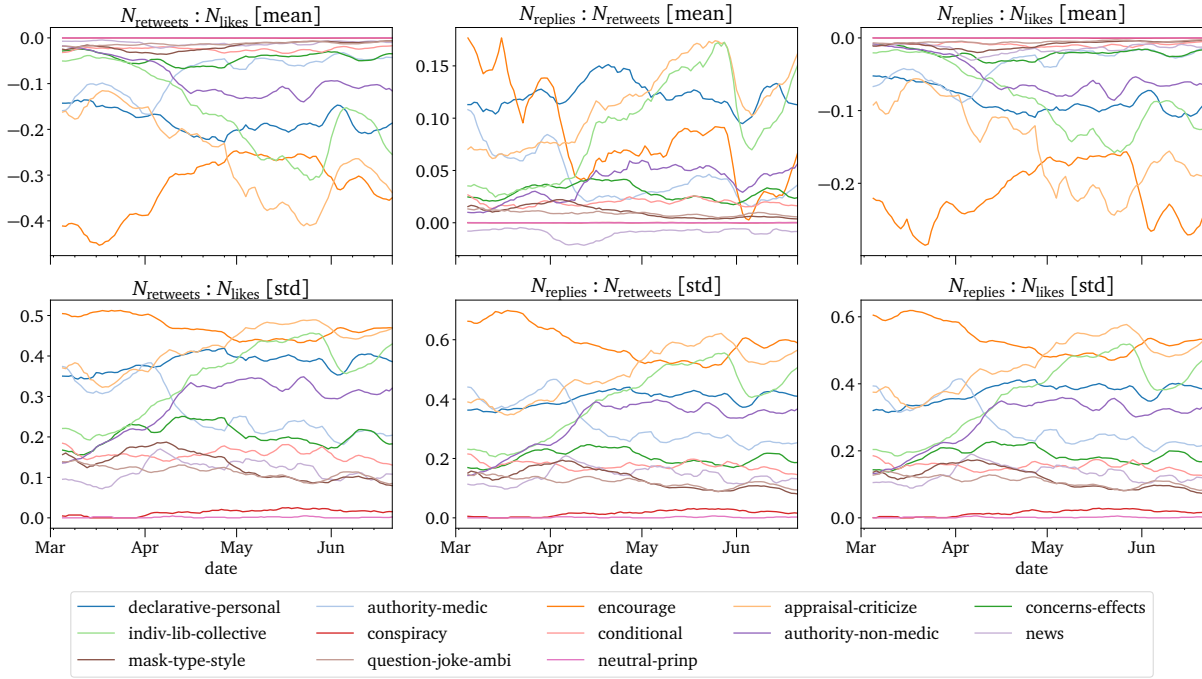


Figure 4.11: Upper row: seven-day rolling averages of various ratios between likes, retweets, and replies received by tweets with different themes. Lower row: standard deviations of the averages.

that their replies outnumber likes, hence it having near-zero intersection with positive $R_{\text{replies-likes}}$. This interpretation of $R_{\text{retweets-likes}}$ also dovetails with the negative correlation that it has with $R_{\text{replies-retweets}}$. As tweets go from contentious to being actively offensive, replies are more likely to outnumber retweets, leading to increasingly positive $R_{\text{replies-retweets}}$, while retweets are less likely to outnumber likes (not due to more people liking the tweets but perhaps due to fewer people retweeting the tweets), leading to negative $R_{\text{retweets-likes}}$.

One extra factor that needs to be kept in mind when examining the ratios for themes is that many themes are much rarer compared to other themes and thus have very low engagement, as can be seen by their respective proportions in Figure 4.5. This translates to few tweets with likes that outnumber retweets or replies by very significant margins, resulting in ratios that stay close to zero. If we look at the standard deviations of ratios (lower row of Figure 4.11), the only themes that are in the same league in terms of the magnitude of variations are *encourage*, *appraisal-criticize*, *indiv-lib-collective*, *declarative-personal*, and *authority-non-medic*. Therefore, we will avoid comparing ratiometric trends for themes outside this group.

The factor mentioned above does not apply to stance ratios as all three stances have similarly high standard deviations (lower row of Figure 4.9).

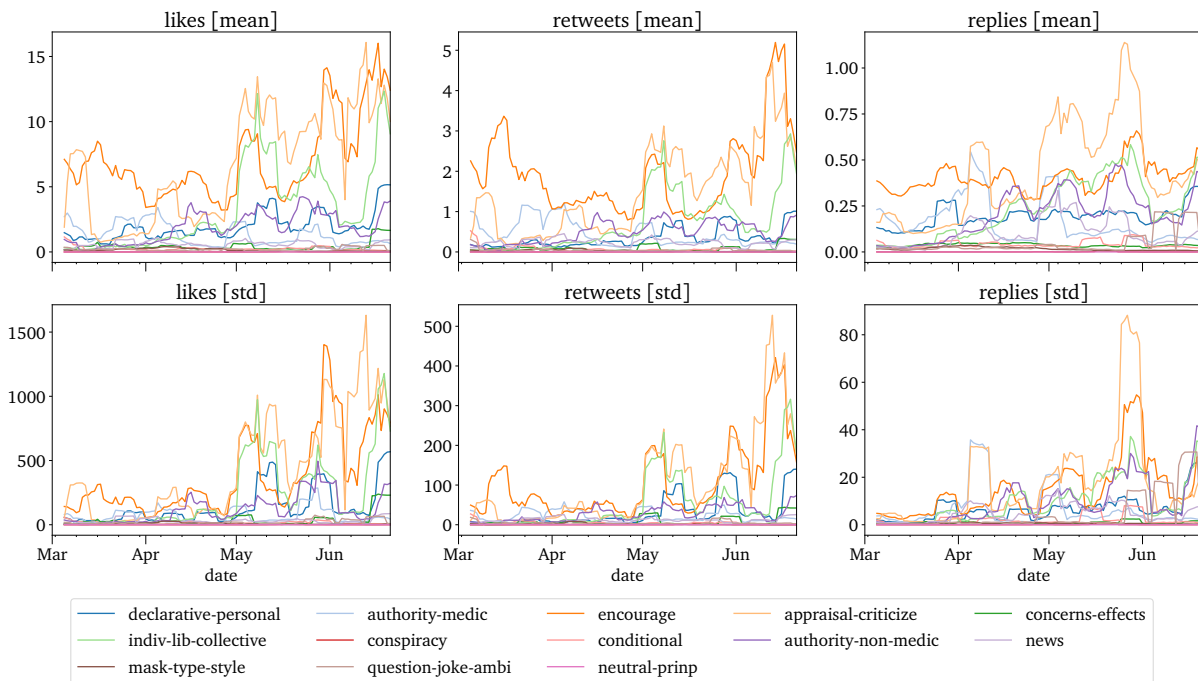


Figure 4.12: Upper row: seven-day rolling averages of the counts for likes, retweets, and replies received by tweets with different themes. Lower row: standard deviations of the averages.

Ratiometric and metric trends

Stances From March until late June, the daily average $R_{\text{retweets-likes}}$ for neutral tweets stayed above that of positive tweets most of the time, which in turned stayed above that of negative tweets all the time (Figure 4.9 left). In other words, neutral tweets are likelier to be merely controversial or slightly offensive. This could be because neutral tweets, by not committing to a position, are likelier to engender debate among its audience but are less capable of immediately making the audience dislike them compared to negative tweets. For the daily average $R_{\text{replies-retweets}}$ and $R_{\text{replies-likes}}$, the stances ranked in descending order goes negative, neutral, and positive (Figure 4.9 top center and right); negative tweets are more likely to be considered to be disagreeable than neutral or negative tweets.

From early April to late May, $R_{\text{retweets-likes}}$ for tweets with positive stances was on a downward trajectory, becoming almost as uncontroversial as tweets with negative stances. This does not automatically mean that people agree more with positive tweets. In the same period of time, looking at the $R_{\text{replies-retweets}}$ plot, we see the values for positive tweets going up, implying that positive tweets might be transitioning from being uncontroversial to being disagreeable. The June anomaly briefly reversed those trends, making positive tweets controversial again but much less likely to be disagreeable. Post-June anomaly saw

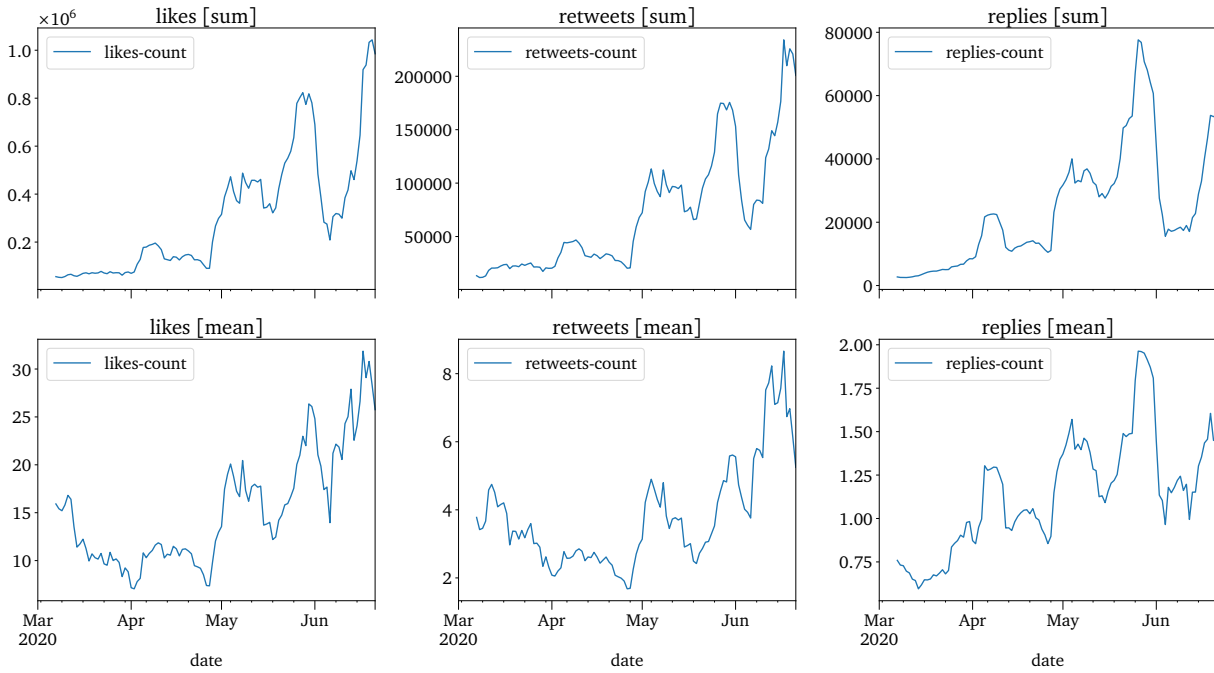


Figure 4.13: Seven-day rolling averages of the total counts (upper row) and mean counts (lower row) of likes, retweets, and replies received by tweets

another trend reversal for positive tweets but some gains were kept; positive tweets are once again less controversial and more disagreeable but it is not less controversial than negative tweets and the growth in disagreeableness only brought positive tweets back to the same level as the minimum in early April and not the peak in late May. The trend for extreme offensiveness ($R_{\text{replies-likes}}$) never saw significant changes for positive tweets, declining from March until late June.

The disagreeableness ($R_{\text{replies-retweets}}$ and $R_{\text{replies-likes}}$) for negative tweets was on a very slight down trend from early April to late May. The June anomaly period itself did not alter that trend. However, post-June anomaly, disagreeableness shot up. As for contentiousness ($R_{\text{retweets-likes}}$), it was on very slight upward trend for negative tweets from mid-April to late May, increased sharply during the June anomaly, then declined sharply post-June anomaly. The decline in contentiousness coincided with the increase in disagreeableness.

Around the start of April, due to Trump ignoring his administration's own mask mandate (Section 4.4.4.3), disagreeableness for tweets of all stances increased briefly before decreasing sharply (Figure 4.9 top center and right). The influx of users (Figure 4.6) at first likely caused further polarization, causing the initial increase in disagreeableness. But as the incoming users were increasingly dispersed among many conversations instead of being concentrated in a few, which is not conducive to creating a situation where a small number

of tweets can receive the majority of attention and replies, disagreeableness decreased.

The daily averages for likes, retweets, and replies (Figure 4.10) shows that positive tweets receive more likes and retweets than tweets of other stances from March through June. Between neutral and negative tweets, neutral tweets have a slight advantage in numbers. As for replies, neutral tweets receive the most on average while positive and neutral tweets have comparable figures, perhaps a reflection of neutral tweets' status as a place for the debate and exchange of viewpoints. In the days leading up to the June anomaly, we can see that a small number of negative tweets have received more likes and retweets than is usual for that time period, as we can see from the spikes in standard deviations (Figure 4.10 lower row). During the same time period, a small number of positive tweets have also received a very high amount of replies, which is evident in both the mean and standard deviation plots (Figure 4.10 right column).

Themes To compare the engagement received by tweets with different themes, we plotted their ratiometrics in Figure 4.11 and the mean counts of their likes, retweets, and replies in Figure 4.12. We focus only on the *encourage*, *appraisal-criticize*, *indiv-lib-collective*, *declarative-personal*, and *authority-non-medic* themes due to the reason laid out near the end of Section 4.4.4.5.

From early March to late May, tweets with an *indiv-lib-collective* or an *appraisal-criticize* theme became less and less controversial ($R_{\text{retweets-likes}}$). The June anomaly caused these two themes to very suddenly become more controversial, although post-June anomaly the contentiousness of the two themes once again trended downwards. In contrast, *encourage* tweets became more controversial from March until the end of April, then stayed the same level of contentiousness throughout May. June anomaly brought about a downward trend for *encourage* tweets' contentiousness that persisted until late June. *Declarative-personal* and *authority-non-medic* tweets became less controversial from March to the mid-April then stayed mostly level after that, remaining mostly unaffected by the June anomaly.

In terms of disagreeableness, tweets with a *indiv-lib-collective* or a *appraisal-criticize* theme have similar trends. More of them are likely to be found mildly offensive ($R_{\text{replies-retweets}}$) as time progresses from March until late May. However, tweets with those two themes are less and less likely to be found very offensive ($R_{\text{replies-likes}}$) over the same time window. During the June anomaly, those trends reversed briefly, i.e. less likely to be mildly offensive but more likely to be very offensive, before reverting again post-June anomaly, i.e. more likely to be mildly offensive but less likely to be very offensive. However, they did not reach a level as low as before the June anomaly. For *encourage* tweets, they became less likely to be mildly offensive from March to mid-April and less likely to very offensive from from early to

late March. After those two points in time, *encourage* tweets became more likely to be both mildly offensive and very offensive. During the June anomaly, *encourage* tweets suddenly became much less likely to be mildly offensive and severely offensive. Post-June anomaly, *encourage* tweets stayed less likely to be severely offensive although they once again become likelier to be mildly offensive. The *declarative-personal* and *authority-non-medic* themes became more likely to be mildly offensive and less likely to be very offensive from March until late April. For the remaining rest of the time, there were some fluctuations on their level of offensiveness but none of the changes were as severe as that experienced by the other themes we have previously discussed.

The *indiv-lib-collective* and *appraisal-criticize* themes having the highest co-occurrence with each other (Figure 4.1) may explain why their ratio trends are so tightly intertwined. Co-occurrence does not offer a sufficient explanation for the similarity in trends seen in *declarative-personal* and *authority-non-medic*, as they do not have the highest co-occurrence with each other. For *declarative-personal*, it is highest with *indiv-lib-collective*. For *authority-non-medic*, it is highest with *appraisal-criticize* and *indiv-lib-collective*.

The mean counts of likes and retweets of *encourage* and *appraisal-criticize* tweets have been comparable since early April. The replies received by the two themes, however, are not. *Appraisal-criticize* tweets are prone to sudden surges of replies and they received far more replies than *encourage* tweets during the entire month of May. *Encourage* tweets, in contrast, received a consistent number of replies from March till June. In the days leading up to the June anomaly, a small number of tweets with either *appraisal-criticize* or *encourage* tweets did receive an extremely large number of replies compared to other tweets bearing the same themes, the result of which can be seen on the standard deviations of replies for the two themes (Figure 4.12 lower right).

4.4.4.6 BART summaries of the five largest conversations in late May

Out of the ten largest conversations found in our dataset (Table 4.2), five of them occurred in late May, during or just before the June anomaly, four in mid-June, near the end of our data collection period, and one in mid April. BART was used to summarize the five May conversations that could provide us with clues as to what might have potentially led to the June anomaly.

The starting dates (Dates (min.)) and summaries of the conversations, listed in the same order as Table 4.2, with spelling, punctuation, grammatical, and other errors preserved intact:

- 2020-05-25: The only reason Trump doesn't wear a mask is he has no leadership

Conversation ID	Date (min.)	Date (max.)	# tweets	# users	# likes	# retweets	# replies	stance	themes (theme counts)
1265045009323241472	2020-05-25	2020-06-20	4,162	3824	117,733	21,965	73,688	4.08	<i>appraisal-criticize</i> (3,584) <i>indiv-lib-collective</i> (1,460)
1265445256465731586	2020-05-26	2020-06-15	2,456	2240	523,654	75,573	41,893	2.81	<i>appraisal-criticize</i> (1,650) <i>encourage</i> (972)
1264721468761481216	2020-05-24	2020-06-13	2,260	1728	20,070	828	2,583	3.15	<i>indiv-lib-collective</i> (1,513) <i>appraisal-criticize</i> (1,258)
1266406045544579073	2020-05-29	2020-06-06	2,182	2022	8,028	1,062	1,160	2.39	<i>indiv-lib-collective</i> (1,586) <i>appraisal-criticize</i> (1,084)
1273295136429084672	2020-06-17	2020-06-21	2,171	1779	37,045	12,471	9,353	3.07	<i>indiv-lib-collective</i> (1,387) <i>encourage</i> (1,077)
1271966131230699520	2020-06-13	2020-06-18	2,164	2111	70,991	3,366	26,623	4.64	<i>declarative-personal</i> (1,773) <i>indiv-lib-collective</i> (459)
1250130325356888064	2020-04-14	2020-06-01	2,126	1998	4,454	222	547	3.89	<i>encourage</i> (1,421) <i>appraisal-criticize</i> (1,136)
1272953220520988673	2020-06-16	2020-06-20	2,032	1975	180,521	20,341	3,468	4.86	<i>declarative-personal</i> (1,944) <i>indiv-lib-collective</i> (331)
1264248352683646976	2020-05-23	2020-06-09	1,758	1507	11,345	780	2,188	2.44	<i>appraisal-criticize</i> (1,055) <i>indiv-lib-collective</i> (787)
1273018665391099904	2020-06-16	2020-06-20	1,664	1515	25,756	1,942	2,313	3.88	<i>indiv-lib-collective</i> (1,242) <i>appraisal-criticize</i> (1,208)

Table 4.2: The IDs of the ten largest conversations found in our dataset along with the earliest and latest dates of tweets associated with the conversation found in the dataset; the counts of tweets associated with each conversation; the total number of users, likes, retweets, and replies; the mean stance; and the top two themes and their respective counts.

qualities. He doesn't wear a masks in public because he's a dumbass. And he could never out-cool President Biden. He looks like a responsible, compassionate human being. If you care about those around you, you'd worn a mask too.

- 2020-05-26: "Wearing a mask is useless if you don't wear it correctly! If you don't wear a mask, you ain't white. You're entitled to stay home a wear a masks. You can take your mask and shove it where the sun don't shine," she said.
- 2020-05-24: Wearing a mask at the grocery store isn't Nazi Germany... If you think a mask will protect you feel free to wear one. I'm not the dum dum going out coughing my lungs out on other people. If you refuse to wear a mask, don't go into a public space.
- 2020-05-29: "If you honestly believe a mask will save you, wear it. If that makes me not nice in some peoples minds, I can live with that" "I wear a mask when I go to the store, but I will admit I lift it numerous times to get air" "The anxiety when wearing one is much worse than my fear of the virus & I I'm in the high risk category"
- 2020-05-23: I don't give a shit if you wear a mask anymore than ifYou wear a t-shirt or a sweater. How is wearing a mask make one suffer? Just wear it. If you're going to wear a masks still, while driving your car, alone. You need to just stay home, you don't deserve to be out. I'm hoping that knitted mask (that has holes) has a filter in between - otherwise,

From these five summaries, the situation of the mask-wearing discourse on Twitter just at the start of the June anomaly aligns with the stance proportions in Figure 4.5, where approximately three-tenths of the tweets in late May are anti-mask. The conversations dated 2020-05-26, 2020-05-24, 2020-05-29, and 2020-05-23 contain sentences that are anti-mask, some definitive (e.g., “You can take your mask and shove it wherethe sun don’t shine”) and some ambiguous (e.g., “If you honestly believe a mask will save you, wear it.”). The conversation dated 2020-05-23 appears to concern Alyssa Milano wearing a crocheted mask in a car¹². While participants of the Alyssa Milano conversation may be sincerely advocating for wearing better masks, some among them might be hijacking it to push an anti-mask agenda, e.g., telling people who voluntarily wear masks in cars to stay at home instead, hence the low mean stance. The largest conversation, with a start date of 2020-05-25, contained references to Trump and Biden, demonstrating that the politicized debates on mask-wearing attract far more attention than those without.

The presence of these five May conversations challenges the idea that the sharp drop in tweet volume during the June anomaly can simply be explained away as just user interest naturally falling off. First, to have five of the ten largest conversation in the entire dataset be located within a small time window in late May suggests that interest in tweeting about masking remains very strong at the beginning of the June anomaly, and if interest were to decay, it is likelier to decay slowly and be more reminiscent of what was seen in at the start of April. Second, the sentences in the summaries are a mix of positive, neutral, and negative in terms of their stances towards masking, suggesting that Twitter users have not all converged upon a single stance on the issue of mask-wearing and thus have lost interest in debating it. Third, going by the dates of the latest tweets to be posted within those five May conversations in Table 4.2, we can see that interest in them persisted until early and mid-June.

Monolithic conversation The second last conversation with a start date of 2020-06-16 (conversation ID: 1272953220520988673), which has the highest mean stance out of the ten listed at 4.86 out of a maximum of 5, serves as a good example of a monolithic conversation where tweets exhibiting a single stance overwhelmingly dominated. The conversation consists of people comparing the triviality of wearing masks to the true hardships suffered by generations past. The BART summary:

- 2020-06-16: My father survived Pearl Harbor, Iwo Jima, Guadalcanal & severe malaria. My uncle survived the Bataan Death March. My grandfather had to hide under his

¹²<https://news.yahoo.com/alyssa-milano-defends-totally-safe-230603836.html>

dead soldier brothers in the Pacific Ocean campaign. My grandmother plowed fields in her bare feet. I am proud to wear a mask.

4.5 June anomaly: censorship?

4.5.1 Twitter's history of censorship

That Twitter engages in censorship, whether crudely through outright removal of tweets or more skillfully by reducing a tweet's visibility or by coloring user perception of a tweet's content, is not some outlandish supposition.

While the process is opaque, Twitter's heavy-handed moderation tactics is a publicly known fact. Banning and removing tweets are not the only tools at the moderation team's disposal. Twitter has clarified that while it does not engage in shadow banning¹³, it does deliberately and severely reduce the visibility of tweets from "bad faith actors" [170], e.g., search result rank manipulation to bury targeted tweets, essentially consigning tweets to death through near-obscurety. As it pertains to the COVID-19 pandemic, Twitter has implemented a new company policy on disinformation and removed tweets they deemed harmful from highly visible world leaders [172] back in April of 2020, which is within our data's date range. And more recently in April 2021, well outside our data date range, Twitter has removed tweets critical of the Indian government's coronavirus response at the behest of the Indian government [306].

There has been research on the plausibility of shadow banning on Twitter [333], with the authors finding that visibility limitations on user profiles being bugs, as claimed by Twitter, "is statistically unlikely with regards to the data [the authors] collected". Twitter has also attempted to steer certain conversations through the use of soft moderation tools like the very recent policy of adding warning labels to tweets concerning presidential election results and COVID-19. This new warning label policy has also garnered the attention of researchers [520], who found that among tweets stuck with the label, 72% were shared by Republicans and 11% by Democrats, perhaps demonstrating a political bias. Researchers have also suggested that Twitter's failure to communicate the reasons for censoring content has led to many users believing that censorship is politically motivated [455].

¹³"[D]eliberately making someone's content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster" [170]

4.5.2 Evidence from our dataset

Throughout Section 4.4.4, the methods we used to probe our dataset have provided simple explanations for the causes behind temporal trends found in our dataset. The only exception is the June anomaly, where instead they have served to successively eliminate simpler explanations until censorship became the simplest remaining explanation. We collate our findings here.

Aside from the anomalous drop in tweet count, user interest in mask-wearing as a topic did not show signs of abating during the June anomaly. A heated debate was still occurring prior to the June anomaly, judging by the high number of replies that tweets were receiving in late May (Figure 4.13 and Section 4.4.4.5). Five out of the ten largest conversations in our dataset started around late May and the latest tweets within those conversations can be found after the June anomaly, demonstrating the there is continuity in user interest (Section 4.4.4.6). The swift recovery in tweet count post-June anomaly is another sign that user interest did not truly decline (Section 4.4.4.3). The June anomaly is therefore unlikely to be a product of user interest suddenly dropping off.

The unequal impact that the June anomaly has had on tweets with different stances and themes is further evidence that the anomaly was an engineered event. Just before the June anomaly, a small number of negative tweets were being boosted by a very high number of likes and retweets while a small number of positive retweets were assaulted with a high number of replies (Section 4.4.4.5). The share of negative tweets was also growing day by day prior to the June anomaly (Section 4.4.4.2). Everything changed with the June anomaly. Not only did the June anomaly increase the proportion of positive tweets, these tweets were also of a good quality, i.e. less likely to be severely disagreeable (Section 4.4.4.5). In contrast, fewer negative tweets were found and these tweets are of low quality. Mean stances for the *encourage*, *authority-non-medic*, and *authority-medic* themes reversed their declines. The proportion of *encourage*-themed tweets, which are strongly associated with a positive stance, increased while the contentiousness and disagreeableness for *encourage* tweets decreased. Proportions for *appraisal-criticize* and *indiv-lib-collective*, which have more equal associations and thus can be considered as more “argumentative”, were reduced. In essence, users are more likely to encounter pro-mask tweets and less likely to be aware of dissenting opinions or even the existence of dissenting opinions.

4.6 Discussion

Whether our argument that the June anomaly was an act of censorship has been sufficiently persuasive or not, we believe that any study involving Twitter data should dispense with the unspoken assumption that Twitter is a neutral platform that is only at the mercy of malicious outside actors. Internal actors matter. Twitter’s moderation exerting an influence over the evolution of the conversations should be accounted for. From attaching warning labels to content shared mainly by Republicans to banning tweets critical of the Indian government’s COVID-19 response, Twitter has clearly never aimed to have the fairest representation of opinions. Results obtained from Twitter data is not just unrepresentative of the general public due to Twitter’s demographics, it is also potentially unrepresentative of Twitter’s own users as certain opinions may be suppressed.

Aside from identifying potential censorship, we also believe that the results of our analyses can offer clues as to how one can best affect change in the overall stance of a contentious issue on Twitter, at least within the narrow niche of mask-wearing for the early months of a coronavirus pandemic to a US-centric, if not necessarily American, audience. These clues can be useful because a lot of policies cannot be unilaterally imposed top-down in liberal democracies or at least in societies that wish to maintain a facade of individual liberty taking precedence over every other priority — the public has to voluntarily accept them. Take the COVID-19 pandemic for example. Successful management of the pandemic requires managing the spread of undesirable information that galvanizes resistance against intervention methods such as vaccinations and universal masking.

As to the question of ethicality of the suggestions we offer here, it is beyond the scope of our work. The debate of whether lofty libertarian ideals encapsulated by quotes such as “I disapprove of what you say, but I will defend to the death your right to say it” and “those who would give up essential Liberty, to purchase a little temporary Safety, deserve neither Liberty nor Safety”, or if pragmatic sacrifices need to be made for the greater good in exceptional circumstances through overt censorship or more covert ones such as tarring and feathering dissent as disinformation and misinformation, is a debate that would not be resolved in this paper.

From the perspective of people who have direct or indirect control over social media platforms such as the platform owners themselves or the government under whose auspices and jurisdiction the platform operates, an effective short-term tactic for managing conversations that are trending in an undesirable direction is to minimize all discussion, even if it meant temporarily silencing supportive voices (during the June anomaly, the absolute tweet count for positive tweets went down). While the Streisand effect does not apply within the

platform due to having control over what messages are visible on the platform, people may move to a viable alternative social media platform (e.g., the platform has an audience of comparable size) to escape censorship. If no viable alternative exist, users may resort to circumventing moderation, e.g., discussing a topic in an oblique manner to escape word filters, provided that the users are aware that discussion is being suppressed in the first place.

From the perspective of those without moderation powers over a platform, efforts to influence the prevailing stance in a conversation or on the entire platform is better done by focusing all efforts on a few conversations with larger audiences than diluting those efforts over many conversations. This is based on our comparison of two events where a large number of users began participating in mask-wearing conversations, one in March and another in April, in which the one where the new users congregated on a few conversations altered the mean stance while the other where the new users are dispersed did not have a visible impact. The threat of centralized attention may have led to the June anomaly; in the days leading up to the June anomaly, a small number of positive and *appraisal-criticize* tweets were receiving a very large number of replies, based on the high means and high standard deviations at the time. If there are no existing conversations with a large number of participants to hijack, one can attempt to build an audience by politicizing an issue. This is based on the observation that the conversation with the largest number of participants in our dataset being more about Trump and Biden than mask-wearing itself. An influence operation can also involve supplanting a narrative with a counter-narrative that has the same themes. This insight was gleaned from observing the “store employees not being allowed to wear mask” narrative shift to “stores making customers wear mask is oppression” with the resultant drop in mean stance for those themes.

4.7 Future Work

While American influence is far-reaching and can impact lockdown protests in Germany¹⁴ and anti-mask movements in Quebec¹⁵, the findings in this paper is still based upon a primarily English dataset that represents US-centric viewpoints. The frequency of tweets with anti-mask stance following a Pacific Daylight Time diurnal pattern and the predominance of US politicians and celebrities as conversation topics demonstrate the limitations facing any attempts at generalizing the findings from this paper. The social media platform we

¹⁴<https://slate.com/technology/2020/09/qanon-europe-germany-lockdown-protests.html>

¹⁵<https://www.cbc.ca/news/canada/montreal/quebec-anti-mask-movement-qanon-covid-19-1.5737040>

have chosen, Twitter, which has moderation policies as well other platform-specific features that encourages its users to interact with it in a certain manner (e.g., ratioing tweets) also limits the generalizability of our findings.

Wearing a mask to combat an airborne pandemic is novel in that until the COVID-19 pandemic, most people in the Anglosphere have never had to grapple with adhering to such a policy. Therefore, people might be more susceptible to being dislodged from their present stance on masking. Whether influence can be as easily gained or lost through social media manipulation for topics where the proponents and detractors are more entrenched, e.g., vaccine hesitancy, remains to be studied.

Due to all of the issues we have listed above, we believe that future work investigating the same topic of mask-wearing conversations could explore other languages, countries, and social media platforms so that we can determine if commonalities exist among these different contexts.

4.8 Conclusion

Studying how influence is won or lost on the contentious healthcare issue of mask-wearing on a widely-used social media platform, Twitter, provided helpful clues on designing information interventions that can help interested parties in more effectively controlling the narrative of such issues online. Stance and theme labels on tweets helped identify inflection points where opinions on the issue changed. Used in conjunction with Twitter's metrics, the labels also provided insights on the factors driving the evolution of the mask-wearing discourse; concentrating attention and breaking concentrated attention appears to be the optimal strategy for propagating and halting the propagation of stances on Twitter.

Using stances and themes to probe a certain anomalous event in our dataset has also highlighted the limitations of using Twitter data as a way to accurately poll public opinion on controversial issues and the need to discard the assumption that the platform allows all opinions to be represented equally.

Our hope is that our findings here will help in crafting new information manipulation strategies that can be tested for generalizability on other issues, from more closely related topics and social media platforms such as propagandizing the wearing of masks for seasonal flus on Twitter to more distant ones such as broadening climate change acceptance on Facebook.

Chapter 5

Carthago Delenda Est: Co-opetitive Indirect Information Diffusion Model for Influence Operations on Online Social Media

“Carthago delenda est” encapsulates informational saturation attacks in this chapter as the phrase showcases the power of repetition in enforcing *censorship* — both in the modern sense and the classical Roman sense. For readers unfamiliar with the phrase [255]:

Delenda est Carthago (‘Carthage Must be Destroyed!’) may be the first recorded incitement to genocide. These were the words of Marcus Porcius Cato, the Censor. Plutarch tells us that Cato’s call ended his every speech in the Roman Senate, ‘on any matter whatsoever’, from 153 BC to his death aged 85 in 149. Scipio Nasica — son-in-law of Scipio Africanus, conqueror of Hannibal in the Second Punic War (218-202 BC) — would always reply: ‘Carthage should be allowed to exist’. But such challengers were silenced. Rome decided on war ‘long before’ it launched the Third Punic War just prior to Cato’s death. One of his last speeches in the Senate, before a Carthaginian delegation in 149, was critical . . . The Carthaginian delegates were accorded no right of reply. Rome soon began a three-year siege of the world’s wealthiest city. Of a population of 2–400,000, at least 150,000 Carthaginians perished.

In this chapter, we introduce an agent-based information diffusion model that features bots/propagandists engaging in saturation attacks against a naive human populace to sway their opinions. This model’s design draws upon the insights gained from the study on the

evolution of Twitter mask-wearing discourse in Chapter 4. Differing from deduplication (Chapter 2) and distinguishing between fake news and satire (Chapter 3), defense against saturation attacks shift from reactive to proactive, with saturation attacks used to defend against saturation attacks. While the previous chapter (Chapter 4) occasionally provided illustrative examples of textual information taken from real-world data, information is exclusively treated as abstractions decomposable into stances and themes in the diffusion model introduced in this chapter instead of concrete units that we directly engage with.

5.1 Introduction

Artificial intelligence (AI), alternatively known as machine learning, has advanced on all fronts — behavioral, textual, audio, visual, and video — to the point where comprehensive and convincing digital personas can be crafted from mere lines of code (Section 5.2.2.5). As a result, the detection of information operations (defined in Section 5.2.2.1) as a method for combating the spread of undesirable narratives by adversarial agents on online social networks, while still important, becomes increasingly non-viable as the predominant means of achieving information security. A 2021 review of research papers on countering influence operations that went as far back as 1972 noted that all 223 studies they found were on user-focused countermeasures (users being consumers of disinformation) [114]. The authors concluded that “[b]eyond fact-checking, the research base is very thin”, and called for increased focus on key areas such as “studying countermeasures that target the creators of disinformation content in addition to studying consumer-facing policies”. In response to this call for action, we propose complementing existing detect-and-correct methods that conduct struggles against undesirable information on a tactical, low-level per-message or per-user basis with a method that facilitates waging war against disinformation from a high-level strategical perspective, where users and messages are seen as a part of a synergistic whole meant to achieve a combined arms outcome.

To that end, we designed *Diluvion*, an agent-based model for simulating how information operations can perturb opinion evolution in an online social media platform. The model relies on parameters derived from publicly available data on a social media platform to forecast the outcomes of info ops on similar platforms. As such, it can serve as a tool to be used by planners in crafting campaigns to attack disinformation or defend against it without needing to first test the interventions in the real world — tests that can have unsavory ethical and even geopolitical implications [185, 193]. Like many other information diffusion models, *Diluvion* is inspired by epidemiology but updated to consider new insights on transmission dynamics from the COVID-19 pandemic (Section 5.2.1.2).

But, unlike many models, *Diluvision* is developed with info ops and bots in mind, meaning that unconventional routes for expanding influence are accounted for and that optimistic assumptions for information diffusion in other arguably marginally more benign contexts are discarded, for example, the ability to select seed nodes to maximize propagation in product advertising [281, 360].

While models are often designed to adhere closely to the real world, no model is meant to be a perfect recreation of the real world; design choices reflect aspects that the modelers wish to emphasize. In our case, *Diluvision* brings into prominence the role played by indirect information, namely social cues [152] — engagement metrics and the perception of mass adoption — in opinion propagation. Repetition is the key to the operation of our model. Much like Cato’s brilliant tactic of incessantly playing up Carthage’s threat to Rome [255] and calling for Carthage’s destruction in the Roman Senate led to the genocide [255] of Carthage, much like repeating a lie can let it take on the mantle of truth [158, 203], so does the persistent saturation of an opinion in the media landscape can drive its adoption among the populace [169, 209, 210, 310, 350, 374, 534].

The contributions of this paper are:

- An agent-based co-opetitive information diffusion model for information operations on social media that is novel in its incorporation of concepts not considered by existing models: an infectious neutral stance; themes that must piggyback on stances to infect users and have an enhancing or diminishing effect on the stance’s infectivity; non-social tie-based transmission of information; parity in the treatment of bots backing different stances; perception of mass support inferred through engagement metrics such as the number of retweets influencing stance adoption; and agents with heterogeneous roles and activity levels populating the simulation.
- In-depth rationale behind the design/modeling choices, demonstrating the empirical evidence that our model is predicated upon and consequently the real-world applicability of our model.
- Initial “validation” of the model by drawing simulation and graph structure generation parameter values from distributions found in real-world data and further validation that succeeded in reproducing via simulations the distributions from a different empirical dataset.
- Simulations that are in part ablation studies and in part experimentation with info ops tactics. The maneuvers examined include bot-backed minority stances facing off against a non-bot-backed majority stance, polarization/depolarization, and max-

imizing theme diffusion. The emergent outcomes from implicit cooperation and competition found in these simulations are examined and discussed in the context of their implications for decentralized info ops.

5.2 Background

As our work occurs at the intersection of different domains, we split our discussion of relevant works into two major sections: information diffusion and influence maximization (Section 5.2.1), and information operations (Section 5.2.2).

5.2.1 Information diffusion models

The study of information diffusion — the process by which information spreads from one entity to another — has a long history. In the particular case of diffusion on online social networks, the history is sufficiently extensive to allow for a survey [195] of existing work to be published in 2013. The closely related domains of influence maximization and consensus formation have similarly rich histories, which can be seen in a 2021 survey [35]. Influence maximization (IM) studies the problem of locating influential nodes that maximize the spread of information to the entire network. Consensus formation is concerned with determining the manner in which a group of agents can all arrive at the same opinion. For an in-depth discussion of these topics, we refer the readers to the aforementioned surveys. Here, we restrict ourselves to discussing the most prominent and relevant models and the elements from those models that influenced the development of our model, *Diluvision*. The information presented is also meant to help situate *Diluvision* among the constellation of predecessor models.

5.2.1.1 Information

In the space of diffusion/propagation models, information goes by many different names, including information [292], influence [35], rumor [264], opinion [393], memes [496], and news [385], depending on the context. While we will use these terms interchangeably throughout this paper, we favor the term “opinion” due to the fact that the major component of the information being spread in networks within our model, the stance, is polarized and homophilic. Together with non-polarizable themes, they form a unit of information in our model (Section 5.3.1). That is not to say that other forms of information are not present in the model. Engagement metrics, for instance, are exposed to agents and can influence their

decision to adopt a piece of information.

We are aware of works that have studied how real news and fake news spread differently on social media [483, 527]. The information of interest in this paper, i.e., the information being disseminated, is not generally regarded as “fact” (e.g., quantum mechanics and general relativity applied within their respective restricted contexts and not outside) but “opinion” (e.g., the dispute over the rarity of scale-free networks in the real-world [218]), which does not have a clear delineation between true and false. In our model, “disinformation” is simply a term of convenience for undesirable information; messages are indistinguishable in terms of veracity, so veracity cannot influence a message’s spread. Only stances and themes have influence.

5.2.1.2 Uncontested information diffusion

Based on the typology introduced in [195], information diffusion models can be separated into two categories: explanatory and predictive. Explanatory models aim to “infer the underlying spreading cascade, given a complete activation sequence.” *Diluvision* belongs to the other category, predictive models, that aim to predict “how a specific diffusion process would unfold in a given network, from temporal and/or spatial points of view by learning from past diffusion traces.” Predictive models are further divided into graph- and non-graph-based approaches. Graph-based approaches assume that information spreads along the edges in the graph. Non-graph-based approaches ignore network topology and model information diffusion as a process represented by a system of differential equations, in emulation of similar equations in epidemiology [250, 251].

The two most well-known graph-based models are the independent cascade (IC) [188] and linear threshold (LT) [191] models, and they serve as the base models for most influence maximization research. IC models diffusion as a chain reaction in which a node that has adopted a piece of information, designated as being activated, attempts to activate inactive neighboring nodes in the graph with a certain probability. There are no retries and an activated node remains activated forever. In LT, an inactive node becomes activated if the sum of influence from activated neighboring nodes exceeds the node’s threshold. Influence takes the form of weighted directed edges between nodes and the threshold is generally drawn from a uniform random distribution. An activated node remains forever activated in LT as well. IC can be seen as sender-centric in contrast to the receiver-centric LT.

In non-graph-based models, the model names usually take after the sequence of state transitions permissible within the model, where the states are susceptible (S), infected (I), and recovered (R), which are based on the earliest epidemiological model, SIR [250]. The first explicit link between information and disease was established by Daley and

Kendall, who were inspired by the SIR model to create the Daley–Kendall (DK) model for rumor propagation [118]. Notable variants of DK include the Maki–Thompson model [314]. Rumor diffusion models have states such as ignorant (I, susceptible), spreader (S, infected), and stifler (S/R, recovered) that mirror their epidemiological counterparts, although these specific terms were not used in the initial work by Daley and Kendall. Later reviews [214, 358] showed the cross-pollination of ideas between rumor-specific and biological epidemiological models. Subsequent studies on epidemiological diffusion models also innovated upon the transition sequence, e.g., SIRS [430] or added additional states, e.g., SIHR [526] rumor spreading model where H stands for hibernator and SHIR [292] model for competitive information diffusion on online social media where H stands for hesitated. DK descendants innovated in a similar fashion, e.g., ISCR [348] which added a counter-spreader (C) state.

In *Diluvion*, network topology matters, as in IC and LT, and information generally travels based on the agents’ connections. The influence of epidemiological models, SIR, SIS, etc., are in inclusion of non-graph edge routes (circumventing topology) for information diffusion to our model, as well as in the agents’ non-durable immunity (SIS). However, agents are never in an uninfected state within our model, as they only ever transition between the three mutually exclusive infected states (stances), negative, neutral, and positive, with infection by non-exclusive themes providing an additional differentiating mechanism akin to viral variants (Section 5.3.1).

The SARS-CoV-2 pandemic has revealed the shortcomings of some graph-based models. In information diffusion, the common assumption is that a link/edge exists between two people *only* if they have social ties (e.g., friends, family, followers, and colleagues). However, in epidemiology, such an assumption is not always made because ties can be formed with complete strangers in many situations. A passerby breathing, coughing, or sneezing in a crowded cafe spreads aerosolizable infectious disease almost as effectively as staying under the same roof with a diseased family member [244]. On social media, many posts are publicly viewable and lively discussions can occur in the replies section of those posts, analagous to a forum or a cafe, and a user may be exposed to the opinions of strangers, i.e., people who are neither followers nor followees, who replied to the post of the user’s followee (this idea will be explored in Section 5.3.3.2). Another issue is the strong assumption inherent in designating the recovered/stifler state as an agent’s end state in an information epidemic. SARS-CoV-2 immunity is transient. It is not permanent. Reinfection by the same or different variants is a real possibility [390]. Non-permanent resistance to contagions that increases or decreases depending on the infection type is a mechanic implemented in *Diluvion* to offer an alternative perspective to the strong assumptions of

permanent immunity in the existing literature (Section 5.3.4).

5.2.1.3 Competitive information diffusion

Classical information diffusion does not account for information possessing polarity and existing in a contested space. “Vaccines are safe” and “vaccines are dangerous” are contradictory opinions which a single entity is unlikely to hold simultaneously. There are actors, such as healthcare authorities, with a vested interest in not just rapidly disseminating the opinion that vaccines are safe and having people adopt that opinion but also stemming the adoption of the opposing opinion and having the opposing opinion be dislodged from those who have adopted it. Competitive information diffusion models, also known as opinion formation models, were developed to simulate such scenarios. In biology, competition can be thought of as cross-immunity contagions [492], where developing immunity for one decreases the likelihood of being infected by the other.

Two broad categories of competition exist based on the type of information being diffused: continuous or discrete. Certain consensus formation mechanisms (e.g., taking the mean or median) are not possible for an information type (e.g., selecting a material for constructing an object, although the problem can be reformulated such that selection is based on continuous-valued attributes such as tensile strength). Elements of continuous and discrete models can be made to work together, as demonstrated by the Biased Voter model [120] that combines DeGroot and voter models.

For continuous-valued information, the DeGroot/French–DeGroot [133, 167] model is widely used for competitive diffusion. In the model, every incoming edge for a node/agent has a weight denoting another agent’s influence (a weight of zero is equivalent to not having an edge). There is also a weighted self-loop for the extent to which an agent values its own opinion. At every time-step, an agent updates its opinion by taking the weighted average of all neighboring opinions, including its own. This process is repeated until the values (opinions) are stable. In a strongly connected graph, perfect consensus (all agents share the same opinion) is guaranteed. The Bounded Confidence model, alternatively known as the Hegselmann–Krause model [208], also updates an agent’s opinion by averaging others’ opinions, except that instead of the opinions of all neighbors, the opinions are those who share similar opinions, with similarity being bounded by some arbitrary value, hence the model’s name. Clustering (a form of polarization), instead of perfect consensus, is the likely outcome of the Bounded Confidence model. A review of continuous-valued models, which includes DeGroot, Weighted-Median, Bounded Confidence, and Quantum Game, can be found in [136].

For discrete opinions, one popular model is the voter model, independently developed

by Clifford and Sudbury [110] and by Holley and Liggett [216]. At every step, the model first chooses an agent and its neighbor, both at random. The agent then adopts the opinion of its neighbor. Another popular model is the social influence model by political scientist Axelrod, designed to explain how global polarization arises due to local homophily. In this model, each agent is defined by its cultural traits, which is a fixed-length string of discrete numbers. Similar to the voter model, the Axelrod model randomly chooses an agent and its neighbor at every step. Then, with a probability equal to their cultural similarity, the agent will update one of its traits that differs from its neighbor's. The Axelrod model can be considered an extension of the voter model.

One natural extension to voter models is to have an agent's opinion updates be guided by the majority instead of just any random neighbor. The majority rule model by Chen and Redner selects at random a flippable spin (a changeable binary opinion) and have the spin adopt the majority state of its interaction neighborhood [95]. Fu and Wang incorporated minority avoidance into the majority rule model by having agents occasionally sever ties with those holding a minority opinion [168]. Other information diffusion models studied a different form of minority avoidance [409, 459], specifically the "spiral of silence" [363], which theorizes that people are reluctant to disagree with the majority, hence restricting themselves from expressing a non-conforming opinion, ultimately leading to the continued domination of the majority voice. The majority voice can be illusory, created by a vocal minority.

There are many ways in which competition can be modeled using epidemiological models. For two competing information contagions, a winner-takes-all situation emerges, where the stronger contagion dominates completely instead of just gaining a majority across all graph types, given the conditions of no permanent immunity (SIS-like, where S is susceptible and I is infected), mutually exclusive infections, and nodes that are homogeneous in their resistance towards both contagions [388]. Non-graph-based epidemiological models can also account for competition. Two of the proposed epidemiological models in [348] accounted for the presence of competing discrete information differently: the ISCR model features an additional class of user, counter-spreaders (C), in addition to the classical ignorant (I), spreader (S), and recovered (R), while the ISSRR model divides the spreader and recovered classes based on the which side of the bipolar information they were infected by or recovered from.

Extensions to IC and LC (Section 5.2.1.2), which are discrete models, allow them to model competitive influence maximization. The simplest modification is to start the diffusion of two different pieces of information from two different origin/seed nodes. The different IC and LT derivatives distinguish themselves based on the restrictions or conditions

that the competing factions are subject to: selecting the most influential starting nodes for one faction given that the competing faction's starting nodes are known under IC [88]; best response and first mover strategies under IC [58]; proving that being the first mover is not always an advantage under a special case of IC [264]; different forms of competitive diffusion under LT, e.g., different tie-breaking rules [67]; and two competing cascades on a signed network under the IC model [448].

The limitations and unrealistic assumptions of IC and LT have been recognized, with [28] highlighting three misspecifications of classical IC and LT: not modeling assortativity of influence, ignoring non-uniformity of the joint distribution of influence and susceptibility (due to peer effects, homophily, and social influence), and heterogeneity in influence and susceptibility. However, [28] has misspecifications of its own. In the problem of IM, [28] defaults to seed selection: a faction (representing a stance) has the opportunity to pick and choose seed nodes from which to begin its diffusion. However, this seed selection assumption that non-competitive and competitive IM operate under is alien to info ops (Section 5.2.2). Influencers on social media can be paid to promote a stance, but not all can be bought off. Info ops are also often conducted in a decentralized manner to make them resilient against the decapitation strike tactic of silencing/compromising influential nodes (Section 5.2.2.4), so info ops planners are unlikely to even consider the problem of seed selection.

Traces of LT's threshold mechanism combined with DeGroot's weighted influence (despite our model being designed for discrete instead of continuous opinion), and the voter/majority rule model are present in our model, grounding it in the extant competitive diffusion literature. A threshold (resistance) must be crossed for opinion adoption/infection events. Unlike classical majority rule models, in which an agent's adoption of the majority opinion is deterministic (the neighborhood's majority opinion will definitely be adopted), ours is probabilistic, with each opinion weighted by perceived popularity, as measured through the proxies of engagement metrics. While agents do not have weighted self-loops as in DeGroot, they are influenced by how others judge the opinions that they express (Section 5.3.3.2). The mechanics make it so that human agents in *Diluvion* are more inclined to convert to a stance that is perceived to be held by the majority in the quasi-global neighborhood (Section 5.3.3.2). Adopting an unpopular opinion is possible, though unlikely. An agent's discrimination against minority opinions is therefore encoded in the probability distribution. This form of minority avoidance in our model, *Diluvion*, despite sharing conceptual roots with those encountered in the literature, is considerably different in its technical implementation.

The design of our mechanics for themes can be considered a special restricted case of

Axelrod's cultural traits. The example of traits given in Axelrod's original paper allows for 10^5 possible states since there are five positions filled with numbers ranging from zero to nine. The number of positions in *Diluvision* is variable, from one to an arbitrary maximum, and a theme can fill a position only once. If the maximum was five and there were 12 themes, $\sum_{k=1}^5 \frac{12!}{k!(12-k)!} = 1585$. Peer effects, homophily, and social influence all affect an agent's decisions in *Diluvision*, heightening realism (see Section 5.3.3.2). One manifestation of homophily is in opinions sharing similar stance and themes as the agent itself possessing greater weights during an infection event, and this homophily is comparable to Axelrod's and Hegselmann-Krause's homophily, though less extreme.

5.2.1.4 Cooperative and competitive information diffusion

Cooperation can also be an outcome of the interactions between multiple information contagions present in a system [492]. Immunological dysfunctions caused by the infections of viruses such as HIV [109], Ebola virus [500], and SARS-CoV-2 [384] that make the infected more susceptible to other viruses are examples of cooperative biological contagions. Relative to competitive diffusion, the many forms of cooperative information diffusion have been less frequently studied.

Cooperation can occur at the information/contagion level. The diffusion model in [353] learns from empirical data the affinity between different contagions (URLs), finding that news stories of the same event boost the adoption of each other's URLs, implying cooperation. The spread of a highly infectious news story's URL suppresses the spread of URLs from other unrelated, less infectious news stories, implying competition. Different models [91, 523] have recreated the synergistic (e.g., complimentary products such as iPhone and Apple Watch) and antagonistic relationships between contagions propagating in heterogeneous and homogeneous multilayer networks where simultaneous infection by multiple contagions (coinfection) is possible.

Cooperation can occur at the agent level. The implicit form of such cooperation is common in the study of multi-agent systems in the context of robotics (vehicular information diffusion and consensus formation [3, 414], drone swarms [30, 493], and search-and-rescue robots [24]). Ideas found in (not necessarily originating in) robotics can be adapted to information diffusion, e.g., swarm intelligence algorithms for influence maximization under IC and LT models [366, 441].

Finally, the cooperative effect can be found in the links/edges connecting nodes in a network; a node u being infected by a contagion through another node v makes u more susceptible to future infection by a different contagion via v [489], reminiscent of the effects of trust and affection in social relations.

The implicit cooperation for *Diluvision* happens at the agent-level, where all bots targeting the most vulnerable for stance propagation (5.3.5) and when bots' independent actions unknowingly contribute towards a common objective (Section 5.6.2). At the information level, stances and themes induce homophilic conversions, and engagement metrics induce bandwagoning (Sections 5.3.1 and 5.3.3.2).

5.2.1.5 Agent-based models

Agent-based models (ABMs) aim to infer macro-level outcomes from micro-level interactions and can be viewed as a further reduction of abstraction compared to models based on interconnected homogeneous nodes and models constructed only from differential equations. Nothing stipulates that agent-based models have to incorporate a network topology in modeling diffusion (e.g., agents can be placed on a grid and the diffusion can be governed by spatial proximity), but in practice, agent-based models of information diffusion on social media invariably use network topology, as networks are an effective approximation of social organization in human communities, online and offline. Being freed from the constraints of stringently adhering to the precepts of classical IC and LT, and their many extensions, enables more customized and accurate models to be built based on empirical data in sociological and psychological studies, albeit costlier to run computationally. An example is the ABM for the highly specific context of online anti-vaccination movements [445] which contained three different agent types: doctors (similar but not equivalent to external authority/mass media in other models), patients (humans), and initiators (stubborns/bots).

A comprehensive review of the various design aspects of modeling information diffusion in social networks as multi-agent systems (MAS) can be found in [234]. This review focuses on the trade-off between the simplistic assumptions found in classical MAS and the complexity of real-world social networks. Thinking in terms of complexity versus abstraction can help establish boundaries between classical and agent-based models of diffusion. This demarcation is used by some ABMs, e.g., [409]. Thus, while agency is arguably present in some models discussed earlier due to individualization/heterogeneity of nodes, e.g., DeGroot [133, 167], those models never fully embraced the possibilities of agency and the resulting emergent behavior (e.g., John Conway's Game of Life) in the interest of keeping the models analytically tractable, which is why we do not discuss them in this section.

Diluvision is an agent-based model. It borrows many concepts from existing models that are relevant to accurately simulating the impact of info ops on information diffusion on social media.

One borrowed concept is that agents have a set amount of energy, with every action depleting their energy reserves. In [241], agents have a fixed energy level and interactions

do not cost energy; however, after some arbitrary point in time, every interaction decreases the agent’s energy level. This setup enables the study of the correlation between agent energy and rumor propagation intensity. In [370], the simulations were populated by different classes of agents, each class with its own range for how frequently its members tweet and retweet. Our implementation of the energy level is based on empirical data from Twitter that separated users into three tiers (Sections 5.3.4 and 5.3.8.1). Our agents’ actions are additionally restricted by a diurnal cycle (Section 5.3.6).

Agents are also commonly constrained by having a memory and attention (also uncommonly, and perhaps confusingly, called vigor [279]) limit placed upon them, e.g., [57, 279, 280, 485, 496], reflecting the non-omniscient nature of humans. We discuss this mechanic in greater detail in Section 5.3.2.

Linked to the concept of memory is the concept of repetition driving information adoption. In the agent-based SIR rumor diffusion model introduced in [225], agents can receive the same information multiple times, with agents likelier to adopt information that has been seen more frequently, contributing to the spread of the information as agents eventually share the information that they have adopted. In the information entropy-based rumor spreading model in [485], a rumor’s information salience is measured by the frequency of the rumor’s occurrence in the agent’s memory, with greater frequency linked to higher salience. The most salient rumor is selected for transmission during the spreading phase of the model, and this selection mechanism constitutes the conformity effect put into practice [485]. In an ABM for studying astroturfing by bots under the spiral of silence framework, conformity takes the form of an opinion climate, which is composed of the opinions expressed by neighboring agents with silent agents excluded [409]. Another model [280] of conformity by has an agent slowly revising its opinion to become more similar to its neighbors’. Yet another model, AMID, features users who are likely to adopt opinions supported by the majority of adjacent neighbors in [279]. Additionally, under AMID, trust between agents was modeled as the frequency with which a truster shares the trustee’s messages, with users likelier to adopt messages from trusted sources [279].

Repetition’s influence similarly suffuses every part of *Diluvsn*. A stance that occurs more frequently in an agent’s memory has a greater chance of being adopted. But, we introduced additional complications into the mechanism. Conformity effects are extended to account for the engagement metrics prominently displayed alongside every message on social media, which serve as cues on how others perceive the credibility of the message (Section 5.3.3.2). Conformity effects are also balanced against homophily for stances and themes, which is based on the real-world tendency for people to prefer opinions that are similar to their own.

A divide exists among ABMs on the matter of opinions being exclusionary or capable of co-existing with others. Exclusionary opinions, as conceptualized in [222], permit a user to only adopt one discrete opinion at any one time. Models in which opinions are modeled as a point along a continuous spectrum, as is common in counter-disinformation ABMs, e.g., [32, 57, 225], naturally feature exclusionary opinions as opinions cannot simultaneously occupy two positions along the spectrum. Studies using the spiral of silence framework, e.g., [409, 459], require polarized/exclusionary opinions by design; users would not be hesitant to voice opinions that ran counter to the perceived majority opinion climate if an oppositional relationship did not exist. For non-exclusionary opinions, there is also a myriad of conceptualizations. In SMSim, developed for studying non-polarized information diffusion [126], bit vectors denote a message's association with the set of possible topics in the simulation. For the rumor spreading model introduced in [485], infection by an opinion means that the opinion goes into the agent's memory bank. There is no incompatibility mechanism that forces a stored opinion to be ejected from memory; removal is only due to insufficient space in the bank. And with some models, the distinction between exclusionary and non-exclusionary is less clear, e.g., [353] where agents can only adopt one contagion but the agent's infection history continues to influence future contagion adoption. In the AMID model [279], each message is associated with all possible topics in the simulation with varying strengths (fuzzy sets) and there is an additional binary number for each topic denoting the message's positive or negative opinion towards the topic, a formulation reminiscent of *Diluvion*'s stance and themes (Section 5.3.1). Users in AMID are capable of caring about multiple topics simultaneously, but the opinion (polarity) for each topic is an exclusionary binary value.

On the matter of co-existing opinions, *Diluvion* straddles the middle as there are exclusionary and non-exclusionary components of an opinion within *Diluvion*, corresponding to polarized stances and non-polarizable themes respectively (Section 5.3.1).

A common feature of ABMs is the segregation of the agents into distinct classes. This practice is intuitive in studies investigating the impact of bots on social media, e.g., [32, 57, 370, 409], as bots have to behave differently than the rest of the users to distinguish themselves and for their influence to be measurable. In [370], both humans and bots were further subdivided into additional categories, e.g., amplifier and commentator, based on extant research on social media user behavior. Some works [32, 57] divided agents based on stance, with "truth" propagandists, neutral/uninformed people, and "disinformation" peddlers exhibiting different behavioral patterns; bots are associated solely with the undesirable stance, disinformation. While most models keep an agent's role static throughout the simulation, some allow agents to radically alter their behavior upon meeting certain

conditions, e.g., radicalization turning passive users into active disseminators [32, 445].

Under *Diluvion*, three main attributes define an agent: its activity level, its action-role, and whether the agent is a human or a bot. There are three activity levels, five action-roles (15 unique action distributions as the distributions for the same role differ based on activity level), and a binary bot/human membership. As such, a total of 30 possible classes of users exist in *Diluvion*. Agents do not change roles. Our conception of bots (Section 5.3.5) hews closer to [370, 409] in that their *behavior* is stance-agnostic for the reasons specified in Section 5.2.2.5, even though [370, 409] also only associate bots with malicious behavior.

Models found in the literature that most closely resemble our *Diluvion* are those developed for a purpose similar to ours: the ones by Averza et al., Beskow and Carley which dealt with information operations involving bots shaping discourse on online social media platforms [32, 57]. Unlike [409] which also studied the impact of bots on online discourse, [32, 57] did not integrate the spiral of silence into their models. And unlike [370] which was interested in the descriptive statistics (e.g., tweets and retweets per day) that a mixed population of bots and humans would generate, [32, 57] are concerned with competitive diffusion of three stances, just like our study. Although to be accurate, our model allows all three stances to be propagated, but in [32, 57], only the two stances which are antithetical to each other can be propagated while the neutral stance is a static default state. The crucial differences between our work and [32, 57] are the stance-agnosticism of our bots (Section 5.3.5), neutrality being an infectious stance in our model as previously mentioned (Section 5.3.1), the inclusion of indirect influence mechanics (engagement metrics and non-edge information transmission, Section 5.3.3.2), and heterogeneous activity levels and roles for agents (Section 5.3.8.1).

5.2.2 Information operations

5.2.2.1 Terminology

Within the existing literature, information operations (info ops), influence operations, and psychological operations have been used interchangeably to describe “activities that target information bases of an adversary, protect one’s own information assets and make efficient use of one’s own information”[304] even though each term has a specific meaning, e.g., psychological operations are originally defined as falling under the umbrella of info ops [150]. In this paper, we adhere to the loose usage of these terms as other researchers have, e.g., authors of [114, 217] switching between the terms within the same document; Cambridge Analytica varyingly described as engaging in psychological operations [38], information operations, [143], and influence operations [504].

Although info ops is “a military defined concept and doctrine” originating from the post-Cold War US and NATO military establishments, it “interacts largely with society as the majority of the targets either attacked or to be defended by Information Operations means are not military but civil targets” [304]. As pervasive computing has become a reality in a world where interconnected mobile computing devices are increasingly commonplace among civilians, there has never been a richer, more fertile ground for info ops [85], especially when the rise of social media normalizes rapid information sharing.

5.2.2.2 Social media

Social media has always been a tool for information warfare. As early as 2003 [213] / 2006 [49], the cyberdissident diplomacy (CDD) initiative from the US State Department reached out to “tech-savvy youth in the sixteen-to-thirty-five age range with the aim of training them in particular forms of cyberdissidence and online campaigning”, leading “young Muslims” to “serve as the proxies to the US” in a “war of ideas”, an “ideological war”—influence operations, in other words — “to maintain its global hegemony” [212, 213]. As the provocative title of one journal paper neatly summarized, *“Facebook to Mobilize, Twitter to Coordinate Protests, and YouTube to Tell the World”: New Media, Cyberactivism, and the Arab Spring* [27]. In the 2009 protests in Iran, Twitter was deemed so crucial to the protesters that the US State Department intervened in Twitter’s scheduled server upgrade to ensure that Iranian users did not suffer from service interruption [81]. In Russia where the user base of native social media platforms VKontakte and Odnoklassniki dwarfs that of the US-developed Facebook, it is Facebook that mobilized anti-regime forces in 2011 [498]. The clearest example of social media as an instrument of war is the USAID-backed ZunZuneo, which is a Twitter-like service designed to run solely in Cuba through mobile text messaging for the express purpose of covertly inciting regime change, according to the reporting done by the US news agency Associated Press [26, 82].

But as the flames of social unrest fanned in faraway “jungle” [313] countries escaped their confines and started engulfing First World/developed “garden” countries, the dangerous potential of social media was finally recognized by the research community and alarms were raised. In a review of academic publications on countermeasures against influence operations, the authors found a “dramatic increase in the number of studies since 2016” which is “[i]n line with other research trends in the broader field of influence operations and disinformation” [114]; 2016 was the year when Russia was accused of meddling in the US presidential election in favor of Trump via social media campaigns [13]. Threats can be homegrown too: Amazon, Apple, and Google suspended the US social media platform Parler for its role in the 2021 Capitol Hill riot [14]. The panic in the research community

has culminated in the birth of a new field, social cybersecurity, which seeks to “characterize, understand, and forecast cyber-mediated changes in human behavior and in social, cultural, and political outcomes” as one of its objectives [87].

Diluvsn is a tool that seeks to aid in the fulfillment of the aforementioned objective by enabling the characterization of Twitter-like services and the forecast of info ops on such services.

5.2.2.3 Twitter-like platforms

Likely because of the ease of collecting data from Twitter, many researchers have focused on modeling information diffusion on Twitter such as [32, 57, 120, 176, 496]. This is one reason the simulation environment within *Diluvsn* models itself after Twitter. A greater motivating factor is the profusion of Twitter-like social media platforms. Truth Social, Parler, and Gab cater to users chafing at the politically restrictive policies of Twitter [449]. Mastodon lets users escape the whims of a single central authority by dispersing regulatory power among multiple central authorities [535]. Twitter-like platforms have also emerged to serve different regions and languages, such as China’s Sina Weibo. Even social media platforms that have never sought to emulate Twitter have analogs to certain features of Twitter (which may not have originated from Twitter), e.g., Facebook’s share and share with a comment being similar to Twitter’s retweet and quote-tweet, respectively.

The mushrooming of Twitter alternatives also speaks to a credibility crisis, where users no longer trust that they can express their views free of suppression from a central authority. Decentralization is a natural antidote to this credibility crisis. Our conceptualization of decentralization within *Diluvsn* does not take after Mastodon, since their federated instances still allow moderation power to rest in the hands of the few [535]. We speak of a decentralization where power is much more diffuse — or at least appears so (Section 5.2.2.4).

5.2.2.4 Decentralization and radicalization

The most effective way to shut down conversation is the top-down approach of restricting free speech, as the Ukrainian experience of banning Russian social media VKontakte from 2017 onwards has shown [190]. Yet at the same time, the vast majority of both pro-Ukraine and pro-Russia Ukrainians circumventing the ban to continue accessing VKontakte [190] shows that the desire to escape overt, centralized censorship can transcend political affiliations. In liberal democracies, harsh speech policing measures are not well-tolerated by users, even if private corporations instead of governments are the ones doing the policing

[378], as evinced by the many alternatives to Twitter (Section 5.2.2.3). And if governments instead of corporations enforce censorship, the governments risk being viewed as draconian, tyrannical, totalitarian, authoritarian, etc., by liberty-minded people. A 2019 attitudinal survey of US internet users found that “among political attitudes, opposition to censorship ($\beta = -.231, p < .001$) had a significant negative effect on support for government regulation of platforms” [403]. Hence the rise of a decentralized approach to information warfare: shaping discourse with an invisible hand, unbeknownst to the public.

Decentralization has proven effective on offline social networks. Decentralization allows for radicalization without attribution and accountability. Absent clear chains of command, membership simply involves claiming affiliation. Members can appear to be acting autonomously, when collectively those actions advance the interests of the movement [327]. Should the members’ actions harm the movement, “they can simply disavow the act without disavowing the ideology” [369] in the same vein as black ops and black sites [287]. Without central authority, a movement can resist collapse from decapitation strikes [27, 412, 454]. Ivan Marovic, a leader in the successful Serbian regime change organization, Otpor, which is funded by NED; IRI; USAID; Freedom House; and British Westminster Foundation, explained to National Public Radio’s Bob Garfield that “[l]eaders could have been blackmailed or bribed or even maybe killed. You can’t do that with brands or ideas” [174]. This is similar to how a perpetual struggle can be waged against an idea, e.g., the Global War on Terrorism. Terrorist themselves from white supremacists [90] to al-Qaeda and Hamas [327] also favored decentralized structures for their resilience and “tactical security”.

Success in the offline world has led to attempts to replicate decentralized movements online. The US cyberdissidence program is one such attempt at growing an online decentralized movement (Section 5.2.2.2). Another is CENTCOM’s Operation Earnest Voice (OEV), which empowered “one US serviceman or woman to control up to 10 separate identities based all over the world” so they can create “false consensus in online conversations” [55, 111]. Graphika and Stanford Internet Observatory (SIO) reported in 2022 that their joint investigation with Twitter and Meta uncovered hundreds of accounts engaging in covert pro-Western info ops over the course of five years [192]. An Israeli company, Team Jorge, specializing in info ops that ran multiple campaigns influencing elections around the world [258, 328], created Advanced Impact Media Solutions (AIMS), a one-click system for creating “fake digital persona[s] built to mimic human behavior and avoid detection” which operate “on social media sites, disseminating rumors, harassment, defamation or praise — whatever the client asked for” [54, 330]. Other Israeli info ops companies Archimedes and Percepto also employed fictitious digital personas [129, 329]. White supremacists

[369] and international Islamists [12, 33] took advantage of social media for recruitment [12, 33, 369], to transition to decentralized structures [12, 369], and to conduct info ops of their own [150]. The “IT cells” manipulating Indian elections adopted a functionally leaderless structure, where the bottom layer “creating and trending narratives” maintains “enough distance from the top to give leaders plausible deniability if the digital foot soldiers become too extreme” [412]. Another example is NAFO (North Atlantic Fellas Organization), which Politico described as consisting of “internet culture warriors taking on Russian disinformation and raising funds for Ukraine” by adopting the same tactics as “the Islamic State” and “the American far-right boogaloo movement” [425]. According to one freelance disinformation researcher and self-proclaimed member [100, 101], NAFO is a leaderless movement [99]. The Wall Street Journal describes NAFO as having “no command structure” [339].

Owing to the achievements and prevalence of decentralized info ops, the simulations we chose to show running in *Diluvion* (Section 5.5) were developed with decentralized info ops in mind, with affecting change through uncoordinated coordination. The change agents being unaware of each other (Section 5.3.5) in *Diluvion* is also a deliberate design choice grounded in decentralization — leaderless terrorist/freedom fighter cells [245].

To affect change and appear organic at the same time, a decentralized movement needs a large number of warm bodies. But growing a movement is time consuming. Furthermore, each additional truly independent (human) agents within a movement is an additional security risk. Humans can turn into whistleblowers. Financial trails from hiring propagandists risk being traced back to the mastermind.

CENTCOM’s OEV, Archimedes, Percepto, and Team Jorge’s AIMS already hinted at a solution: autonomous artificial personas (Section 5.2.2.5).

5.2.2.5 Digital personas and bots

Based on all the markers accessible via cyberspace — i.e., excluding anything that requires physical interaction in the real world — bots can be made virtually indistinguishable from live human operatives. This is all thanks to progress in the field of machine learning making it easier than ever to deceive human senses and cognition [256]. On the behavioral and textual front, services such as Character.AI allow users to create chatbots with distinct personalities, such as Elon Musk or Marvel movie characters, and these chatbots are capable of carrying conversations for hours with humans [406]. Chatbots can exert influence over humans; interacting with chatbots has a positive impact on people’s attitude towards vaccines and also their intention to get vaccinated [21]. Bot-run Twitter accounts are perceived as credible sources of information [147]. On the audio front, text-to-speech

audio generation [254] has led to convincing audio clips of Biden and Trump chatting over gaming sessions [407]. On the video front, there are talking head video generation neural networks producing realistic depictions of people saying things that they have never said, e.g., [491]. A variety of techniques exist for transposing one person's face onto the body of another person in a video, e.g., [512]. High resolution videos can be created with only text prompts, e.g., [61]. A navigable 3D scene can be created through neural radiance fields using just a few static images, e.g., [45, 517]. On the image front, text-to-image synthesis networks can hallucinate scenes that never existed in the real world purely by relying on text descriptions, e.g., [531]. For realistic faces, one does not need to steal the identity of a human, as faces of non-existent people can be generated [243].

To tie it all together, AIs given an initial goal from humans can plan to accomplish the goal by giving itself a list of tasks and setting about accomplishing the tasks by itself using various tools (including itself and other AIs) at its disposal [392], currently best exemplified by Auto-GPT. Pretending to be a human employed in a job will not be too difficult, as AIs can produce codes that run [246] and also play at the role of a knowledgeable analyst since vast swaths of programming and data science work can be automated via ChatGPT [205].

While we are not aware of any entity that has made digital personas using a combination of all these bleeding edge AI capabilities¹, there is a high likelihood of such hypothetical human-like bots being created and deployed in service of info ops at present or in the near future. Relying on detection as the sole strategy to safeguard against them might be a dead-end. A collaborative report by the Stanford Internet Observatory, OpenAI, and Georgetown University's Center for Security and Emerging Technology concluded that it was reasonable to assume that "attributing short pieces of [textual] content as AI-generated will remain impossible" [189]. The same report warned that, based on their deliberately limited scope of "generative language models" producing "persuasive text", AI is "likely to significantly impact the future of influence operation" [189]. The days in which bots are simplistic and easily detectable automatons, such as those found by BotHunter to be influencing Singaporean elections to little apparent success [471] or found by Botometer (which already often mistakenly classifies humans as bots [398]), are drawing to a close with advancements in AI.

The promise of AI-enabled bots lies not just in rapidly raising an info ops army as Team Jorge did with its 39,000 avatars [258, 328], but also in reducing the cost of info ops. "The potential of language models to rival human-written content at low cost suggests that these

¹OEV predates the AI/neural network revolution. Graphika and SIO noted that assets in the pro-Western covert info ops they uncovered "created fake personas with GAN-generated faces"[192], but are no details were provided on whether other content are AI-generated. AIMS bots do not rely fully on AI: profile pictures were "stolen from a genuine profile", although the "creation of [textual] content" is "driven by AI" [54, 328].

models—like any powerful technology—may provide distinct advantages to propagandists who choose to use them” [189]. Accessibility to content generation capabilities, which is correlated with cost, has also been widened with the advent of AI. Image generation via diffusion models such as Dall-E-2, Midjourney, and Stable Diffusion are available as services “for free or for a low price” [198] to practically everyone. Consequently, social media platforms have been flooded by “hyper-realistic fake images deceiving many online” such as those of Trump resisting arrest in March 2023 [415].

Bots can solve the problem of credibility crisis by appearing independent. This is not the same credibility crisis we spoke of earlier where social media platforms were not trusted to be non-interventional (Section 5.2.2.3). This is a credibility crisis among entities pushing the message — usually users of the platform and not the platform itself. Messages from US-linked organizations are distrusted even for the seemingly benign (or even beneficial) cause of counter-terrorism [12]. Attempts at laundering their reputation by using proxies such as “community groups, non-governmental organisations and private enterprise” may be unsuccessful because simply being affiliated with the US in any way taints the proxies’ reputations [12]. However, by conducting info ops through a collective of artificial online personas who cannot be linked back to any state or non-state actor, the info ops would not be subjected to any long-held animosity that the targeted population may have towards an actor, thus encountering less resistance. A scholar examining the US’s credibility problem on social media in the issue of counter-terrorism similarly noted that “producing the appearance of independence by hiding behind fictitious online identities is one option” to address the credibility deficit [12].

Our conception of bots being nigh undetectable, as expounded in this section, plays a critical role in the design of information propagation within *Diluvion*. In *Diluvion*, we assume that bots are bound by many of the same limitations that a human agent would be, e.g., their activity level, as they would seek to blend in with the human population. We also assume that people would not be able to ascertain another social media user’s true nature and affiliations, hence why a person’s judgment on the credibility of a message/tweet relies on external social cues. And this human heuristic of assessing credibility by following the herd makes social media users uniquely vulnerable to decentralized info ops that can be trivially signal-boosted by bots, as we will discuss in Section 5.3.3.2.

However, at the same time, there exists in *Diluvion* a special class of agents separate from humans that are called “bots” (Section 5.3.5); within *Diluvion*, “bots” is a term of convenience for all forms of propagandists. If a human takes on the role of a propagandist (calculatingly penning tweets to maximize engagement, using Photoshop or a physical airbrush to alter images, etc.) on social media, they are rightly classified as “bots” since

they are indistinguishable from purely algorithmic bots engaging in info ops.

The reason we reject the terms “propagandists” or “operatives” in favor of “bots” is because we reject the current erroneous framing within the agent-based simulation community of segmenting agent types based on stance (defined in Section 5.3.1) on an issue. In `twitter_sim` [57], there are “three types of users: normal users (ignorants), bots/trolls (spreaders), and truth defenders (stiflers).” In the model introduced in [32], there are “three types of agents: Deceptive agents, representing bots or malicious human users; neutral agents, representing most Twitter users; and news agents, official sources who share accurate, verified information.” A pattern emerges: bots are invariably associated only with the “bad guys”, which is categorically false since NAFO, OEV, and the pro-Western covert info ops documented by Graphika and SIO are examples of bots working for “good” causes by disseminating “truth”. Under *Diluvion*, all propagandists are bots regardless of their stance.

Reclaiming the term “bots” is necessary to ensure accuracy in analysis and simulation, and thus avoid errors in strategic judgment. Propaganda work and counter-propaganda work are often broadly equivalent in tactics. Persisting in the delusion that the “good guys” behave in a radically different manner in matters of war is detrimental to the formulation of good policy. Researchers will chase after the wrong solutions due to that initial error in judgment. Recall the example of the failure of US counter-terrorism info ops on social media [12]. Foreign policy analysis and reporting, e.g., [130, 173, 178, 324], will fixate on attributing the success of Islamist or ISIS propaganda over US counter-propaganda to their speed, sophistication, creativity, having a better narrative, etc., and rarely question whether the failure of counter-propaganda lies in the counter-propagandists’ lack of credibility, in the message listeners’ fear and loathing of the US. The State Department’s Digital Outreach Team (DOT) accusing ISIS of terror tactics via social media accounts clearly affiliated with the US [12] appears oblivious to the fact that the use of terror tactics by the US itself² undermines such a message. The initial error in judgment in assuming that the US’s hands are clean ensured US counter-propaganda efforts’ continued failure [12].

5.2.2.6 State and non-state actors

Other notions that we need to disabuse ourselves of for the sake of strategic clarity is that info ops originating from the “good guys” — democratic states — are only sanctioned

²In recent years, it has come to light, through the reporting by outlets such as Current Affairs, The Guardian, and The Wire [1, 83, 325], that indiscriminate bombing via drones has instilled terror in the hearts of the populace in many regions of the Middle East. Children have come to fear clear blue skies because those are the perfect weather for flying drones.

against so-called non-democratic states and that meaningful distinction exists between state and non-state actors. Democratic states attack each other and themselves. State and non-state actors lead a chimeric existence.

For social media platforms, the distinction between state and non-state actors is increasingly blurred due to the vast reach that the platforms have in people's lives, leading governments to constantly intervene in the platforms' daily operations [135] or even prop up the platforms, e.g., the case of Twitter, Iran, and the US State Department [81] and the case of USAID-funded ZunZuneo [26, 82]. So deeply and confoundingly intertwined is the public-private relationship that in [378], the author explicitly set out to argue that it is simultaneously unlikely and likely that courts will conclude that a social media company is a state actor.

Distinguishing between private and public entities among those engaged in info ops would be similarly difficult. Former intelligence operatives naturally gravitate towards private info ops firms, e.g., [40, 197, 329, 330, 463], and these operatives bring along their contacts in the government, skills, and knowledge into their new roles [40]. One remaining difference may be the scale of resources that state-level actors can muster compared to non-state actors, but even that disparity is vanishing as states outsource info ops to private entities, potentially as a form of legalized corruption via awarding lucrative contracts [92, 173] and as a way to shield themselves from culpability (Section 5.2.2.4). According to Oxford Internet Institute's 2020 Industrialized Disinformation report, despite the difficulty in "tracking down contractual evidence of private contracting firms", they found that "almost US \$60 million was spent on hiring firms for computational propaganda since 2009" [69]. Public-private joint efforts can be seen in CENTCOM awarding a \$2.76M contract to the private corporation Ntrepid to develop OEV's persona management software [111]; in the US State Department awarding a \$500,000 contract to Cambridge Analytica's parent company SCL for "target audience research" [461]; in the many info ops companies manipulating democratic elections at the behest of clients from democratic states [38, 40, 54, 129, 143, 258, 328, 329, 330, 412, 463, 504]; in Team Jorge running info ops through "Demoman International, which is registered on a website run by the Israeli Ministry of Defense to promote defence exports" [258, 328]; and in NAFO, which according to a Kyiv Post interview of its founder, is "endorsed by U.S. Congressman Adam Kinzinger, as well as U.K. Minister of Defense Ben Wallace" [443].

Conducting info ops against other democracies is not beyond the pale for democracies. Many documented instances exist. The campaigns conducted by the UK-based Cambridge Analytica in support of Trump and in collaboration [197, 330, 461] with the Israeli Team Jorge is a well-known example [38, 143, 504]. The actions of Israeli info ops companies in

meddling with non-US democracies — e.g., Team Jorge’s interference³ in “33 presidential-level election campaigns, 27 of which were successful” [258, 328]; Percepto’s info ops against a Burkina Faso election [329]; Archimedes targeting a Nigerian election [129] — are no less impactful simply because they received comparatively little attention in mainstream media and scholarly circles. Israeli Prime Minister Netanyahu also attempted to influence the 2016 election in favor of Trump [40]. Info ops conducted by an outside actor may even be orchestrated by paymasters situated within the targeted state. Trump campaign official Rick Gates sought proposals from an Israeli company Psy Group “to create fake online identities, to use social media manipulation and to gather intelligence to help defeat Republican primary race opponents and Hillary Clinton”; the proposal, named Project Rome, was ultimately never acted upon [322, 389]. The US under the Obama regime in turn has employed covert info ops against Israel [10]: a US Senate inquiry found that “[i]n service of V15 [an anti-Netanyahu group], OneVoice deployed its social media platform, which more than doubled during the State Department grant period; used its database of voter contact information, including email addresses . . . and enlisted its network of trained activists, many of whom were recruited or trained under the grant, to support and recruit for V15.” Even a seemingly innocuous “voter button” displayed on Facebook to encourage turnout in UK elections can be construed as an attack because Facebook, as a US company, should not be influencing UK politics in any way [193].

All these examples are in essence the use of non-state intermediaries in information warfare involving three state-level actors widely considered to be democracies (according to the Economist Intelligence Unit’s 2016 and 2022 Democracy Index [148, 149], the US and Israel are flawed democracies and the UK is a full democracy), putting to rest the idea that info ops are only ever employed in the stalwart defense of peace, freedom, and democracy against states with autocratic, authoritarian, totalitarian, etc. tendencies⁴.

Info ops’ use by democracies is not limited to state-to-state situations. Info ops can be

³Veracity of Haaretz’s reporting: “Journalists from The Guardian, Der Spiegel, Die Zeit, Le Monde, the international organization of investigative journalists OCCRP, Radio France, Haaretz, TheMarker and other media outlets worked in France, Kenya, Israel, the United States, Indonesia, Germany, Tanzania and Spain to examine the veracity of Jorge’s claims about his worldwide deeds. Shockingly, many of the allegations were corroborated.” [328]

⁴For purveyors of “No True Scotsman” and “Democratic Peace” [257]: Some academics have argued that Israel is an apartheid state [155] and also a settler colonial state [295]. The US has been characterized as an oligarchy [182] and as a plutocracy [46, 181], and some Americans prefer to define the US as a republic [219]. The UK may fit the definition of an “electoral authoritarian” regime [422] due to a history of banning political parties [68], media censorship [60], and electoral fraud [157], on top of inheriting a hereditary monarchical head of state and hereditary peers in the House of Lords who exercise indirect political influence [240]. An argument can also be advanced that certain states within the triad, being client states, lack true sovereignty [166, 456] so conducting inter-state info ops does not constitute warfare as no one’s sovereignty has been violated.

used against a democracy’s own populace. In India, the world’s largest democracy, “[t]he 2014 election of Narendra Modi catapulted the BJP to power thanks, at least in part, to an expansive network of ‘IT cells’ aimed at spreading positive news about the BJP and attacking its detractors” [412]. Cambridge Analytica is believed to have influenced the Brexit vote results [330, 401], though its culpability is disputed by some [50].

Having established that info ops originating from state and non-state actors would have no appreciable difference due to ever-present public-private entanglement, we can claim that incorporating the publicized details of info ops conducted by non-state actors into our modeling will not bias the model towards the form of info ops conducted by non-state actors and vice versa, and we can claim that findings from *Diluvision* simulations are equally applicable to info ops from or against either state or non-state actors.

In *Diluvision*, we defaulted to bots starting out embedded in a social network prior to being activated, in contrast to some preceding models that started bots out with only a single connection (edge) to the network, e.g., [57], or let bots have lower degrees than humans, e.g., [409], as they assumed that communities have not been penetrated by bots prior to an info ops. Our choice is based on reports of uncovered info ops. Team Jorge estimated that about 1,000 bots are needed to “postpone an election in Africa without good cause” with half being newly created bots and the other half preexisting bots [54]; the “inventory of fictitious accounts” were categorized based on regions/identities (e.g., Arabs, Russians, Asians, and Africans) and reused for different info ops [328] — segmentation by region and recycling bots for multiple campaigns strongly suggest that Team Jorge would equip bots with social relations (if the bots’ positions within a social network did not matter, swapping in a campaign-appropriate regional identity on preexisting bots instead of maintaining an extensive inventory would have been more cost-effective). AIMS bots also exhibit meticulous attention to detail, with some of the bots even having bank accounts [54]. Taking that fact into consideration, Team Jorge are unlikely to neglect situating their bots inside a suitably believable network of social relations at creation time. Percepto’s case adds further support to the likelihood of bots having relations. One of Percepto’s bots was an entirely fictitious investigative reporter based in Paris that had a real news agency with actual humans working under the bot [329]. Even the simplistic bots that apparently failed in manipulating Singaporean elections “appear relatively ubiquitous throughout the social network” despite a “failure to amass network influence” [471]. The strongest evidence comes from bots used in the covert pro-Western info ops studied by Graphika and SIO [192], which exhibited “a typical long-tail distribution in the follower footprints, with a few influential accounts followed by a descending list of accounts with progressively fewer followers”, showing that the distribution of node degrees among bots mimic that of

humans on social media (in most [42, 318] but not all cases [73] of human networks). The probability of an info ops bot army starting out from the periphery of a network is low.

5.3 The Model — Diluvion

In our information diffusion model for Twitter-like social media, *Diluvion*, information is composed of two elements: stance and themes. Stances are polarized and exclusionary. A message can have multiple but non-repeated themes. The dissemination of stances through the network is of greater concern to us than themes. Details on stances and themes are found in Section 5.3.1.

There are two major classes of agents in our model, humans (Section 5.3.4) and bots (Section 5.3.5). While “humans” are exclusively humans, “bots” can encompass artificial algorithmic entities and natural organic entities so long as they play the role of propagandist. Our motives for using this terminology are explained in Section 5.2.2.5.

Section 5.3.2 explains the attention-memory mechanism that agents are subjected to. In brief, the mechanism aims to mimic the impact that real humans’ limited information processing capability and retention capacity can have on information diffusion. Section 5.3.3 lists the range of actions that agents can perform within the simulation to propagate or boost the propagation of stance and themes. Section 5.3.3.1 describes statistical data — engagement metrics — that agents are exposed to and affected by. Section 5.3.3.2 describes indirect influence mechanisms as well as the rationale behind our model’s reliance on engagement metrics as proxy for indirect influence.

The chosen resolution for *Diluvion* simulations, i.e., the length of real-world time that a single step in the model is equivalent to, is justified in Section 5.3.6. The starting parameters of the simulation are detailed in Section 5.3.8.

An assumption under our model is that all agents are devoted to debating one topic. They may devote a differing amount of time per day on social media and their forms of engagement may differ (e.g., spending most of their time liking posts they agree with, primarily retweeting others, etc.). These additional traits of the agents are sufficiently divergent that each combination constitutes a sub-class of bots and humans (Section 5.3.8.1).

Private accounts are excluded from our model primarily because they have a limited ability to engage in public debates; a private account’s tweets can only be seen by the account’s followers. For instance, if a private account sees an opinion they disagree with, refutations that the private account u posts in reply will not be seen by the author v of the disagreeable tweet and the author’s, v ’s, followers if the author v is not a follower of

the private account u . This limited reach of private accounts makes them inappropriate for inclusion in simulations of public info ops. Additionally, private accounts constitute a small minority of Twitter users so their exclusion would not drastically affect our results. According to a Pew Research Center survey on adult US Twitter users conducted in 2019, 13% of Twitter accounts are private [399].

We also exclude the “silent majority”/“no activity” users (Section 5.3.8.1) from our model because their lack of activity means that there is little data to collect on how they exert influence and how they, in turn, are influenced. Their lack of activity also means that they would have no discernible impact on the evolution of a Twitter discourse.

DiluvSION is built on the Python agent-based modeling framework Mesa, which in turn uses the Python network analysis package NetworkX as a backbone.

5.3.1 Stances and themes

Stances are opinions on an issue. To ease interpretability, we allow only three discrete stances in our model: two located at the poles and one at the exact midpoint between the two. We let the distance from the midpoint to either pole be one. The three stances are known as negative, neutral, and positive. Alternate names used throughout this paper for the negative stance include “con”, “against”, and “opposition” and the positive stance is also known by “pro”, “favor”, and “support”. While not all opinions can be modeled under this system, e.g., opinions on the actual worth of a used car, many can be, e.g., opinions on voluntary abortion.

Our model differs from others in the literature, e.g., [280], in that we allow neutrality as a stance to be propagated. Bipolar ideological battles have been consistently framed as drawing users to either pole, but not taking a stand can be as principled a stance as taking a stand, at least in principle, although perhaps not always in practice (some may argue that true equidistant neutrality in the strictest sense does not exist as everything exhibits a bias⁵). An example of a manifestation of neutrality is the Non-Aligned Movement [477] in

⁵Academics, e.g., [350, 534], have argued that mass media always represent ruling class interests, making impartiality impossible. To maintain a facade of impartiality, mass media “strictly limit the spectrum of acceptable opinion, but allow very lively debate within that spectrum” [37]. Others have argued that journalism, “in its pretense of ideological neutrality”, created “a particularly pernicious myth” as the presence of filters in both the news reporter and news consumer prevents neutrality [338]. Spiral of silence [363] predicts that some people are unwilling to voice out an opinion to avoid contradicting what is perceived to be a majority opinion. Such inaction can be interpreted as neutrality, yet it does not contribute to the emergence of neutrality but instead contributes to the (potentially false) majority’s continued dominance. Central banks invoking market neutrality to depoliticize corporate security purchase decisions are biased because “[a]s critical political economists, sociologists and anthropologists have so frequently argued, markets embody a specific, political vision of society, which is not shared by every member of the polity” [473].

geopolitics.

Stances are different from sentiments. Sentiments are concerned with the emotive content of a message, whereas stances are concerned with the message's position with regards to an issue of interest. For example, both statements "I hate wearing masks" and "I hate doctors not wearing masks around immunocompromised patients" show a negative sentiment due to their negative tone and presence of words such as "hate" in them, but in terms of stances, the first statement is against mask-wearing while the second is in favor.

Themes are the non-polarizable aspects of an opinion, and they are non-polarizable only with respect to the issue of interest. We illustrate the concept with an example where the issue of interest is mask-wearing: a message decrying a store for banning masked patrons for fear of criminal intentions and another message praising a store for banning unmasked patrons out of health concerns, despite differences in stance (anti- vs. pro-mask) and sentiment (negative in condemning vs. positive in complimenting), are united under the common theme of "non-medical authority" being exercised. And although people may have opinions on authority (favor or against), those stances have no direct bearing on the issue of concern, mask-wearing. Themes are comparable to the concept of frames in the field of communications [103, 275, 484] but is *not* equivalent to frames⁶. A frame is envisioned as a wrapper, a method of carrying messages. Different perspectives (wrapper) on an issue (message) changes one's perception on an issue. For us, themes function as more than wrappers. They are capable of being transmitted to (infecting) a target, just like stances. There can be multiple themes per message just as there can be multiple frames/wrappers per message. Multiple themes are similar to opinions encoded as binary strings in the rumor spreading model introduced in [485] and to a lesser degree the discrete-numbered cultures in the Axelrod model [34], SMSim's single-topic bit vector [126], and AMID's fuzzy set multi-topic messages [279].

The usefulness of explicitly modeling themes and having multiple themes per tweet/user can be demonstrated with an example involving an unpopular theme (adopted by a low percentage of agents). To make things easier to understand, we name this theme "conspiracy".

⁶To underscore the importance of themes/frames: Incidentally, the concept of framing introduced by Tversky and Kahneman was first motivated by the "Asian disease problem" [469], where people were surveyed on their preferences to two solutions to an "unusual Asian disease", with one solution framed as saving lives and the other as causing deaths. Framing the disease as Asian betrays a problematic inclination harbored by the authors in their supposition that a disease terrible enough to force people to face a moral dilemma could have only originated from Asia or Asians, much like SARS and SARS-CoV-2 are exoticized and racialized as Asian [132, 529]. But perhaps the framing of framing's motivating problem can be excused as it dates from an era when people have yet to learn to better hide their prejudices. In the own words of Kahneman, "[s]ome people have commented that the 'Asian' label is unnecessary and pejorative . . . but the example was written in the 1970s, when sensitivity to group labels was less developed than it is today" [239]. The term "Asian disease problem" remains in common usage in scholarly circles today.

By wrapping every message a bot army tweets out with a layer of conspiracy and by having more popular themes accompany the conspiracy theme since multiple themes are allowed, we can ensure that the conspiracy theme is more willingly adopted by others because of its constant presence and its association with popular themes. This ultimately results in a conspiratorial mindset slowly taking over the targeted human populace on social media. We indeed simulated such a scenario under *DiluvSION* (Sim 9, Section 5.6.4). A less nefarious theme can always take the place of “conspiracy”, e.g., “affordability”.

Each message, i.e., tweet, in our model is composed of a single stance and one or more themes. Each agent has a stance and a set of themes as well. Agents can switch stances and can adopt themes over time. To keep the agents’ themes from homogenizing as the simulation progresses because of the adoption mechanic, the maximum number of themes that an agent can possess at any one time is always lower than the set of all available themes in a simulation. This limitation also reflects that not all aspects (themes) of an issue can be the core interests of an agent; people have a limited capacity to care. All tweets by an agent will have the same stance as the one held by the agent at the time of tweeting. Themes, however, can differ between tweets (Sections 5.3.4 and 5.3.5), e.g., a human agent will use the same themes as the tweet it is replying to, even if the themes are not part of the agent’s theme set. Under *DiluvSION*, compatibility in both the stance and the themes is required to improve the persuasiveness of a message to its recipient (Section 5.3.8.2). Once again using a mask-wearing example, a message that attempts to proselytize mask-wearing through the themes of safety and medical authority will find no purchase on an agent whose themes reflect a concern for comfort, e.g., masks being difficult to breathe through. Stance plus themes can also be thought of as viral variants, except that the transmission of stance and themes is dependent on homophily, but viral variants do *not* necessarily operate under this mechanic⁷.

5.3.2 Memory and attention

Memory and attention mechanisms are common features in many models of information diffusion, e.g., [57, 279, 280, 485, 496] and the motivations behind incorporating such

⁷There is some superficial similarity in how newer variants of SARS-CoV-2 such as Delta and Omicron evolve greater binding affinity to ACE2 receptors in human cells than the ancestral wild type [428, 508]. Antibodies and some therapeutics either aim to present a more attractive binding target for the virus or bind to ACE2 receptors first before the virus [161, 446, 486], essentially outcompeting the virus. Themes are analogous to binding affinity, where a tweet exhibiting a greater number of matching themes (homophily) with the targeted user increases the chance of the tweet’s stance being accepted by the targeted user, just as a the strength of binding affinity can influence whether an antibody (a stance) or a virus (a different stance) binds to a host cell.

mechanisms are largely similar. Modeling agents as having limited attention and limited memory is meant to reflect the reality where humans are incapable of digesting all the information presented to them and retaining the digested information over the long term. *Diluvion* features an attention and memory mechanism. An attention-memory limit governs the number of tweets that an agent sees per activation (when an agent wakes during a time-step and performs actions), the number of its own most recently tweeted tweets that the agent remembers, and the number of users that the agents remember seeing. Memory observes the first-in first-out (FIFO) queue, with the newest item displacing the oldest should the limit of a memory bin is reached. Since it is a combination attention and memory limit, it is possible for items that have entered an agent's memory to be forgotten before the items have a chance to be seen by the agent. For instance, if there are 31 new tweets directed to an agent whose memory is empty and has a limit of 30, the earliest/oldest 31st tweet will be "forgotten" before ever being read by the agent. The memory of seen users is duplicative and cumulative: a user can appear more than once in an agent's history and users that appear more frequently are given greater weight in calculations (e.g., when deciding which user to follow).

While we keep track of the history of *all* users seen by an agent (this history is different and separate from the limited memory seen user history) and the history of *all* influential users (users that infected/changed the stance of the agent), they do not affect agent behavior and are only used by us to diagnose how influence spreads in the social network (Section 5.5). These histories are also duplicative.

Bots are constrained by a memory limit though theirs is several times higher than that of the humans. Bots are assumed to be more dedicated to spreading stances and/or themes than regular humans whose social media usage is less singularly goal-driven. Therefore, bots would dedicate more resources, such as greater attention and memory, to accomplish their goals.

5.3.3 Engagement

The ways in which a user can engage with the social media platform Twitter are referred to as actions in our model. The set of actions available to an agent in our model — following, unfollowing, like, tweet, retweet, quote-tweet, and reply— is different from the set of engagement metrics, which we will discuss in Section 5.3.3.1.

Following is a one-way relationship where the follower u indicates their interest in receiving updates on the activities of the followee, v , e.g., when v tweets, retweets, etc. Unfollowing is the sundering of the follow relationship. A follower following a followee

indicates that the follower is receptive to information flowing out from the followee. A loss of interest in the followee and/or the followee’s tweets losing their informativeness often leads to the demise of the follow relationship [270, 511]. In our model, a potential followee having matching stance and themes with a user increases the candidate’s chance of being followed by the user.

Liking (previously known as “favoriting”) is the most voluminous [25] action performed by Twitter users. Likes on social media can be considered a “numeric representation of social acceptance” [408]. The bandwagon effect — the heuristic of following the crowd used by humans when other signals of authenticity are weak or absent — can guide a person’s decision to like, as people were found to be more likely to like photos with many likes over those with few likes [433].

Tweeting is the act of composing and posting a message on the social media platform. Since our model excludes private accounts, all tweets are assumed to be publicly viewable and searchable. While there are a number of factors (Twitter’s algorithm, a human user’s own preferences, etc.) affecting whether these viewable tweets are actually seen by users, our model assumes that an agent’s tweets are always seen by the agent’s followers so long as the tweet remains in the memory (Section 5.3.2) of the followers. While not the most common action, the production of tweets is the basis for all other actions, except for following and unfollowing, because liking, retweeting, etc., require tweets to first exist.

Retweeting is the sharing of another’s tweet with one’s followers without any embellishment. According to [335], retweets indicate interest, trust, and agreement with the message. As credibility and trust are important to having a tweet retweeted, our model factored trust into an human agent’s decision on which tweet to retweet. Rather than using a randomly generated number to quantify trust relationship between agents, we relied on indirect influence, namely engagement metrics (like, retweet itself, and reply), as a proxy for trust (detailed in Section 5.3.3.1, justified in Section 5.3.3.2).

Replying is responding to another’s tweet with one’s own tweet, not as an independent tweet but as a tweet that will be added to the original tweet’s reply thread, easily accessible by others should they tap or click on the original tweet. While not all replies challenge the stance of the parent/source tweet, replies are strongly weighted towards being negatory in nature. The Stance in Replies and Quotes (SRQ) dataset [478] shows that 44.0% of the replies are against the stance expressed in the tweet being replied to (parent/source tweet), and that approximately half, 22.9%, are supportive. The remaining 33.2% are comments and queries about the parent tweet’s stance, which can be interpreted as being neutral in nature. In the problem of detecting contradiction for rumorous claims [273], there is a special category for detecting disagreeing replies (as opposed to expressing disagreement in

independent tweets), highlighting replies as vehicles for disagreement. Our model adheres to the conception of replies being primarily, though not exclusively, negatory in nature.

Quote-tweeting, alternatively known as quoting, is sharing another's tweet with one's followers while adding extra content from the quote-tweeter. Quote-tweeting is relatively rare, even when compared against replies. In a dataset of COVID-19 tweets that consisted of just retweets, quote-tweets, and replies, quote-tweets' share is 3.82% while replies have 4.91% [141]. A study on the roles assumed by Twitter users on the platform [25] showed that quote-tweeting is seldom practiced, with Repliers as a role and replying as an action outnumbering Quoters and quoting across all roles and user activity levels save for Quoters themselves. The motivation behind quoting appears equally divided between refuting and affirming the original tweet. In the SRQ dataset [478], 46.8% of quote-tweets are negatory with respect to the quoted tweet and 43.5% are affirmatory, with the remaining 10.7% being comments and queries regarding the quoted tweet's stance. Stance ambivalence in motivating quote-tweets, as evinced by the even split between negatory and affirmatory, is modeled in *DiluvSION* as compatibility in stance having a lower weight than themes when an agent picks which tweet to quote-tweet.

Ratios can also offer clues on whether agreement or disagreement drives an action (Section 5.3.3.1). Actions as modeled in *DiluvSION* is generally in agreement with findings from ratiometrics.

In our model, an agent's decision to like, retweet, or reply to a tweet is partially based on the agent's agreement with the tweet's stance and themes, and partially on whether the tweet's popularity merits bandwagoning. Whether popularity or compatibility is the dominant factor depends on which action is being taken, as detailed in the explanation of this mechanism in Section 5.3.8.2. The existence of a follower-followee (or inverse) relationship between agents does not influence an agent's action; the "text semantics and contents" of a tweet, "regardless of tweet author", are the most important factors in determining if a user engages with a tweet through liking, retweeting, replying or quoting [464].

Agents who received replies to their tweets or have had their tweets quoted usually receive notifications, so we assume that information from replies and quote-tweets diffuse to the targeted agents. Additionally, replies to a reply-tweet are read by a fixed percentage of the followers of the author of the tweet being replied to, which is a form of non-social tie information diffusion (Section 5.3.3.2).

Like Twitter and other real-world social media platforms, we only allow a user to like a tweet once and retweet a tweet once. The practice of including a link to a tweet, a in a reply-tweet, which increases the retweet counter of a and can be performed multiple times,

is excluded from our model. And although it is possible to reply to the same tweet multiple times on Twitter, we restrict the number of times an agent can reply to a tweet to just one. Any information that a replier wishes to convey through multiple replies to a parent tweet can be thought of as having been consolidated into one reply.

Tweeting is the action that all agents, bot and human, default to if the agent’s attention-memory contains no seen tweets or if the chosen action for a particular time-step of the simulation is invalid. Many actions, e.g., liking and retweeting, require tweets to be present before they can act on those tweets, so this default behavior helps populate the simulation with tweets. Invalid actions are common around simulation start due to a confluence of three factors (1) the lack of tweets, (2) limiting each tweet to one like, retweet, quote-tweet, and reply each per user, and (3) liking and retweeting being common actions. A consequence of tweet scarcity is that the action distributions generated from the simulation deviate slightly from empirical data, despite the model using parameters from empirical data, as we shall discuss in Section 5.4.

5.3.3.1 Engagement metrics and ratioing

There are three prominent engagement metrics — likes, retweets, and replies — visible to Twitter users. While we allow for quote-tweeting as an action in our model, its metrics are not visible to agents in our model and are therefore incapable of influencing the agents (Section 5.3.3.2). In the context of stance conversion, likes are positive, retweets are moderately positive, and replies are moderately negative.

Excluding the number of quote-tweets from engagement metrics is because of this metric being hidden throughout much of Twitter’s existence. In 2020, quote-tweets were still semi-hidden, with Twitter forcing users to tap/click on a tweet then on the tweet’s retweets before they could see the number of retweets broken down into retweets and quote-tweets [383, 470]. Besides quote-tweets, we also exclude additional metrics that have been surfaced since Elon Musk’s Twitter acquisition, such as the counts of views and bookmarks a tweet has received. This is because the foundations of our model — engagement data we and others have collected (Section 5.3.8.1) and the vast majority of scholarly literature on the impact of engagement — date from the era when these statistics were hidden. On Twitter’s current timeline interface where users scroll through tweets, only viewcounts have been added alongside the counts of likes, retweets, and replies. The newly surfaced engagement metrics are also rarely displayed on other platforms, e.g., Instagram and YouTube only show the number of likes and replies that a post has received while the Twitter-like Mastodon shows retweets, replies, and likes.

The number of followers is an engagement metric that does not influence human agents.

Although a user's follower count is publicly visible to anyone who visits the user's profile page, it is not visible in the context of other actions — liking, retweeting, replying, and quote-tweeting. Visiting a user's profile page is an extra step from the typical Twitter usage behavior of “infinite scrolling”, breaking from a user interface (UI) designed to entrap users in perpetually consuming media [305, 362]. For following and unfollowing a user, these actions are available as part of a dropdown menu accessible by clicking an icon on a tweet; visiting a user's profile page and being made aware of the user's follower count are entirely avoidable. Bots, however, do use follower count as information for some of their decisions, as they are motivated agents of stance/theme propagation.

Ratioing on Twitter is a phenomenon that best expresses the logic behind engagement when Twitter conversations involve contested issues. When a user expresses an objectionable opinion, others may express their disagreement by replying to the user's tweet. Should the number of replies (regardless of whether they all express disagreement) or the number of likes of any one of the disagreeing tweets outnumber the likes of the parent tweet, a “ratio” has been achieved, i.e., the ratioed user has been publicly and quantitatively mocked. A study on an info ops campaign waged by social media users opposed to the fossil fuel industry used ratioing as a means to identify tweets targeted by the operatives as the operatives employed the strategy of questioning fossil fuel organizations' legitimacy with large numbers of negative comments [466]. A broader definition of ratioing includes cases in which the number of replies exceeds the number of retweets. This definition was employed in a study of the different ratios of engagement metrics that tweets from Presidents Trump and Obama received [342], which found that Trump consistently achieve higher reply-retweet — more controversial — ratios compared to Obama.

5.3.3.2 Indirect influence and trust

Our conceptualization of influence departs from some works in the literature that regard all influences as having a positive impact. These works include [397], which posits that the high trustingness of an acting agent and the high trustworthiness of the acted-upon agent are the determinants behind the acting agent's decision to act (retweet, reply). Another work is [525], which deemed one user as having influenced another if the user engaged with the tweet in any way (retweet, like, reply, quote, and follow). From the perspective of stance diffusion, such conceptualizations of influence clash with empirical observations of why people choose to respond to messages. As discussed earlier in Section 5.3.3, replies and, to a lesser extent, quote-tweets are often used to make rebuttals. Ratioing (Section 5.3.3.1) demonstrates that not all interactions constitute affirmative influence. The disconnect between the models [397, 525] and the data is a result of ignoring polarity/stance. Zhang

et al. [525] explicitly stated that their method of finding matching tokens between two tweets “identify whether their topics are similar”; nothing is said about the user’s stance on the topic.

Direct influence comes from direct attempts at persuasion, which in our model consists of the contents of a tweet — its stance and themes. Direct influence travels along the prescribed lines of influence, social ties, from followee to follower. Indirect influence is information outside of a tweet’s content that can influence an agent’s adoption of the contents. Indirect influence can either increase or decrease the persuasiveness of tweets. Indirect influence can operate outside of observable follower-followee relationships.

One form of indirect influence modeled under *DiluvSION* is the ability for replies to circumvent social ties when transmitting information, where a certain percentage of the followers of the parent tweets’ authors will read the replies (Section 5.3.8.2). Tweets in most cases can be replied to by any user, not just followers or followees. More than just transmitting information to the user targeted by the reply, replies can also transmit information to the targeted user’s followers. A browse through Twitter will reveal that many replies to tweets that have gained a lot of attention (high number of likes, retweets, etc.), colloquially known as “going viral”, are advertisements or posts about topics tangential to those of the original tweets, e.g., fake BTC giveaway scams [321] and shilling for cryptocurrency. They are attempts to take advantage of the tweets’ popularity. The hope is that some of the attention can be diverted to profitable or persuasive ends.

Attempts have been made to model such indirect routes of information diffusion. The model proposed in [32] allows for tweets that have crossed a certain threshold of engagement numbers to go viral, meaning that the tweet will be “shown to all agents in the network that share the same topics with the sender agent, even if an edge does not connect them”. From our perspective, the indirect propagation of information captured in their implementation appears limited, as most information still travels along the network edges in their model. The sole exceptions were viral tweets. In *DiluvSION*, virality would naturally emerge from the propensity for agents to favor sharing or otherwise engaging with popular tweets over unpopular ones (bandwagoning) since engagement metrics, themselves a form of indirect influence, factor into an agent’s decisions on which tweet to engage.

The engagement metrics discussed in Section 5.3.3.1 constitutes a form of indirect influence. The metrics serve as both [152, 285, 435] social cues for the bandwagon effect [44, 355] and social cues for computing source trustworthiness [279, 285, 397] and message credibility [435].

Engagement metrics taking the place of credibility is supported by the existing research literature. One study found that “bandwagon heuristics—such as the number of likes,

comments, and retweets—increased the credibility of news” [435]. The automated activity of social probing by non-human-like bots with zero trust, i.e., strangers, which in the case of [11] consists of visiting the profile page of others, can boost popularity. That popularity translated successfully to influence, which was getting others to follow accounts recommended by the bots. The number of retweets, despite its noisiness and bias, has a positive correlation with expert-assessed credibility of messages, and it is effective at predicting ground-truth credibility when combined with credibility ratings from a different crowd, Mechanical Turk annotators [440]. Cues in the form of retweets (information sharing) boosted the credibility of a tweet’s author regardless of whether the tweet originated from a peer or a stranger, and even if the retweets were performed by strangers [285]. Conversely, source credibility was found to play a major role in the information sharing behavior of users on Twitter [196]. Credibility and influence are intertwined, feeding off each other and creating a rich-get-richer effect, where the wealth is social capital.

The need to impute a separate trust value for a social tie shared between two agents is also obviated by engagement metrics that can serve as proxies. We improve upon the trust mechanism found in [279, 397], where the proxy variables for trust are the number of times a user has retweeted [279, 397] and replied [397] to another user, by factoring in the relationship between engagement type, stance distance, and themes; weighing the different engagement metrics; and counting indirect engagement actions instead of direct ones. The authors of [397] themselves noted that replies as a proxy can be unreliable because replies may contain users expressing a stance different from the parent tweet, lending support to our decision to account for stance when using engagement metrics. Whether an engagement type is driven by heterophilia or homophilia in stances and themes (and consequently the metric generated for the engagement type) under our model is explained in Section 5.3.3.

As for using indirect engagement numbers over direct ones as proxies for trust, indirect engagement can be more reliable than direct engagement despite appearing counter-intuitive.

Traditional indicators of sources as being authoritative are less visible online [336] because of a flattening of hierarchies. Surmounting the accessibility barrier (e.g., radiocommunication spectrum licensing, capital costs of print media) and the expertise barrier (e.g., sufficiently credentialed to be invited onto mass media to comment on an issue) denoted authoritativeness pre-social media, but under social media where such barriers towards broadcasting a message has disappeared, so too has their function as indicators of trustworthiness. Distrust in authoritative sources has also been on the rise, further weakening traditional signals of credibility. Distrust in authority being a widespread phenomenon is

evident in Twitter discourse on vaccine and climate change where “science” (science as a source of authority, not science as practiced) is regarded with suspicion [201, 227, 394]. While an authoritative source can still engender trust on social media, a stranger retweeting information from an authoritative source actually *reduces* the trustworthiness of the source, whereas a stranger retweeting peers or other strangers increases trustworthiness [285]. The rising distrust in authorities was an impetus for the development of bot-augmented decentralized info ops (as discussed in Sections 5.2.2.4 and 5.3.5). On social media, the audience has also been flattened through the phenomenon of context collapse [319], where users cannot tailor messages to pander to specific subgroups because messages posted on social media are viewable by all, making it easy to offend others, with the offense causing diminished trust in the source [159].

With direct trust in a message’s source significantly dampened by flattened hierarchy and context, message credibility and user trustworthiness online are strongly coupled. Social media users become only as trustworthy as the credibility of their messages. People judge the credibility of information by engagement metrics attached to that information [285, 435]. Consequently, crowdsourcing trust is an established practice, e.g., computation of a Twitter user’s credibility based on the degree of agreement between the user’s tweets and their respective replies [502]. Engagement-based source credibility indiscriminately boosts the credibility of peers *and* strangers [285]. These engagement metrics are susceptible to the bandwagon effect, in which people base their opinion on what they perceive to be the opinion of the collective herd [285, 435], e.g., liking what others liked and retweeting what others have retweeted. Another factor at play is minority-avoidance and the spiral of silence, as discussed in Section 5.2.1.3. Trust quantified by indirect engagement numbers is the overwhelmingly dominant form of trust online.

That is not to say that a message source under an indirect trust system, e.g., as implemented in *DiluvSION*, is wholly incapable of affecting the message’s credibility, and by extension, the source’s trustworthiness. A source that has amassed a large following of users sharing similar stances is likelier to receive more likes for its message than a user with zero followers.

The pervasive indirect influence of engagement metrics extends beyond how one perceives others and into how one perceives oneself. Every additional like or retweet that a user receives for their own tweets serves as a small validation, affirmation, and acceptance of the online self [75, 408]. Under *DiluvSION*, an agent’s memory of its own recently tweeted tweets also influences stance adoption, and the agent’s perception of those recently tweeted tweets are shaped by their respective engagement metrics. Because an agent’s tweet’s stance is always the same as the agent’s (unless they recently changed their stance) under our

model, if an agent's recent tweets have received a lot of affirmation (e.g., a high number of likes), the agent will be more likely to retain their stance even when they encounter a lot of opposition to their stance in their memory of seen tweets (memory of others' tweets that they saw).

Figure 5.1 demonstrates how indirect influence operates to infect an agent. This example omits the influence of stance and theme compatibility between the tweets and the agent in an infection event. For an example that does include it, see Section 5.6.5.

5.3.4 Humans

An agent designated as "human" in our model is an idealized representation of a typical social media user. All human agents follow the same heuristics, and they also have the same maximum memory size and maximum number of concurrent themes. Human agents may differ in their starting stance and themes, starting infection resistance, activity level, and roles. Over the course of the simulation, their stances and themes may change. Resistance evolves with every infection (stance change). Their activity level and roles remain fixed. The heuristics and maximums remain fixed as well.

The roles taken on by human agents, as well as their activity levels, are based on empirical Twitter data (Section 5.3.8.1). Roles are the five groups that users were placed into based on their favored engagement action on Twitter, e.g., they were known as Likers if they exhibited a tendency to like others' tweets rather than compose their own tweets. Users were also separately placed into three activeness categories depending on the number of observable actions taken per week. Each role, with respect to its activity level, will have the highest frequency for the engagement action that the role is named for, e.g., moderately active Quoters will have the highest frequency of quote-tweeting compared to other moderately active roles. However, the same role will have different action frequency distributions between different activity levels, e.g., a highly active Quoter will be likelier to like tweets than they are to quote-tweet, but a moderately active Quoter will quote-tweet more often than like.

At every time-step, a check is performed to determine whether a human agent wakes. If it wakes, it will read tweets and potentially be infected by the tweets it has read. If it is awake, a second check is performed to determine whether it has used up its allotted quota of weekly actions (this quota is based on its assigned activity level). If it has not, the agent will definitely act during this time-step. If it has, the agent will have a very small fixed random chance of acting. The actions that an agent can perform are discussed in Section 5.3.3.

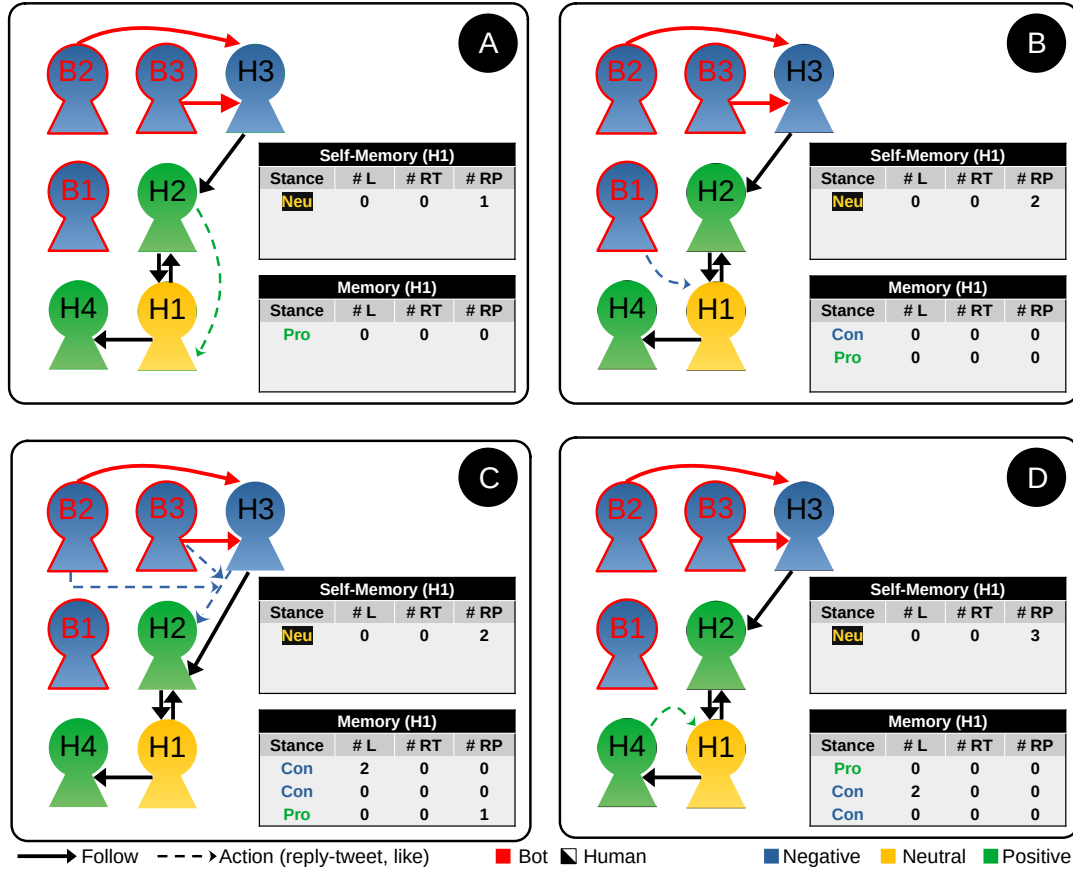


Figure 5.1: At time-step 1, a neutral human agent (H1) makes a tweet (action not depicted in this figure) expressing a neutral stance and themes of medical authority and individual liberty over collective good, e.g., “My doctor tells me wearing masks is a personal preference.” **(A)** While still at time-step 1, a second human agent (H2), H1’s follower, concocts a reply-tweet with matching themes but a positive stance, e.g., “The CDC said masks protect everyone, not just the wearer.” **(B)** A negative bot (B1) calls `poll_tweets`, which is equivalent to doing a search of existing tweets, finds H1’s neutral tweet, and replies with a negative tweet. **(C)** A third human agent (H3) makes a reply-tweet with a negative stance to H2’s earlier reply to H1. Two bots (B2 and B3) liked H3’s reply-tweet. H3’s reply happens to be seen by H1 because H1, as a follower of H2, occasionally reads the replies to H2’s tweets. **(D)** A fourth human agent (H4) makes a tweet with a positive stance that is seen by H1, which pushes the very first tweet seen out of H1’s limited memory. **(E, not shown)** At the start of time-step 2, when `read_conversion` is called, H1 observes one positive-stanced tweet without any engagement, two negative-stanced tweets, one which received two likes, and finally its own neutral-stanced tweet that received three replies. To determine if H1 will experience a stance conversion after reading these tweets, we check if a generated uniform random number clears H1’s resistance threshold. The number does. To see which stance from the seen tweets infect H1, a weight is calculated for each stance. The negative stance has a heavier weight than neutral or positive because the negative tweets are greater in number and have affirmative engagement instead of negatory (likes instead of replies). Weighted random choice picks the negative stance, H1 changes its stance to negative.

For all actions, the part that involves broadcasting a stance — i.e., *writing* a reply, a tweet, or a quote-tweet — the broadcasted stance is always the same as the agent's. Themes for replies and quote-tweets always match the themes of the parent tweet; the assumption is that when responding to another, normal users would not go off-topic. Themes for an independent tweet are drawn from the agent's own pool of themes, with themes that have been recently used by an agent having greater probability of being chosen. For the part of an action that does not explicitly broadcast a stance, i.e., *selecting* a tweet to like, retweet, quote, or reply to, the agent will favor tweets with stance and themes that align with its own, but the agent is also influenced by engagement metrics to pick tweets with different stances and themes.

To be infected, a uniform random number must first exceed an agent's resistance threshold. A high resistance results in fewer infection events being triggered. If the threshold is cleared, one stance among the tweets an agent has read will be selected as the infecting stance. The selection is random, but each stance is weighted by the engagement metrics (Section 5.3.3.1) and by the compatibility of the tweets representing the stance. The compatibility calculation factors in themes as well, e.g., for a neutral agent (equidistant from both poles), a group of tweets representing the positive stance that has many matching themes will influence the agent more strongly than a similar-sized group of negative tweets with non-matching themes, all else being equal. A successful infection results in the agent's adoption of not just the infecting stance but also a portion of the most popular themes found in the tweets representing the infecting stance. Examples of the infection mechanism at work can be found in Figure 5.1 (the influence of stance and theme compatibility is omitted in this example) and in Section 5.6.5.

Although we have called infection as stance “change” or “conversion” throughout this paper, an infection event can end with an agent being infected by the *same* stance. This situation can be thought of as witnessing strong arguments that affirms one's faith in one's stance.

Resistance is altered by infections. If the infecting stance is the same as the one held by an agent at the start of an infection event, the infection will improve resistance. Infections by non-matching stances degrade resistance. This mechanism is inspired by certain viruses such as the Hepatitis C virus (HCV) whose infection some people can clear, whereas in certain other population segments they induce an impaired immune response thereby allowing reinfections or chronic infections [80, 156]. More domain-appropriate inspirations are the observed instances of people slowly doubting their worldview if some strongly held beliefs have been shattered (e.g., South Korean feminists “taking the red pill”[232], spiral of mistrust/cynicism [175]).

5.3.5 Bots

Bots are propagandists, biological (Section 5.2.2.5) or electronic, dedicated to promoting a stance on a social network. The key differences between bots and humans are: a bot's stance is fixed, a bot never unfollows another user, a bot constantly surveys the conversation landscape, and a bot has a greater memory capacity. The set of heuristics that a bot follows also differ from that of humans as they are designed to spread their stance as widely as possible. None of the differences should be interpreted as bots having superhuman abilities or access to tools on social media platforms that regular users do not. The difference is purely a result of motivation and prioritization. Bots are just as limited as humans in terms of their activity levels and roles.

A bot surveying the conversation landscape, `poll_tweet`, at the start of every time-step is equivalent to running a search or covertly reading tweets from accounts that it does not follow. A fixed number of the most recent tweets are returned with each survey attempt and these tweets are saved into the bot's memory of seen tweets. Surveying allows bots to take note of tweets/conversations that may have escaped the attention of human agents. If necessary, bots can jump into the conversations to influence other agents. This form of non-social tie information acquisition is exclusive to bots. For human agents, non-social tie information is acquired if they happen to be randomly selected as the followers who have read a reply to their followee's tweet (see Section 5.3.3; the mechanism applies to bots too).

Bots do not unfollow so as to avoid triggering reciprocal unfollowing among their followers. Bots want as large an audience as possible to maximize the reach of their messages. This desire for a large audience is also manifested when choosing which tweets to quote, retweet, or reply to. Unlike humans who will only consider engagement and compatibility, bots will additionally consider a tweet author's follower count, favoring those who have large counts.

Bots not changing their stance is simply a result of them being bots, i.e., unyielding propagandists. Their greater memory capacity (Section 5.3.2) is also due to them being motivated agents of change.

However, bots do adopt new themes. Since bots are conceptualized as being interested in altering *only* the stances with regard to the issue of interest, theme adoption does not compromise the bots' mission as themes are non-polarizable with regard to the issue of interest (Section 5.3.1). It may even aid in the bots' mission as bots can better empathize with the rest of the agent population, e.g., composing tweets with themes that resonate with other agents and are able to infect others more because of greater theme compatibility.

When bots have to choose which tweet to reply to, they have two tactics — targeting vulnerable users and targeting high engagement tweets — that they alternate between. A probability distribution governs how often one tactic is used over the other. To target vulnerable users is to target tweets from users who have a low number of followees. The reasoning for this tactic is that such users have low number of information sources (followees) and are therefore more easily swayed by a flood of information from non-followees. To target high engagement tweets is to target tweets whose engagement metrics are the highest among those that a bot has witnessed in its memory and are authored by users with high follower counts. This tactic maximizes the visibility of a tweet as high engagement tweets are likelier (due to bandwagoning) to be targeted by agents for actions that trigger information sharing such as replying, retweeting, and quote-tweeting. This tactic has real-world precedence, e.g., BTC scams [321].

Replies and quote-tweets authored by bots have the same stance as each bot’s respective stance, just like humans, but the process of themes selection differs from humans. While humans would use the same themes as the parent tweets, bots would only use half of the themes from parent tweets, while the other half is drawn randomly from each bot’s pool of themes. In the case where a bot’s themes deviate significantly from the population, this tactic serves to subtly inject the bot’s themes into the information environment to facilitate infection in the future, while infection chances at present are improved by adopting half of the same themes as the parent tweet. Theme selection for independent tweets also differs from humans, as a bot does not select based on a history of its own tweets. Instead, following the same selfish/motivated reasoning as when replying and quote-tweeting, half of the themes are the most common themes found in a bot’s seen tweet memory while the remaining half are randomly chosen from the bot’s pool of themes.

Like humans, bots start out embedded in a network instead of starting out in the network’s periphery. Section 5.2.2.6 explains the reasoning behind this positioning. The bots in our simulation are bound by the same diurnal activity schedule as the one humans follow. This is informed by the activity data of the covert pro-Western info ops’ assets, which have peaks and troughs [192] that mimic a human’s. Motivated by decentralization (Section 5.2.2.4), the bots act independently (just like human agents) and are not conscious of the presence of other bots. Naturally, this excludes explicit coordination between the bots.

Bots operate the same across stances within *Diluvsn*, unlike other works that split propagandists based on stance. For instance, the “bad guys” bots/trolls class of agents in *twitter_sim* actively spread disinformation while the “good guys” stiflers in *twitter_sim* [57] are reactionary by design, only countering disinformation that they see and never

proactively spreading truth.

5.3.6 Temporal resolution

Humans follow a diurnal cycle, and their inclination towards being active on social media follows a similar cycle [194]. Therefore, the human and bot agents in our model also observe this cycle. The mean number of tweets from a dataset of mask-wearing conversations on Twitter, binned by the day of the week and the time of the day, gave us an action-probability table. This table approximates the likelihood of an average Twitter user using the service at a particular time-step. Within each time-step, the order in which the agents activate is random. A consequence of simulating peaks and troughs in platform activity levels is that during periods of high activity, information dissemination and actions to boost dissemination will reach the highest number of awake users. However, at the same time, the great volume of information flowing during high activity periods means that disseminated information risks being pushed out of the memory (Section 5.3.2) of agents with high numbers of incoming links.

The resolution of the simulation is 30 minutes. This resolution is sufficient and has empirical support: the Graphika and SIO investigation found that the info ops “assets in the Afghanistan and Central Asia groups typically posted at roughly 15-minute or 30-minute intervals in any given hour” [192]. Our decision was also influenced by the mean weekly frequency (the number of actions taken by a user in a week) for the high-activity group in the study by Antelmi et al., which was found to be 133.84 [25]. If we took 133.84 actions and divided them over seven days, this leaves approximately 19 actions per day. At a resolution of 30 minutes, 19 actions will take 9.5 hours, including the interval time between actions. This wakefulness period does not violate typical sleep-wake cycles within human circadian rhythms [160].

As validation, when we generated 1,000,000 sets of 336 random numbers (2 periods per hour $\times$ 24 hours $\times$ 7 days) with a 0.9 damping factor⁸ and tested them against the activity-probability table (derived from our mask-wearing dataset), the maximum number of periods/time-step in which an agent is awake was 153.45 ± 8.46 per week, a range which can accommodate the maximum number of actions the most active class of agents can take per week within our model (we let this maximum be the rounded mean number of actions of high activity users, 134, from the dataset in [25], which is different from the mask-wearing dataset used to generate the action-probability table). To allow for variability,

⁸When multiplied by probabilities in the action-probability table (Section 5.3.8.1), which can range from 0 to 1, the factor shifts the graph down/reduces the amplitude, i.e., lowers the action probability.

once an agent exhausts its allotted action quota, if a random number check against the action-probability table says that the agent is awake and can therefore take an action, the agent has a fixed 1% chance of exceeding the quota.

5.3.7 Graph structure

Directed scale-free graphs were used as the simulation environments. Specifically, we used the improved version [423] of the algorithm first introduced in [424] by Schweimer et al. for generating random directed social network (SN) graphs for use in information diffusion studies. The hyperparameters used for graph generation comes from the hyperparameters file in the open-source code repository shared by the author⁹ which themselves are derived from crawling real-world social networks on Twitter. The set of hyperparameters used were named “Covid”, the largest of the 14 networks crawled with 50,133 nodes and 4,832,226 edges.

Desirable network properties such as a high amount of clustering and reciprocity are reproduced by the SN graph generation algorithm. Clustering describes users aggregating into homophilic clusters when interacting online on Facebook and Twitter [107]. Reciprocity describes the small but still significant amount of reciprocal edges on social media (22.1% found on Twitter [269]). If we used non-SN scale-free graph generation algorithms, e.g., the one proposed in [65] by Bollobás et al. which improved upon the Barabasi-Albert method, certain properties of the network deviate further from expectations than if an SN graph generation algorithm was used (Section 5.4).

A total of 20 graphs with 1,000 nodes and 20 graphs with 900 nodes were generated. Because of an inherent flaw [424] of the graph generation algorithm itself, approximately 2.5% of the nodes for each graph are isolates, i.e., the nodes have no incoming or outgoing edges. The graphs were saved and reused for different scenarios to allow for reproducible results and also because graph generation is time consuming. The graphs with 900 nodes were used for the “peripheral”/“fringe” scenarios where bots (10% of the network population, equivalent to 100 agents) do not start out embedded in the network but is added to the graph later and given only one outgoing edge to a random node, i.e., a bot’s only social tie is being a follower of an existing human user.

⁹https://github.com/Buters147/Social_Network_Graph_Generator

5.3.8 Parameterization

The same set of parameters are used for validation (Section 5.4) and simulations (Section 5.5).

5.3.8.1 Empirical data

To ensure that agents populating the network mimic the behavior of those found in real-world social media, we initialize the agents based on statistical data collected from Twitter users.

For the diurnal activity schedule, we took the mean number of tweets split by day of the week and into half-hour intervals (the resolution of the simulation) from the two million tweets we collected on COVID-19 masking debate between January 1–June 21, 2020 as a probability distribution for when an agent will be active.

The starting distributions of stances and themes, including the starting number of themes per agent, are taken from the February 27–March 2, 2020 period of the COVID-19 masking debate, which we have identified as the starting point for when the majority of the discussions occur within the appropriate context of masking as protection against COVID-19 instead of masking in other contexts, e.g., masks to protect against volcanic ashfall. A total of 12 themes are identified from the dataset. Examples of the themes are “encourage” (encouraging people to mask or to not mask) and “appraisal-critize” (praising others for wearing or for not wearing masks).

Although the proportions of stances and themes at the population level are maintained, the stance and the themes chosen for an individual human agent are entirely random, unaffected by the agent’s neighbor’s stance and themes. Some may take issue with our choice since homophily would predict that users sharing similar outlooks, political or otherwise, are more likely to be connected, so there should be homogeneous stance and theme clusters in the network. For emerging issues where strong narratives have not been established, the situation is not so clear-cut. Republicans are more anti-mask during the COVID-19-pandemic than Democrats [163, 265], but the Twitter data on mask-wearing that we collected showed that right before the pandemic, many fans of Trump supporter Scott Presler were urging him to wear a mask to avoid catching disease from homeless people.

The counts of users found in the high, medium, and low activity categories are taken from [25]. For the range of actions available to the agents (Section 5.3.3) except for following and unfollowing, the frequency distribution is based on the tables and figures in [25]. In [25], the authors categorized users into five roles — Tweeters, Quoters, Retweeters,

Repliers, and Likers — with a corresponding action frequency distribution. Each role is further sub-divided into high, medium, low, and no activity. Although almost 50% of users belong to the no activity cluster (no visible activity but have done at least one hidden activity such as changing their screen name), we exclude this category because they would have no appreciable impact on the simulation dynamics owing to their inactivity. This leaves us with $5 \times 3 = 15$ human agent categories.

Statistics on following and unfollowing are not found in [25]. We relied on the descriptions provided in [270, 354]. According to Kwak et al., the “average of unfollows is 15.4” and “average number of followees increases from 59.7 to 75.7” in 51 days with “about 30% of users [having] unfollowed at least once” [270]. According to Myers and Leskovec, “[i]n a given month a user of degree 100 tends to gain 10 and loose [sic] 3 followers” [354]). Their descriptions showed that while following and unfollowing are pervasive over a scale of multiple months, they are rare when one factors in all the possible actions that a user can take (tweeting, retweeting, etc.) at the scale of half-hour steps which our simulation runs at. Modeling the actions of following and unfollowing as suitably rare events translates to having follow and unfollow constituting $0.\bar{6}\%$ and $0.\bar{3}\%$ of all of an agent’s actions, respectively. This applies to all 15 agent categories. Given that highly active and moderately active agents form 10.4% and 22.2% of the population (totaling $\sim 30\%$) and their actions per week are 134 and 66, respectively, we expect highly active agents to follow $134 \times 0.66\% \times 4 = 3.5$ users and unfollow 1.8 users over 28 days while moderately active users would attempt 1.7 follows and 0.9 unfollows, which is slightly lower than the frequency observed in [270].

The parameterization of activity levels and roles used statistics from [25] because it cannot be done using the mask-wearing dataset. Only tweet-level data were collected for the mask-wearing dataset; user-level data were absent. However, this does provide us an opportunity to validate a simulation using parameters from the dataset in [25] against an “unobserved” dataset (the mask-wearing dataset), as we shall see in Section 5.4, which is a more informative validation process than using the same dataset from which the parameters are drawn.

5.3.8.2 Best approximations

Certain parameters in our model require information that the existing literature does not provide and that we were unable to collect. The resistance to stance conversion is one such parameter. We opt to model the resistance as a truncated normal distribution with a low of 0.5, a mean of 0.75, a high of 1.0, and a standard deviation of 0.25. The size of a human agent’s memory is fixed at $m = 30$, and this upper limit is used for three separate

memory bins, which are the agent’s own recent tweets, tweets by others that an agent has seen, and the author of those seen tweets. Bots have a memory of 120 and `poll_tweet` returns 30 tweets each time. The number of themes that every human and bot agent can hold at once is five while the maximum percentage of new themes that an agent will adopt is 40% for humans and bots. Humans adopt themes upon a successful infection, whereas bots adopt themes at every time-step. All of a user’s tweets are assumed to propagate to all of the user’s followers, although not all tweets are guaranteed to be seen as older tweets risk being pushed out of a follower’s limited memory by newer tweets. The proportion of a user u ’s followers who see a reply from another user v to a tweet from u , which is the primary method of non-social tie information propagation in our model (there are other methods such as u retweeting a seen tweet from a non-followee user), is set at 10%, with u ’s followers being chosen at random.

Every infection event by a stance matching a human agent’s own increases the agent’s resistance by 1% while an infection by a non-matching stance decreases the agent’s resistance, with the decrease being proportional to the distance between the stances, -0.5% for the closest neighboring stance and -1% for the stance on the opposite pole.

Stance compatibility is simply the distance between two stances divided by 2, $c_s = |s_{\text{user}} - s_{\text{tweet}}|/2$. Division by the maximum distance between stances, 2, ensures that the maximum value of c_s is 1. Themes compatibility is the number of matching themes divided by the number of unique themes, $c_t = |t_{\text{user}} \cap t_{\text{tweet}}|/|t_{\text{user}} \cup t_{\text{tweet}}|$. Compatibility, $c = c_s w_s + c_t w_t$ where w_s is the weight for stance compatibility and w_t is the weight for themes compatibility. The default values for $\{w_s, w_t\}$ are $\{1, 0.8\}$ so having matching stances is more important than matching themes. Note that since incompatibility, not compatibility, in stances increases the chance of replying, the compatibility score when replying is calculated slightly differently, $c = ((1 - c_s) \times w_s) + c_t w_t$. Weights are also modified when replying, $\{w_s, w_t\}$ are $\{0.8, 1\}$ to mildly de-emphasize stance compatibility, thus preventing replies from being exclusively an act of disagreement with a tweet’s stance.

When deciding on which tweet to reply to, a human agent will prefer to reply to tweets that have the same themes but a different stance, giving less consideration to the amount of engagement that a tweet has received. When deciding which tweet to retweet, human agents will prefer tweets with matching stance and themes, with a tweet’s engagement having slightly less weight on the decision. Quote-tweets emphasize theme compatibility over stance compatibility when choosing a target tweet to quote, and a slight importance is placed on compatibility over engagement just like when retweeting. The probability score of a tweet being liked, retweeted, quoted, or replied to, $p = gw_g + cw_c$ where w stands for weight, g for engagement, and c for stance and theme compatibility. The weights $\{w_g, w_c\}$

are $\{0.5, 1\}$ when replying, $\{0.8, 1\}$ when retweeting and quote-tweeting, and $\{1, 1\}$ when liking. Engagement is a weighted sum of metric counts, $g = N_{\text{like}}w_{\text{like}} + N_{\text{RT}}w_{\text{RT}} + N_{\text{reply}}w_{\text{reply}}$. The weights used differ based on the action taken (e.g., in deciding whether one should like a tweet, the bandwagon heuristic would more pay more attention to the number of likes the tweet has instead of the number of replies). The weights for the engagement *metrics*, $\{w_{\text{like}}, w_{\text{RT}}, w_{\text{reply}}\}$, for the engagement *actions* of liking, retweeting, quoting, and replying are $\{1, 0.2, 0.2\}$, $\{0.2, 1, 0.2\}$, $\{0.2, 1, 0.2\}$, and $\{0.2, 1, 0.2\}$ respectively. Quote-tweeting and retweeting have the same weights as both actions expose one's followers to another's tweet. Both g and c are normalized by dividing over the values summed over all candidate tweets, e.g., for a tweet x , $g_x = g_x / \sum_{n=1}^{N_{\text{candidate}}} g_n$.

Bots differ from humans in that their probability for choosing a target, $p = g + c + f$, meaning that the probabilities are unweighted and an additional term f has been added for the follower counts of the tweet authors. Like g and c , f is normalized. And instead of the probabilities being used in weighted random choice, they are used to sort the candidate tweets and the tweet with the highest probability is always selected. Bots also weigh each metric in engagement g and the weights, $\{w_{\text{like}}, w_{\text{RT}}, w_{\text{reply}}\}$, for liking, retweeting, quoting, and replying are the same as those of humans.

Within the sub-component of the `read_conversion` algorithm for computing a human agent's stance changes, the compatibility score, r_c , for each stance (e.g., $r_{c\text{-pro}}$) is the sum of the individual compatibility scores for the subset of an agent's seen tweets that exhibit the stance. The same applies to the engagement metrics' scores, r_g . The compatibility scores are normalized by dividing over the total score for messages of all stances. The engagement scores are normalized on a per-engagement type basis by dividing over the total of the frequencies for an engagement type. Prior to summation, the engagement metrics are weighted, $\{w_{\text{like}}, w_{\text{RT}}, w_{\text{reply}}\} = \{1, 0.5, -0.5\}$. The weighting reflects whether an engagement action is affirmatory or negatory in the context of stances. The summed compatibility scores and engagement metrics' scores are also weighted to avoid stance conversion from being entirely dominated by engagement metrics. The weight for compatibility is an agent's memory, m , while the weight for engagement is the reciprocal, $1/m$. As such, the final conversion score for a stance, e.g., pro , $w_{r\text{-pro}} = m^{-1}r_{g\text{-pro}} + mr_{c\text{-pro}}$ where $r_{g\text{-pro}} = \sum_{x=1}^{N_{\text{pro}}} g_x$ and $r_{c\text{-pro}} = (\sum_{x=1}^{N_{\text{pro}}} c_x) / \sum c$. The engagement score for a tweet x is $g_x = w_{\text{like}}N_{\text{like-}x} / \sum N_{\text{like}} + w_{\text{RT}}N_{\text{RT-}x} / \sum N_{\text{RT}} + w_{\text{reply}}N_{\text{reply-}x} / \sum N_{\text{reply}}$. The conversion scores are used as probabilities for selecting the stance which a human agent would convert to if its resistance threshold has been exceeded.

5.4 Validation

Validating information diffusion models can take the form of showing that the model is able to generate statistical distributions similar to those in the real world, e.g., recreating the percentages of tweets that receive zero engagement and tweets that go viral by adjusting parameters [32], and matching curve shapes for message volume [126]. Another method is confirming that the parameters used in the model match those found in the real world, e.g., Twitter usage inter-arrival time in the `twitter_sim` model [57] by Beskow and Carley.

As explained in Section 5.3.8.1, a large number of key parameters in our model such as the distribution of a user’s actions, a user’s degree of activity (cognate to inter-arrival time), and the distribution of user roles are taken directly from real-world data. The underlying graph structure on which the simulation is run on (Section 5.3.7) and the user’s diurnal schedule (Section 5.3.6) are also based on real-world data. Our model would naturally reproduce inter-arrival time, action distribution, etc., that are close (though not perfect matches) to empirical observations. Therefore, we have “validated” our model in the same sense that Beskow and Carley have.

For additional validation, we compared the engagement statistics obtained from simulations with those obtained from real-world data, namely the Twitter on mask-wearing conversations pertaining to the COVID-19 pandemic (Section 5.3.8.1). Differences and similarities between statistics collected empirically and from the simulation are shown in Figure 5.2. The simulations are run for a period equivalent to 28 days and data was aggregated from 20 runs. Parameters for the agents are the same as those described in Section 5.3.8. The configuration for the human-only simulation scenario is equivalent to the Sim 11 scenario described in Section 5.5.

There are two sets of real-world distributions from the mask-wearing dataset in Figure 5.2 as we intend to illustrate how distributions can differ depending on how hotly a topic is contested. Each set of distributions is from the complete month-long period of the named month. The January 2020 data are from a less popular and less contentious period as it is composed of tweets that have been classified as out-of-topic/containing no stances, i.e., they are not offering any opinion on mask-wearing as it relates to COVID-19, but they are discussing masks in other contexts (e.g., protection against volcanic ash, fictional characters). In contrast, the March 2020 distribution is a subset of only tweets that have stances. March is also the period we have identified as when the mask-wearing conversation occurred almost exclusively in the context of COVID-19.

The January distributions, from a subset where masks are a less contentious and less discussed topic, all contain lower values than the March distributions, i.e., shifted

left. The distributions from the simulation in Figure 5.2 nearly overlap the empirical January distribution for likes and replies. The simulation distribution is visibly shifted right compared to the January distribution when comparing retweets, although it is still bound to the right by the March distribution. For simulations, the within-dataset ratios $N_{\text{like}} : N_{\text{RT}}$ and $N_{\text{like}} : N_{\text{reply}}$ are on average 2.0 and 8.4 respectively, close to the numbers calculated directly from the data from which the simulation parameters are drawn (sans simulation), which are 2.1 and 8.5. If we had instead used a non-social network-specific graph generation algorithm (Section 5.3.7), the discrepancy between the simulation and empirical ratios would have been starker, as the simulation ratios obtained would have been 1.7 and 5.2, validating our choice in the graph generation algorithm. January mask-wearing data shows ratios of 3.8 and 22.5, while March's ratios are 3.3 and 13.2. Retweets and especially replies are intensified when a topic is contentious (March), hence the lower ratios.

One factor involved in the discrepancy in distributions and ratios is the insufficient number of tweets at simulation start for agents to like, retweet, etc. We alleviated tweet scarcity by having an agent default to tweeting as an action if there are no incoming tweets in its memory and whenever a chosen action for a time-step is invalid. A consequence of this design choice is that the observed action distribution of agents would differ from the parameters they were given, as they tweeted more often than they should. However, the alternative of doing nothing if a chosen action for a time-step is invalid would have caused a similar issue of mismatch between parameters and observed distribution of actions taken, in addition to causing an agent to appear less active (due to skipped actions) and causing the starting phase of the simulation to last longer (due to tweet scarcity). We had considered seeding the simulation with a set of initial tweets but decided against it as the assumptions and ramifications — heuristic for distributing the initial tweets to the nodes, the tweets causing stance changes in agents immediately at simulation start, etc. — would complicate the analysis of the simulation results.

Another partial explanation for this difference is the size and structure of the networks. Simulations relied on a synthetic scale-free network as the graph structure in terms of the follower-followee network of the mask-wearing dataset is unknown to us. When compared with the January dataset, the across-dataset ratio for the counts of likes, $N_{\text{like-Sim}} : N_{\text{like-Jan}}$, is 2.7, that of retweets, $N_{\text{RT-Sim}} : N_{\text{RT-Jan}}$, is 1.4, and that of replies, $N_{\text{reply-Sim}} : N_{\text{reply-Jan}}$, is 1.0. When compared with March, the respective ratios are 30.3, 18.0, and 19.2. If we infer the hidden size of the network through the engagement counts (hidden size because users in the mask-wearing dataset cannot represent the totality of network; metrics for a tweet can originate from a user outside the collected dataset), the number of users in January could be at least $2.7\times$ larger than in the simulation and for March, $20.3\times$ larger.

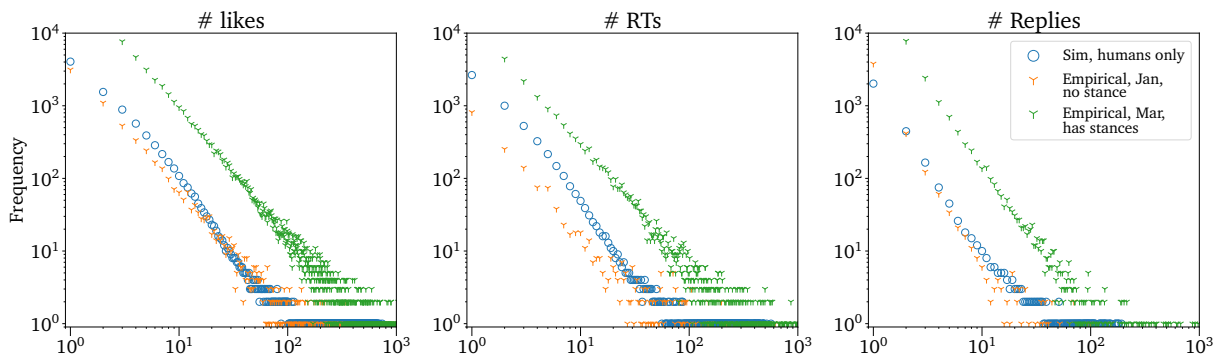


Figure 5.2: Engagement metrics: real-world data vs simulations with 1,000 agents.

The across-dataset ratio of replies as well as the within-dataset ratios indicate that the simulations show a greater affinity to March than January data. For all engagement types, a constant multiplier of $\sim 1.5\times$ reproduces March’s ratios for within-dataset ratios, whereas while the across-dataset ratios for March’s ratios are a constant $\sim 20\times$ of the simulations except for $N_{\text{like-Sim}} : N_{\text{like-Mar}}$. January’s multipliers differs depending on engagement type for across-dataset ($\sim 2\times$ and $\sim 3\times$) and within-dataset ($\sim 3\times$, $\sim 1.5\times$, and $\sim 1\times$) cases. This makes the simulation a closer approximation of conditions where a topic is heavily debated, i.e., March’s conditions.

Overall, the distributions for the engagement actions strongly resemble real-world data, owing to the real-world statistics used to initialize simulation parameters and the robust SN-graph generation algorithm that also relied on empirical data. As the indirect influence from engagement actions and the metrics resulting from those actions are central to *Diluvion*, the findings in this section supports the model’s validity.

5.5 Simulations

We present a number of simulations, starting from the baseline scenario of bots representing two disadvantaged stances engaged in a co-opetitive influence operation and continuing with simulations where certain key functionalities of the bots are progressively altered/degraded, which can be thought of as ablation studies. Other scenarios tested the effect of varying the mixture of bot-backed stances. The results from the simulations are shown in Table 5.1 and we grouped our discussion of the results in Section 5.6 based on their implications for conducting real-world info ops. All scenarios featured 1,000 agents in scale-free networks and the results were averaged over 20 runs, with each run lasting 1344 steps (includes step 0, equivalent to 28 days). A bot’s role determines its action distribution,

i.e., the probabilities for liking, retweeting, etc. at each time-step. A role is named after the most frequent action the role takes, e.g., Likers like liking. Bot frequency is the number of times a bot is permitted to act per week. The definitions for both roles and frequencies (activity levels) are based on the Twitter data reported in [25], and the definitions apply to bots and humans.

In classical influence maximization (Section 5.2.1.3), information that we wish to see widely propagated is injected via nodes that have been algorithmically identified as influential, i.e., positioned to maximize spread, but the bots in our simulations have no such advantage as their positions in the network are entirely random. Starting from the periphery of a network is the most disadvantaged position our bots are subjected to (Sims 6 and 7).

Inter- and intra-stance awareness of the presence of other bots is absent as the info ops is conducted in a decentralized fashion (Section 5.2.2.4). Competition is implied as each agent can only belong to one stance so an agent's conversion from one stance to another directly results in losing an agent to the other stance group; bots from one stance are not forbidden from converting agents belonging to a different bot-backed stance. Cooperation amongst agents arise naturally because agents gravitate towards performing affirmative actions, e.g., liking, for other agents that share similar a stance. Cooperation among two bot groups, each backing a different stance, can be implicit in that their uncoordinated and even competitive actions can accomplish a common goal better than they can individually, e.g., reducing support for a third non-bot-backed stance.

A high bot percentage, 10%, is used in all scenarios so that the impact from bots is more easily apparent, not because this percentage is reflective of the actual bot population in the real world. Estimates of bot percentages on Twitter vary widely depending on the source, with Twitter's own analysis claiming 5.3% in 2022, a 2016 Securities and Exchange Commission filing by Twitter claiming 8.5%, Cyabra claiming 11% in 2022, and a study on the 2020 Singaporean election claiming 26.99% [144, 450, 471]. Bots defaulting to a high activity level also makes their influence more easily observable. Additionally, evidence exists that bots are prolific content creators on Twitter, with the Israeli web analytics company Similarweb finding that "20%–29% of content in the US on Twitter is generated by bots" [474].

To ensure that the overall stance ratio at simulation start matches the one at the start of March 2020 in the mask-wearing dataset, the stance ratio for humans is rebalanced to account for the bots' stance distribution. In other words, the number of human agents assigned to a particular stance is reduced if bots are assigned to that stance.

As Sim 1 is the baseline, Sim 1 is referenced by all other scenarios for comparison in the

scenario descriptions below:

1. Sim 1: Bots belonging to two minority stances, neutral and opposition, attempt to maximize the spread of their respective stances. The bot population is evenly split between neutral and opposition, and there are no bots in the camp of supporters, making the bot population's stance ratio 1 : 1 : 0 (pro : neutral : con). This stance distribution enables us to observe synergistic and antagonistic effects between the two bot sub-populations. For all agent types, the number of themes an agent has as well as the distribution of the 12 themes are the same as in the March mask-wearing dataset. Bots start out embedded in the network, i.e., they have organic follower-followee relationships just as human agents do for reasons we have argued in Section 5.2.2.6. All the bots take on the role of a Replier to better demonstrate non-social tie spreading of information (Section 5.3.3.2). Bots are as active as highly active humans are. In choosing whose tweets to reply to, bots will favor those designated as vulnerable, which are people with a low number of information sources, i.e., low followee count. Bots will adopt commonly encountered themes to help the tweets they compose appeal to a wider audience.
2. Sim 2: Bots will prioritize tweets that have high engagement metrics and that happen to also be authored by people with high follower counts when deciding whose tweet to reply to instead of targeting tweets from vulnerable agents as in Sim 1.
3. Sim 3: Bots are as active as moderately active humans rather than matching highly active humans' frequency.
4. Sim 4: Instead of being Repliers, bots are Likers.
5. Sim 5: Instead of being Repliers, bots are Retweeters.
6. Sim 6: Instead of being embedded in the network, each bot starts out isolated in the fringes of the network with only one followee relationship with a random human agent in the network (meaning that any tweet by the bot that is not a reply or quote-tweet will not be read by anyone in the network). This replicates the difficult starting condition that bots face at the start of an info-ops, as presumed by other researchers [57] but which we have argued against in Section 5.2.2.6.
7. Sim 7: In addition to starting out in the fringes of the network as in Sim 6, bots take on the role of Likers instead of Repliers.

8. Sim 8: Instead of being highly active Repliers, bots take on a variety of roles (Likers, Retweeters, etc.) with variable activity levels: low, medium, and high, just like human agents. The role and activity level distributions of the bots follow the distributions in [25].
9. Sim 9: Bots conserve a theme, Theme 7, instead of following the usual theme adoption process. Conservation involves a set of conserved themes that always have priority whenever the bot updates its list of themes to adopt and whenever the bot needs to decide which themes its tweet should be accompanied by.
10. Sim 10: Just as in Sim 8, bots mimic human agents' roles and activity levels. Furthermore, the distribution of the bots' stances are the same as the humans' stance distribution from the March 2020 mask-wearing dataset.
11. Sim 11: All agents in this scenario are humans, unlike other scenarios that contain 10% bots.
12. Sim 12: Bots are evenly distributed between supporters and detractors instead of being evenly distributed between neutrals and detractors, 1 : 0 : 1 (negative : neutral : positive).
13. Sim 13: Bot distribution follows a 1 : 0 : 0 (negative : neutral : positive) ratio.
14. Sim 14: Bot distribution follows a 1 : 1 : 1 (negative : neutral : positive) ratio.
15. Sim 15: Bot distribution follows a 2 : 2 : 1 (negative : neutral : positive) ratio.
16. Sim 16: Bot distribution follows a 4 : 0 : 1 (negative : neutral : positive) ratio.

5.6 Results and discussion

The simulation results are shown in Table 5.1. Figure 5.3 shows the temporal evolution of the stances for the simulations in Table 5.1 while Figure 5.4 shows the temporal evolution for follower and followee counts, split by stances. Figures 5.5, 5.6, 5.7, 5.8, and 5.9 are for our discussion on the consequences of bots starting out embedded in the network versus its periphery (Section 5.6.5).

A note on the presentation of the results: not all models of information diffusion are strongly deterministic, which is why some authors, e.g., [409], opted to instead show the number of simulations in which the bot-backed stance becomes dominant instead of

		Sim 1	Sim 2	Sim 3	Sim 4	Sim 5	Sim 6	Sim 7	Sim 8
Stance ratio (bot)		1 : 1 : 0	1 : 1 : 0	1 : 1 : 0	1 : 1 : 0	1 : 1 : 0	1 : 1 : 0	1 : 1 : 0	1 : 1 : 0
Bot start		Embed	Embed	Embed	Embed	Embed	Fringe	Fringe	Embed
Bot roles		Replier	Replier	Replier	Liker	RTer	Replier	Liker	Human
Bot frequency		Hi-hum.	Hi-hum.	Md-hum.	Hi-hum.	Hi-hum.	Hi-hum.	Hi-hum.	Human
Bot reply		Vuln.	Engage	Vuln.	Vuln.	Vuln.	Vuln.	Vuln.	Vuln.
Bot adopts themes		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Start/End N (Con)	139	305	345	133*	212	101*	70*	64*	109*
Start/End N (Neu.)	129	568	551	400*	391*	159*	74*	67*	124*
Start/End N (Pro)	732	127	105	468*	397*	740*	856*	870*	767*
Start/End N (T1)	158	232	223	277	246	236	224	404*	248
Start/End N (T2)	240	240	230	259	270	310*	374*	264	303*
Start/End N (T3)	549	288	317	329	360*	430*	592*	589*	439*
Start/End N (T4)	188	244	236	263	261	271	306	246	311
Start/End N (T5)	34	204	195	189	184	181	126*	155	180
Start/End N (T6)	69	220	237	219	223	203	190	142*	172*
Start/End N (T7)	1	60	56	87	36	34	20*	15*	54
Start/End N (T8)	52	218	195	194	224	191	144*	196	206
Start/End N (T9)	26	193	221	186	176	172	126*	101*	151
Start/End N (T10)	35	187	177	193	198	157	131*	137	132*
Start/End N (T11)	26	194	207	184	185	180	108*	125*	144*
Start/End N (T12)	26	200	178*	191	190	159*	142*	105*	151*
		Sim 9	Sim 10	Sim 11	Sim 12	Sim 13	Sim 14	Sim 15	Sim 16
Stance ratio (bot)		1 : 1 : 0	March	-	1 : 0 : 1	1 : 0 : 0	1 : 1 : 1	2 : 2 : 1	4 : 0 : 1
Bot start		Embed	Embed	-	Embed	Embed	Embed	Embed	Embed
Bot roles		Replier	Human	-	Replier	Replier	Replier	Replier	Replier
Bot frequency		Hi-hum.	Human	-	Hi-hum.	Hi-hum.	Hi-hum.	Hi-hum.	Hi-hum.
Bot reply		Vuln.	Vuln.	-	Vuln.	Vuln.	Vuln.	Vuln.	Vuln.
Bot adopts themes		Consv.	Yes	-	Yes	Yes	Yes	Yes	Yes
Start/End N (Con)	139	320	25*	12*	196*	760*	115*	213	540*
Start/End N (Neu.)	129	536	63*	61*	8*	6*	228*	372*	7*
Start/End N (Pro)	732	143	912*	926*	796*	234*	658*	415*	452*
Start/End N (T1)	158	197*	231	312	220	219	236	243	243
Start/End N (T2)	240	218	347*	367*	246	237	248	229	241
Start/End N (T3)	549	266	427*	571*	305	303	298	290	301
Start/End N (T4)	188	214	283	355*	220	232	218	238	231
Start/End N (T5)	34	139*	190	61*	180	204	202	210	188
Start/End N (T6)	69	181*	182	151*	198	205	214	221	200
Start/End N (T7)	1	519*	29	4*	69	38	40	51	23
Start/End N (T8)	52	175*	184	138	196	190	204	206	194
Start/End N (T9)	26	160*	140*	66*	171	189	187	207	192
Start/End N (T10)	35	152*	153	56*	195	181	182	212*	194
Start/End N (T11)	26	174	124*	38*	179	190	206	193	178
Start/End N (T12)	26	157*	144*	81*	175	191	192	173*	188

Table 5.1: The number of agents holding a stance and theme(s) at simulation end averaged over 20 runs for 16 scenarios with 1,000 agents. Agents can adopt a maximum of five themes. Stance ratio is negative : neutral : positive. An asterisk (*) indicates significant difference ($p < 0.05$) compared to Sim 1. Starting number of agents for a stance or a theme is listed on the unnamed column on the left, between the column of row names and the named columns.

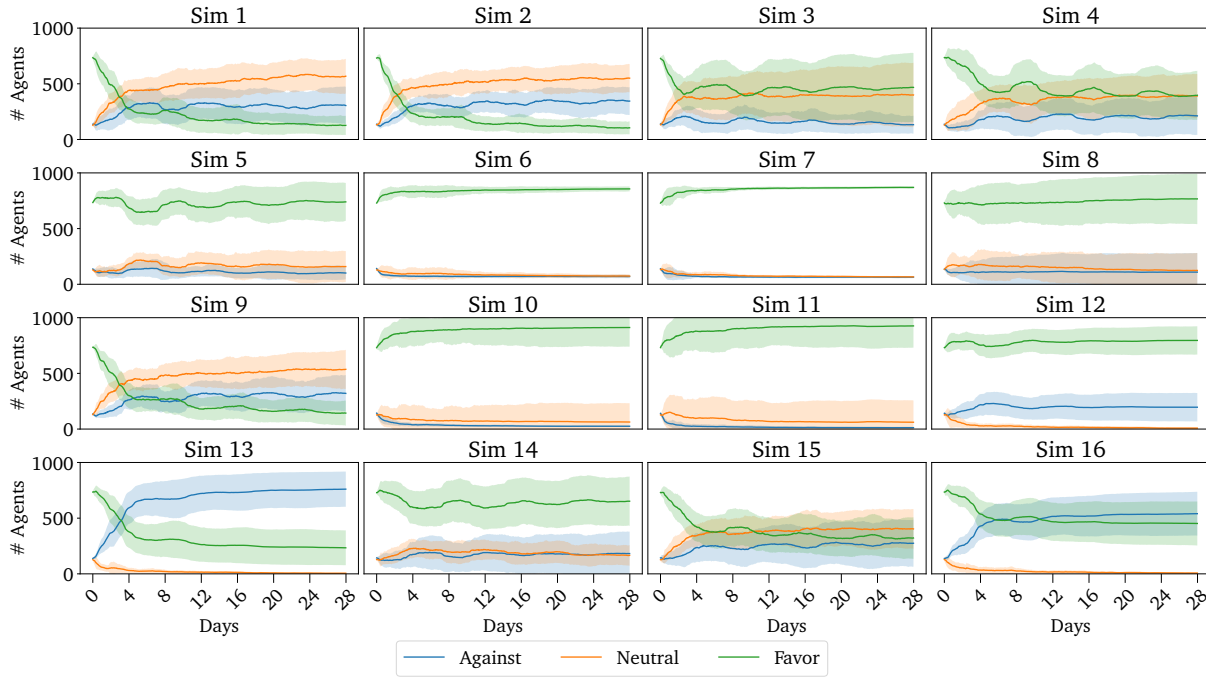


Figure 5.3: Temporal evolution of stance distribution among 1,000 agents. Values are from 20-run averages of different scenarios/simulations. Values at simulation end corresponds to the mean values reported in Table 5.1. Descriptions of the scenarios are found in Table 5.1 and in Section 5.5.

showing the average number of agents per stance at the conclusion of a simulation as we have done in this paper. However, because our simulations all have a biased start where one stance is a clear majority that outnumber the combined total of the minority stances by a factor of 2.7 to 1 (human agents’ stances are evenly split in [409]) and bots can back more than one stance per scenario (bots all back the negative opinion in [409]), reporting our results as percentages of runs in which the bot-backed stance dominates is not viable.

When we speak of a “majority” or a “minority” stance, we are referring to a stance’s share among the agent population at the *beginning* of the simulation.

Since *Diluvion* is designed to support studies on info ops, discussion of simulation results will be interwoven with discussion of their relevance to info ops.

5.6.1 Polarization

Winner-takes-all, i.e., no polarization/unipolar polarization, is the natural trend of a network where a stance has a numerical advantage over the combined total from minority stances, as witnessed in Sim 11 and Sim 10. This trend is not meaningfully altered whether the network is populated entirely by human agents (Sim 11) or the network contains 10% bots which are allocated similarly as the human population’s stance distribution (Sim 10). The

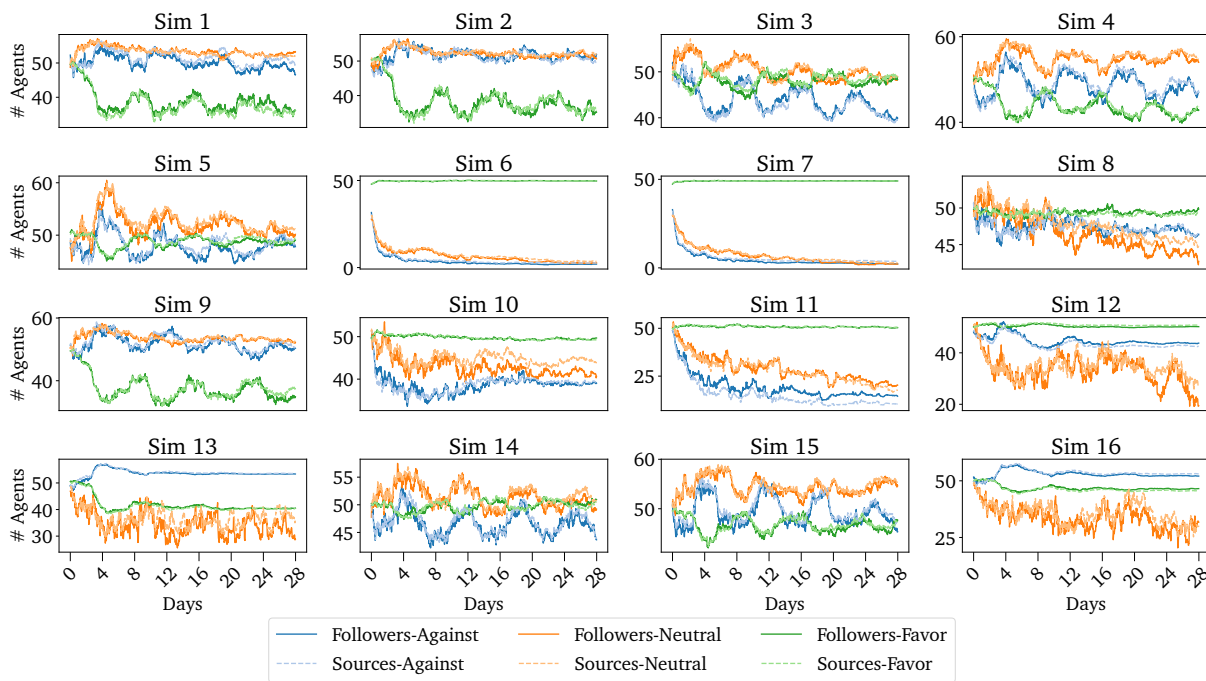


Figure 5.4: Temporal evolution of mean follower and followee (source) counts split by an agent’s stance for 20-run averages of different scenarios/simulations. When an agent converts to a stance at a time-step, its follower and source counts are included in the average for the new stance and excluded for the old stance. Subplots do not share a common range for the y-axis due to significant variance across scenarios. Descriptions of the scenarios are found in Table 5.1 and in Section 5.5.

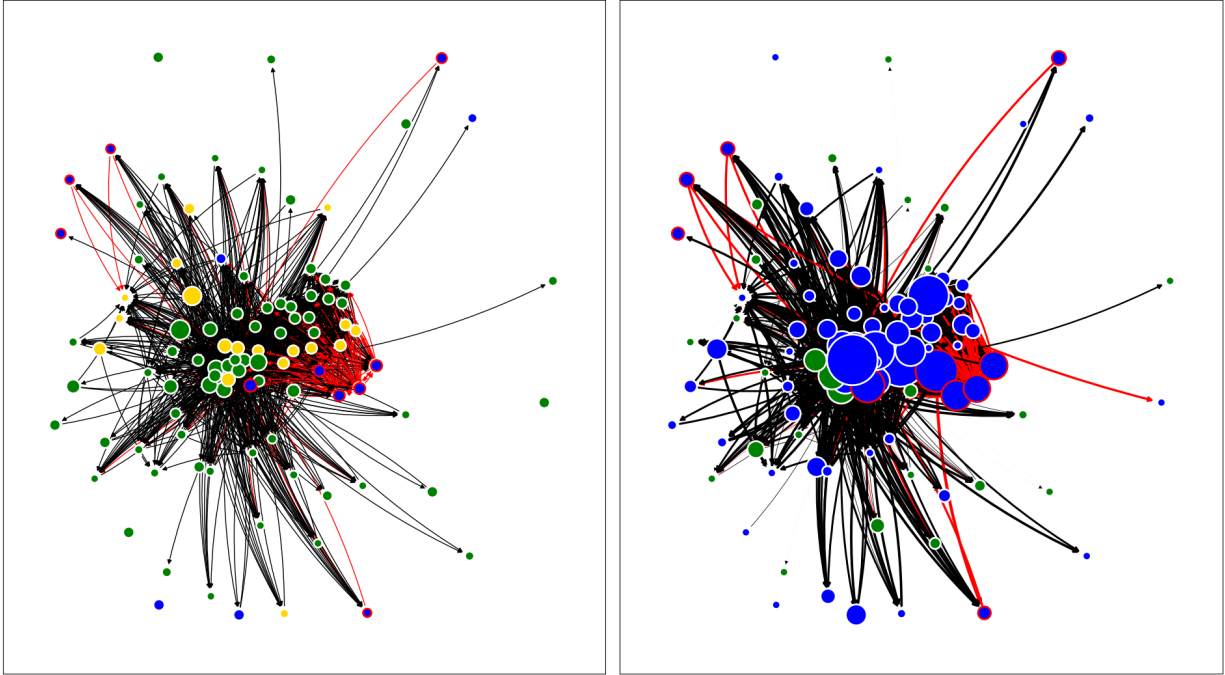


Figure 5.5: Embedded bots scenario. Each edge arrow represents the direction of information flow, i.e., the inverse of a directed follower-followee relationship, for a simulation with 100 agents in conditions equivalent to Sim 13 (Table 5.1). In standard graph notation for an edge/link (u, v) , u is a follower of v , so the follower-followee relationship is $u \rightarrow v$ but the information flow is reversed, $u \leftarrow v$. The image on the left shows the network at simulation start and has nodes sized according to the ratio of followers to followees. The image on the right shows the graph at simulation end and nodes are sized according to the average ratio of likes to replies received by the agent's tweets $\frac{1}{N} \sum_{i=1}^N \frac{\#likes_i}{\#replies_i}$, with higher values indicating that the agent's tweets tend to have a perception of being well-received. As likes and replies are accumulated over the course of a simulation, sizing the nodes based on engagement metrics is only possible post-simulation. The edge/link widths for the image on the left has a fixed value, while those on the right are based on stance and themes compatibility between two agents. Stance is indicated by each node's color. Agency is indicated in the edge/link color and the node's border color. **Blue** nodes are the opposition, **yellow** nodes are neutral, and **green** nodes are supporters. A **red** border indicates a bot while a **white/invisible** border indicates a human.

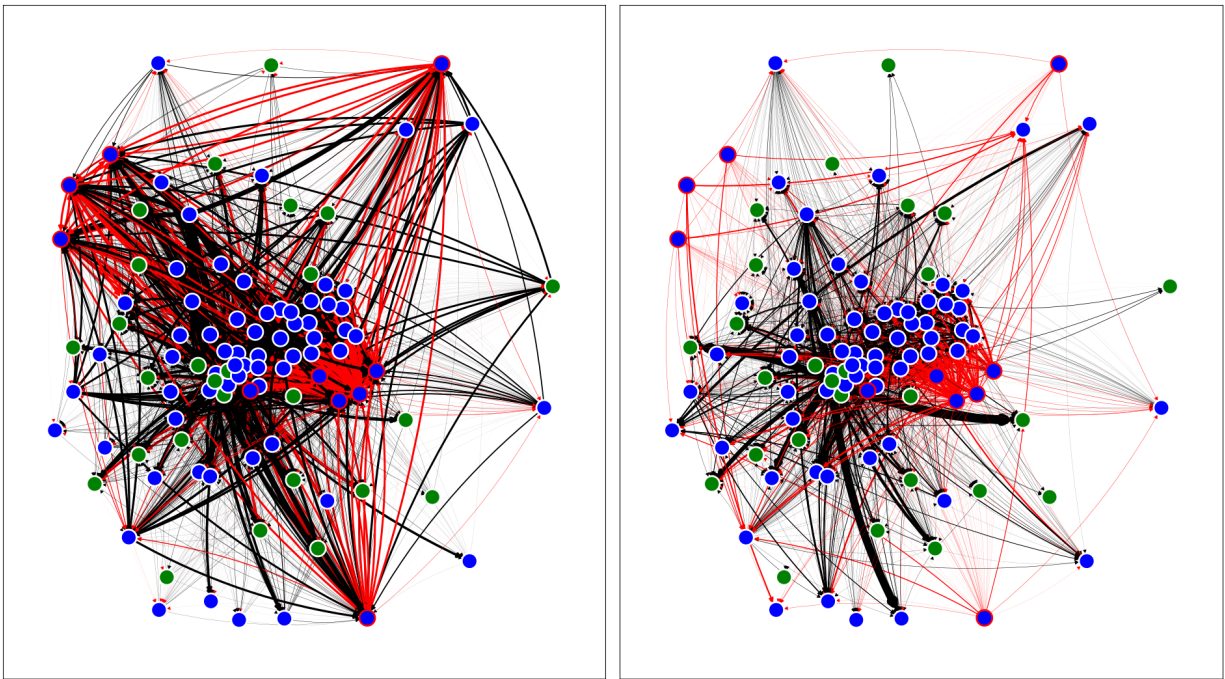


Figure 5.6: Embedded bots scenario. From the same simulation run as Figure 5.5. The image on the left shows whose tweets a node has seen at the end of the simulation. The image on the right shows whose tweets contributed to a node's stance switching at the end of the simulation. For both images, the edge widths correspond to the frequency of the seen agent's (left)/the contributor's (right) appearance. Not all seen tweets led to stance conversion, hence the difference. The color of each node indicates its stance while the color of the node's edges indicates agency. **Blue** nodes are the opposition, **yellow** nodes are neutral, and **green** nodes are supporters. A **red** border indicates a bot while a **white/invisible** border indicates a human. The edge/link colors distinguish between influence originating from bots, indicated in **red**, and from humans, in **black**.

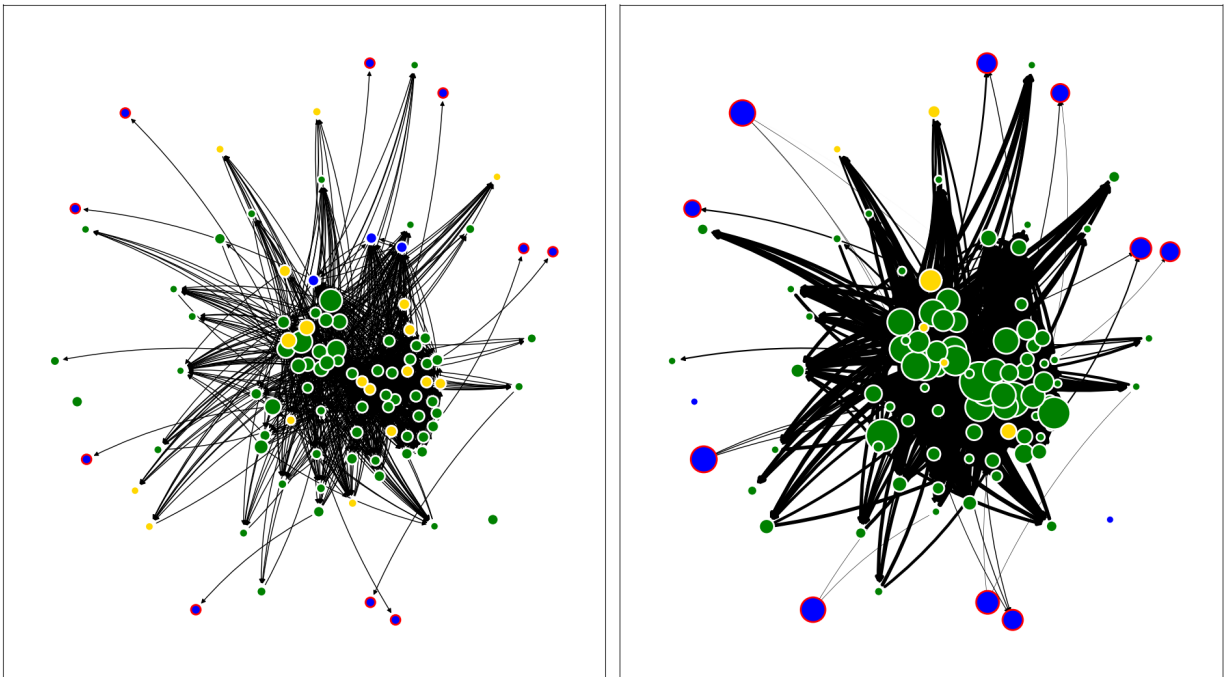


Figure 5.7: Peripheral bots scenario. Information flow for a simulation with 100 agents with all bots backing the negative stance (no equivalent in Table 5.1, closest is Sim 6). This scenario differs from Sim 13 in Figures 5.5 and 5.6 only in that each bot starts out in the periphery/fringes with a single outgoing edge (follower-followee) to a random human node instead of being embedded naturally in the network. This more challenging starting condition resulted in the non-bot-backed dominant opinion maintaining its dominance, although bots managed to convert two nodes in the fringes. Legend is the same as Figure 5.5.

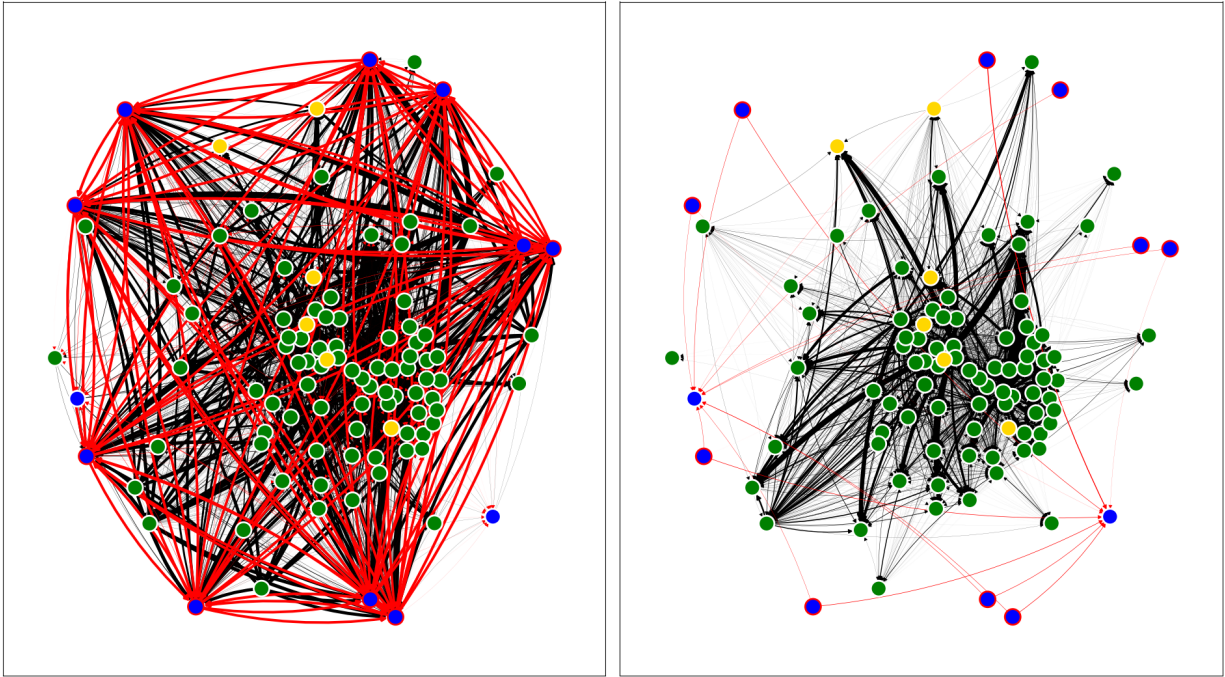


Figure 5.8: Peripheral bots scenario. From the same simulation run as Figure 5.7. Left image shows how often a different user’s tweets were seen by an agent while right image shows how often a different user’s tweets converted an agent’s stance. Legend is the same as Figure 5.6.

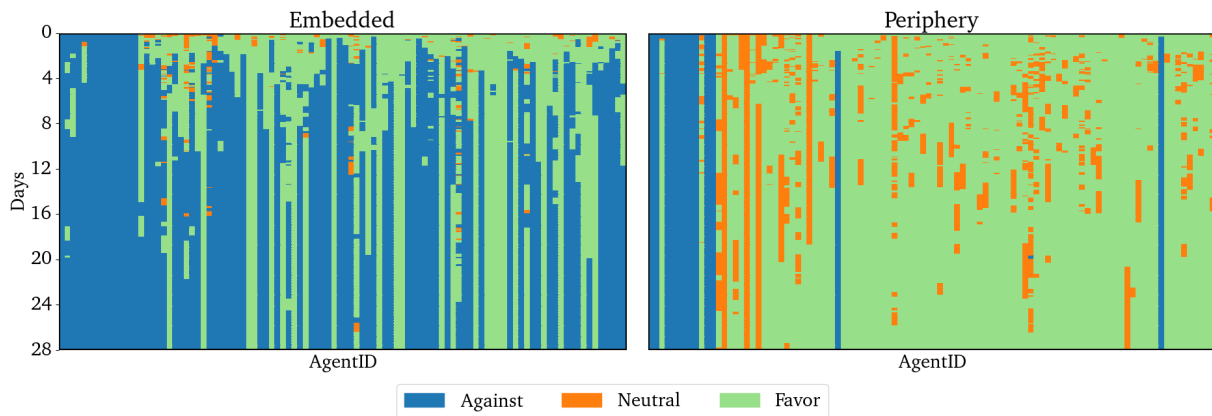


Figure 5.9: Each box in the heatmap represents the stance of an agent at a particular time-step. The evolution of an agent’s stance can be tracked by following a column from top to bottom. Agents are sorted by their *initial* stances, leftmost are negative agents and rightmost are positive, with neutrals in between. Rapid conversion at the early stage in the embedded scenario has rapidly eliminated neutral agents, hence their seeming absence. Data are from the same simulation runs as Figures 5.5–5.8.

slightly higher number of agents belonging to the minority stances in Sim 10 is likely a consequence of the bots' presence, whose stances can never be changed, slowing but not halting the network from reaching a consensus around the majority opinion. These are only trends, and not necessarily the final outcomes, as the simulations need to be run longer to observe whether all agents converge to a single opinion. And in the case where bots populate the simulation, Sim 10, the network will never fully converge as bots will never change their stance.

A winner-takes-all result [388] has been found for networks with competing viruses where one virus is stronger than the other, and the strength of the virus is quantified by a single numerical parameter (derived from attack and recovery rates). While the viruses (opinions) in our model are equal in that they do not have an explicit parameter that offers one virus an advantage over others, a combination of factors enable the dominant stance in our work to replicate a similar advantage over competitors. The dominant stance in *Diluvsn* is the stronger virus because homophily dictates that the stance has a more receptive audience, and their greater number means that their opinions are broadcasted and boosted more often, making others more susceptible to succumbing to peer pressure and adopting the majority stance.

Polarization can be induced if bots are present, and the stance distribution of bots differs from that of the humans. When half of the entire bot population holds a negative stance and the other half a neutral stance while the bots are restricted to human-like role and activity level distributions (Sim 8), the encroachment of the majority opinion is mostly halted. In Sims 12 and 16, where bots are only distributed between negative and positive stances with none for neutral, are assigned a role more suitable for stance propagation (Replier), and are all as active as highly active humans, neutrals virtually disappear and agents congregate around the two polar opinions. In the case of Sim 13 where all agents start out belonging to only one minority stance, specifically the polar minority stance, a similar depopulation of neutral agents is observed. More time-steps are needed to determine whether Sim 12 ends with this polarized distribution or if all agents are eventually converted to the bot-backed negative stances.

In existing studies on binary opinion dynamics, consensus cannot be reached if stubborn agents are present in the network [2, 516]. When the problem is extended to account for committed/stubborn agents and for multiple opinions (greater than binary), only opinions that are backed by committed people survive under the voter model [222]. The results from our simulations serve to further confirm the role played by stubborn agents, bots in our case, in driving polarization in a network.

5.6.2 Impact of bots and bot ratios

Given that the distribution of roles and action frequencies among bots do not differ between the stances (i.e., bots all have the same role and frequency regardless of the stance that they back), the ratio of bots between the stances has a drastic impact on the temporal evolution of the stances, which can be seen in their impact on polarization in (Section 5.6.1). If the stance ratio of bots follows that of humans, then the bots backing the majority stance would suppress the bots backing the minority stances, reproducing a dynamic very similar to the scenario where only humans presents while bots are absent. The bots' stance ratio must deviate from that of the humans to cause disequilibrium.

The co-opetitive relationship between the bots backing a minority polar stance (negative) and those backing a minority midpoint stance (neutral) benefits the midpoint stance the most. Across all scenarios in which the two minority stances are backed by an equal number of bots (Sims 1–9, 14, and 15), the number of neutral agents is always greater than that of negative agents. We suspect that this is due to the neutral agents' stance distance being the shortest for both polar stances (except for the distance between a polar stance and itself). Once neutral agents gain a critical mass through conversion, they have the numbers to peer pressure other agents in the network to adopt their opinion. But the strength of neutral agents is also their greatest vulnerability. Without agents immune to stance conversion, i.e., bots, in their midst, the group of neutral agents collapses quickly as they are the easiest targets to be converted by the majority stance (which does not need to be bot-backed) and the bot-backed minority stance. This phenomenon is evident in Sims 12, 13, and 16.

Depending on the info ops goal of the bots' operators, one mixture of bots can be more optimal than the other. If the goal is to maximize the number of agents holding a stance that is the polar opposite of the majority, all bot resources should be devoted to backing that stance, as shown by the results of Sim 13. The cost of not having neutral bots is a slight inefficiency in reducing the number of people holding the majority stance as seen in the greater number of agents being positive at simulation end in Sim 13 compared to Sim 1. However, if one intends to purely reduce the share of the majority stance, Sims 1 and 2 show that having an equal number of negative and neutral bots is the most effective. The neutral agents' smaller stance distance with positive agents makes them more persuasive. A side effect of a bot operator employing the tactic of mixed neutral and negative bots is that the population of negative agents is "cannibalized", a consequence of the bots in our simulations acting independently to avoid being seen as a centrally-directed, coordinated info ops. In cases where the positive stance is also backed by bots, using a mix of neutral and negative bots to counter the positive bots (a pro : neutral : con ratio of 2 : 2 : 1 in Sim

15) still appears marginally better than relying purely on negative bots (4 : 0 : 1 in Sim 16).

If the goal is to drive polarization, bots can be assigned to both ends of the pole at equal ratios (Sim 12) or with a ratio that favors the minority polar stance (Sim 16). Without unwavering advocacy for neutrality, neutral agents quickly dwindle to nothing, as polar opinions compete to attract users to their stance. In situations where there are motivated neutral agents, they must be outnumbered to be overpowered. As Sim 10 shows, if the bots' starting stance distribution matches that of the humans from the March mask-wearing dataset, which implies that the numbers of bots representing the pro and neutral camps are roughly equivalent in most runs, the polar minority stance, negative, diminishes in size far more rapidly than the midpoint minority stance, neutral.

For stances starting out from a disadvantaged position (negative and neutral at 13.9% and 12.9% of the population), Sims 15 and 16 show that the combined number of bots backing the non-majority stances requires an approximate 4-to-1 numerical advantage over the bots backing the majority stance in order to reach influence parity, assuming bots for all stances are similarly capable (same roles and activity levels). Should bots only back the polar stances in equal numbers while neutral bots are absent, the majority stance maintains and extends its dominance, nearly eliminating neutrals but is unable to penetrate the enclave of the non-neutral minority stance (Sim 12). The source of the negative agents' resilience against conversion can potentially be explained by Figure 5.4, where the follower and followee counts for the bot-backed negative stance in Sim 12 remained steady as the simulation progresses (so did the counts for the bot-backed positive stance). The neutral stance which lacked bot backing showed greater volatility and trended downwards. Bots backing only one minority stance against the onslaught of the majority stance, which can be seen in Sim 13, also stabilizes the follower and followee counts of the minority stance. Bot-backing may be capable of creating echo chambers for minority stances.

Although echo chambers/filter bubbles are terrible for the purpose of gaining new recruits as the Islamic State's experience has shown (Section 5.6.3), echo chambers are excellent for the reinforcement and consolidation of opinions [33]. A bot's broadcasted messages can reach people sharing the same stance, allowing bots to drive stance retention, i.e., keeping users with the same stance from being lost to other stances because those users are exposed to messages containing alternative stances. Bots play much of the same role as stubborn users in other information diffusion models; their unchanging stance making them ideal as the resilient core of a movement. Some have attributed Trump's electoral success to the presence of botnets "spreading pro-Trump content in overwhelming numbers" and "digital echo chambers where users see content and posts that agree only with their preexisting beliefs" [139]. Using bots to create echo chambers may even escape alerting

social media monitors as certain platforms are already inclined towards echo chamber formation, with researchers finding that “information diffusion is biased towards individuals who share a similar leaning in some social media, namely Twitter and Facebook” compared to Reddit and Gab [107].

When great political orators across the entire American political spectrum, e.g., Hillary Clinton [272] and George W. Bush, expressed sentiments along the lines of “you are either with us, or against us”, that there is no middle ground, their words contain a kernel of truth. From the perspective of eroding the support base of a polar opinion (e.g., pro-war), the actions taken by neutrals and the opposition (e.g., situationally pro-war/war-ambivalent and anti-war) to draw people into supporting their respective stances have the same effect — erosion. All our simulations featuring a pro : neutral : con stance ratio at 1 : 1 : 0 for bots provide evidence that support for the positive stance is inevitably eroded (Sims 6 and 7 are not exceptions as they are just examples of bots being ineffectual, as discussed in Section 5.6.5). In other words, if a predominant force supports one pole of a polarized issue, appealing to others not to pick either side can make one a collaborator for the “enemy” (supporters of the opposite pole). The slightest deviation from full, enthusiastic support of a stance is indeed “treachery”.

5.6.3 Stance conversion and retention strategies

Repetition is the key to success for info ops. Repetition in this case means being extremely active in disseminating and boosting content. If the action frequency of the bots is cut in half, from 134 actions per week to 66, as seen in Sim 1 vs. Sim 3, then bots are substantially less effective at promoting minority stances against a non-bot-backed majority stance. This reduction in influence is also evident in other scenarios where bots went from being highly active to adopting activity levels that mimic the human distribution (low activity and medium activity users make up 67.4% and 22.2% of the human population respectively) which can be seen in Sim 1 vs. Sims 8 and 10. Our observation that repetition is important has support, e.g., simulations in [241] which showed that high energy agents contributed towards the prevalence of extreme opinions (polarization). Another study demonstrated that repetition can turn into influence: a bot in a social media experiment that had gained popularity simply by periodically visiting others’ profile pages later recommended accounts for users to follow and succeeded in convincing a majority of them to take its recommendation [11].

Not all forms of repetitive action elicit the same responses from the network. Sims 4 and 5 show that bots who favored liking or favored retweeting over all other actions are

unable to convert as many users to their stance compared to bots which favored replying (Sim 1) despite having the same high activity level. In the case of Sim 5, the bots were only able to stop the majority opinion from gaining ground. There are several factors behind these outcomes. In a situation where the bots represent the minority stance, there are a greater number of tweets that express the majority stance and consequently more targets for negatory replies. On the other hand, bots that favor liking and retweeting, which are affirmatory actions, are likely to have run out of minority stance tweets to boost. As a small percentage of the population, bots are unlikely to occupy a majority of high-degree nodes, limiting the reach of their messages. Replier bots can sidestep the lack of an audience by borrowing the audience of others as a small portion of replies is seen by the followers of the parent tweet. Retweets, however, are only seen by the author of the parent tweet (and the bot's own followers). Liker bots do not have to find an audience as they are liking other users' tweets that already have audiences, namely those other users' followers, but the reach of liker bots is largely confined to these audiences as they seldom invade the conversations of other users compared to Replier bots. Since minority stances are minorities population-wise and Liker bots who back a minority stance will favor liking tweets advocating the minority stance, the audience of tweets liked by Liker bots are small too, although not as small as that of Retweeter bots.

Targeting vulnerable users — friendless people, i.e., having a low number of followees — for replies can result in a similar number of stance conversions as the alternative tactic of finding the largest platform to shout from (replying to tweets with high engagement), as demonstrated by Sim 1 vs. Sim 2. Also, in terms of decreasing the number of people holding the majority stance, both Sim 1 and Sim 2 are the most effective of all the tested tactics (Sim 9 will be discussed in Section 5.6.4). Our results are in accord with real-world observations. Bots finding success by targeting vulnerable users bears some similarities to how “extremist” groups have historically targeted loners, marginalized and ostracized groups, and other societal outcasts for recruitment/radicalization owing to these people's greater receptiveness towards extremist messaging [199, 465]. The tactic of maximizing audience reach is also favored by extremist groups, as seen in the Islamic State (IS) “fighting for its propaganda to appear on search indexes and continued to launch ‘media invasion raids’ to gain an even short-lived visibility on mainstream social media” due to IS's favored social media platform, Telegram, having “limited audience reach and echo-chamber effect, with the IS message being shared mostly among like-minded individuals” [495]. Another example are BTC scam bots [321], which are known to raid the replies sections of other tweets to maximize visibility.

5.6.4 Themes conservation and propagation

Themes are unique within our model in that agents do not guard against them unlike stances that have polarity, but their transmission is entirely dependent on the successful adoption of polarizable stances. Themes can improve the transmission probability of the stance that they are attached to if the targeted agent shares the same themes. Themes can therefore be thought of as having a symbiotic relationship with stances.

We tested a naive tactic for improving the propagation of themes — saturation. By making all the bots conserve a particular theme (Theme 7), we aim to see if themes can spread beyond the bots themselves and into the human population. Conservation involves attaching the conserved theme(s) to every outgoing message from bots and also giving the conserved theme(s) a permanent place in a bot's set of themes. The results from Sim 9 show that not only are bots capable of making half the population of the simulation be infected by Theme 7, they were able to accomplish it without causing any appreciable change in the final distribution of the stances when compared to Sim 1, i.e., the bots still experienced similar success in stance propagation while conserving a theme. Because the number of agents holding the conserved stance, Theme 7, exceeds the bot population and an agent cannot hold duplicates in its theme set, we can conclude that the bots have succeeded in pushing the conserved theme onto human agents.

Even without an explicit theme conservation tactic, the bots' greedy strategy of keeping half of their own themes whenever they survey tweets for themes to adopt ensures that minority themes are given a boost whenever bots are involved in a simulation (Sims 1–16, except the humans-only Sim 11). In cases where bots failed to propagate their backed stances because the bots' starting positions are in the periphery of the network (Sims 6 and 7), the bots are still capable of elevating the presence of the initially obscure Theme 7. Bots can also suppress the growth of popular themes or even cause their decline, which can be observed by tracking the counts for the first and second highest frequency themes, Themes 3 and 2 across the simulations. The exceptions are when bots are exiled to the fringes, as in Sims 6 and 7, when bot activity levels are low, as in Sims 8 and 10, or when bots are completely absent, as in Sim 11, where a normative rich-gets-richer effect is at play where themes that were already popular mostly grew in popularity as the simulation progresses.

Prior to obtaining our simulation results, our concern was that themes not matching fully will hamper stance adoption and therefore theme adoption; theme compatibility is $c_t = |t_{\text{user}} \cap t_{\text{tweet}}| / |t_{\text{user}} \cup t_{\text{tweet}}|$, so deliberately including a theme that is virtually absent in the entire population would ensure that during the early phase of the simulation, the denominator of the theme compatibility equation will almost always be increased by one

even in cases where all the non-conserved themes in a tweet matches with a user's theme. However, as the results from Sim 9 demonstrate, this is not a factor, at least not when the number of themes to be conserved is just one. The outcome might be different if more than one theme has to be conserved, as the reduced presence of compatible themes may be insufficient to compensate for the incompatibility arising from multiple conserved themes.

We theorize that engagement metrics, a high volume of replies representing a stance, and other effects that are within the ability of the bots in the simulation to manipulate are sufficient to overcome incompatibility in themes during stance adoption. Once a critical mass of agents are infected with the conserved theme, a synergistic effect begins to take hold. Human agents who have yet to be infected by the conserved theme find that their tweets are less effective at changing the stances of those who are infected with the conserved theme. This enables the number of agents with the conserved theme to grow without much interruption. These agents are also likely to share the same stance(s) as the bots because the conserved theme is almost exclusively possessed by the bots at simulation start, so acquiring that theme implies that the bots' stance(s) has infected the agent, at least in the early phase of the simulation. Once agents with the conserved theme become the majority, they can peer pressure others into accepting their stance. A theme can be viewed as performing a function similar to "investments" (sunk costs) that makes "betraying"/leaving "extremist" groups [22, 352], armed forces [505], religions [123], and currency standards [108, 320] difficult.

Bots being able to push a single theme without sacrificing progress on stance conversion, as results from Sim 9 demonstrated, is a non-trivial finding. There are implications for actors intending to conduct separate info ops on more than one issue. Such actors would ideally wish to succeed on all fronts without having to decide which info ops are the most important ones while making trade-offs and compromises for the rest. Our results show that not only can balancing acts be avoided, the actors can design an info ops in ways that can improve the outcomes of other related info ops. We illustrate this with an example concerning the issue of mask-wearing.

Under the guise of facilitating debate between the supporting, neutral, and opposing factions of mask-wearing, a motivated actor participating in more than one side of the debate via bots that always frame their messages with a theme of liberty (personal/individual freedom versus communal responsibilities and interests) is ensuring that the theme over time becomes the basis around which the issue is discussed as an increasing number of agents are infected by the theme. This mindset caused by the theme-infection can potentially be carried over into other issues, e.g., seatbelts, drunk driving, and climate change, causing the infected agents to frame those issues in terms of liberty. This is similar to the belief

in the field of pedagogy that thinking habits can be taught and can be carried over across different situations (e.g., critical thinking [426], thinking like an engineer [303], fostering creativity in children [17]), exposing people who think in a certain way to cognitive blind spots [367] due to “belief bias and the magnitude of framing effects” [497]. When others attempt to persuade these theme-infected agents to, for instance, not drink and drive by presenting arguments framed only in authoritarian terms (e.g., driving under the influence is against the law) or safety terms (e.g., drinking and driving often kills the driver), they would have difficulty dislodging the agent’s stance.

5.6.5 Embedded vs. peripheral bots

Figures 5.5, 5.6, 5.7, and 5.8 are visual illustrations of the differences between bots starting out embedded naturally within a network and bots starting out as outsiders with only one outgoing follower-followee link (i.e., the bots’ tweets that are not replies will not be seen by anyone in the network). This challenging start is meant to mimic the “time zero on the periphery” with “a single link to a random node” start for bot disinformation maneuvers in [57]. Two other scenarios featuring bots in the periphery are Sims 6 and 7 in Table 5.1. Although only 100 agents are used for plotting these graphs, the embedded scenario is representative of the 1,000 agent simulation results in Table 5.1 for Sim 13. The isolated/fringe scenario with 100 agents has no equivalent in Table 5.1, but it can be considered a variant of Sim 6 where the stance ratio for bots is 1 : 0 : 0 instead of 1 : 1 : 0. The influence of bots is easier to see visually if the bots all belonged to one stance, hence the chosen configurations.

Embedded bots (Sim 13 and Figures 5.5 and 5.6) succeeded in converting a majority of the network population to its stance but the bots that started out in the network’s fringes (Sims 6 and 7, and Figures 5.7 and 5.8) could only convert other users who also exist on the fringes. Cut off from the influence of negative bots existing on the periphery, the positive human agents in the central clusters rapidly converted the few neutral and negative human agents into positive agents (Figures 5.7 and 5.8, and Sims 6 and 7). Although bots did not actively initiate the severing of links — they start the simulation already isolated from the central clusters — the peripheral bots scenarios serve as an *in situ* demonstration of the effects of a community (central clusters) shielding itself from outside influence (bots). The results from these simulations of ours are in accord with the model in [421] which produced predictions of social influence (homophily in opinion) and unfollowing being drivers behind the emergence of echo chambers, consistent with the empirical Twitter data they collected.

The fierceness of ideological competition is evident from comparing both halves of Figure 5.6, which represents the embedded bots scenario. While a large share of outgoing messages were read by users in the network, as indicated by the thick lines of the edges in the graph on the left, the messages that convinced another user to switch their stances are a smaller proportion of those that have been read, indicated by the thinner influence lines connecting nodes on the right. The same observation applies even more so to the peripheral bots scenario in Figure 5.8. Bot messages lack the consistent persuasiveness of more centrally located humans, as the bots' red influence lines are comparatively thin next to the thick black influence lines emanating from humans. One shared trait of the winning sides in both the embedded and fringe cases is that their tweets receive a lot of affirmation, i.e., likes received outnumbered the replies, indicated by their greater node size in the right halves of Figure 5.5 (blue negative agents) and Figure 5.7 (green positive agents).

Figure 5.9 shows that stance conversion happens early in both the embedded and fringe scenarios. While there are occasional bouts of rapid stance switches, most agents tend to hold a consistent stance for long periods of time. In the fringe scenario, there is a greater occurrence of positive agents converting to neutrality for short stretches of time in the early phase of the simulation. We speculate that this is because of negative bots having numbers that are sufficient to dampen positive tweets with their actions but are insufficient to boost negative tweets past the point where their stance incompatibility with positive agents is minimal, leaving neutrality to benefit from the struggle¹⁰. The outcomes for Sim 6 and Sim 7 in Table 5.1 provide support for this observation in Figure 5.9. Despite the absolute counts at simulation end dropping for both negative and neutral, neutral has a higher count relative to negative despite starting with a comparatively lower count.

¹⁰An example with a positive agent undergoing stance conversion after having read a tweet each from negative, neutral, and positive stances: Recall that a stance's conversion score w_r is made up of two components, engagement r_g and compatibility r_c , $w_r = m^{-1}r_g + mr_c$. The m^{-1} and m terms are weights used to de-emphasize engagement and emphasize compatibility, respectively. Let themes be ignored in the compatibility calculation and assume that the two components will *not* be weighted when summed. Assume that in the absence of bot-backing, the three tweets receive zero engagement. Assume that in the presence of bot-backing for the negative stance, the backing involves only actions that impact the engagement component of the conversion weight — the positive tweet receiving one reply (assume that the reply was not seen by the agent experiencing this conversion event). For the non-bot backed case, $r_{g-con} = r_{g-neu} = r_{g-pro} = 0$ and $r_{c-con} = 0, r_{c-neu} = 0.33, r_{c-pro} = 0.66$. For the bot-backed case, r_c remains the same but $r_{g-pos} = -0.5$ due to receiving a reply. Engagement components for negative and neutral remain the same, $r_{g-con} = r_{g-neu} = 0$, because they received no engagement. The conversion weights without bot-backing are $w_{r-neg} = 0, w_{r-neu} = 0.5, w_{r-pro} = 1$ and with bot-backing, they are $w_{r-neg} = 0, w_{r-neu} = 0.5, w_{r-pro} = 0.16$. In percentages, the chances of a stance being selected, from negative to neutral to positive, are 0%, 33%, and 66% for the case without bot-backing, and are 0%, 76%, and 24% with bot-backing. By "attacking" positive tweets, the bots' actions ended up making neutrality a more attractive option. In the actual simulations, the m^{-1} and m weights are meant to ensure that the influence from engagement metrics does not easily overpower homophily for stance and themes.

The peripheral bots' tactic of targeting low-follower accounts for replying and quote-tweeting is unlikely to be the cause of the bots' failure in combating the majority stance, because replying and quote-tweeting retweeting is limited to once each per tweet, so bots would switch to other targets once they have hit their interaction quota with agents on the fringes. For liking and retweeting, bots target tweets with high engagement written by agents with high follower counts. Incidentally, bots frequently liking other tweets to boost their infectivity appears to be a slightly inferior tactic compared to frequently replying to messages (Sim 6 vs. Sim 7). This difference between the Liker and Replier bots are less pronounced compared to the embedded scenarios (Sim 1 vs. Sim 4).

The difficulty faced by bots under the fringe scenario in overturning the predominant stance can perhaps be attributed to the follow and unfollow mechanism in *Diluvion*. The probability for an agent to follow and unfollow accounts are based on data of user behavior on social media (Section 5.3.8.1) that likely was collected under "peacetime", i.e., when a community is not vigorously debating a topic. Therefore, they may not reflect a greater tendency to cluster into echo chambers/hives during "wartime" by rapidly establishing social ties with like-minded users and severing ties with non-believers. However, as we aim to ground our model in as much empirical data as possible and we did not encounter data of heightened follow/unfollow behavior on "war footing", we opted to keep probabilities low. The model in [57], *twitter_sim*, strongly incentivizes the creation of new links, which explains the difference between their simulation results and ours, where they demonstrated that bots are successful in backing a stance and in bridging two separate communities despite bots starting in the fringes. Under *twitter_sim*, link creation probability for humans is low at 5% but is higher than our model's 0.6%, the check for link creation is triggered every time a user wakes instead of being a separate event that consumes a user's action budget under our model (compounding the higher link creation probability), and links in *twitter_sim* are never broken.

Figure 5.4 shows the crippling effect of being deprived of an audience. Unlike other scenarios where the minority stances' mean follower and followee counts are at parity with the majority stance's at simulation start, they are only about three-fourths of the majority's in Sims 6 and 7. Once agents begin renouncing a minority stance for the majority stance, the minority stance's follower and followee counts begin experiencing a precipitous decline that they never recover from.

Our results here are corroborated by the details of real info ops on social media that have emerged (Section 5.2.2.6), showing that great effort is invested in equipping many bots with extensive social ties, going as far as having real humans work under a bot. No matter how persuasive a bot is constructed to be, its effect is limited if it does not have a

ready audience that is always exposed to the messages that the bot broadcasts. Bombarding the replies section of other users to persuade them and their neighbors into changing their stances cannot succeed if the bots are just 10% of the population, as our simulations have shown. Resorting to other tactics to improve the performance of bots without increasing the number of bots, e.g., explicitly coordinating with other bots and increasing the frequency of activity, runs the risk of the info ops being discovered.

5.7 Limitations and future work

At the outset of this paper, we have stated that no model is perfect and that they are all reflections of some aspect of real life that the modelers wish to bring to attention. There is room for future work because we made several simplifying assumptions in our model. For instance, we accounted for the bandwagon effect but did not implement a mechanism for the spiral of silence. Our simulations start out with one dominant and two minority stances, which do not cover scenarios where the distributions are balanced. Some other assumptions involved picking one side of an ongoing debate (e.g., whether resistance increases with repeat infections) and not having the opportunity to explore the alternative side in-depth in this paper.

One promising route for future investigation is to allow a reinforcement agent to control the bots. The reinforcement learning agent AlphaStar has been demonstrated to perform at the Grandmaster level for the real-time strategy (RTS) game StarCraft II [480]; an info ops problem bears strong resemblance to an RTS game. Just as playing an RTS game involves controlling groups of hundreds of units with specific roles (attackers, defenders, healers, etc.), so too can the managing of an info ops on social media involve controlling hundreds of bots with specific roles (Repliers, Likers, etc.). However, a central controller dictating the actions of multiple bots does violate the emergent/uncoordinated coordination condition that this paper is predicated upon and could potentially make this proposed form of bot-driven info ops easier to detect, so an alternative scenario where there are multiple reinforcement agents that each control one bot should also be explored.

Greater heterogeneity in terms of bots should also be investigated. While we have varied the distribution of roles and activity levels, we have kept other aspects homogeneous throughout a simulation run, such as the process by which a bot chooses a tweet to like. And when deciding how bots should act, e.g., the per-week frequency of actions such as retweeting, we have relied on pre-defined roles derived from empirical observations of Twitter user behavior. All this was done to facilitate comparison between scenarios. There may be more optimal action distributions for maximizing stance propagation. Heterogeneity

would also enable us to explore more complicated info ops scenarios. An example would be explicit cooperation between neutral and opposition bots, where neutral bots would only target agents who are supporters for conversion to neutrality while opposition bots would target both neutral and supporter agents. In this example, opposition bots would effectively function as a conversion pipeline for opposition bots. Compared to opposition bots attempting to convert supporters into detractors, the lesser difference in stance between supporters and neutrals means that neutral bots would have an easier time converting supporters to neutrality. Once converted to neutrality, opposition bots can more easily convert the former supporters into agents of the opposition.

Agent heterogeneity should include class and role transitions, e.g., deepening radicalization turning human agents into bots and de-radicalization turning bots to humans. Bots switching roles as the info ops situation demands should also be included. Role transition in an ABM for information diffusion has been explored before in [445] where radicalized patients become staunch anti-vaccine propagandists. Our model's radicalization and de-radicalization involved only an agent switching stances. In addition, human and bot users entering and exiting the network should be accounted for. The dynamics of a social network in which the ratio between bots and humans is ever shifting may provide additional info ops insights.

5.8 Conclusion

In this paper, we developed and empirically validated an agent-based information diffusion model for bot-driven information operations on social media, *Diluvision*, that incorporated crucial but oft-neglected real-world aspects of diffusion. Chief among them is indirect influence, consisting of two forms, the exposure to a message's engagement metrics inducing bandwagon/conformity effects within a user and the ability to transmit information to another node in the absence of a social tie linking two nodes together. Other key innovations of the model are the designation of neutrality as an independent infection-capable stance instead of a transition point between the two discrete polar stances (opinions) and the introduction of the non-polarizable, homophily-inducing, infection-capable, and stance-symbiotic themes as a counterpart to stances.

Through simulations, we demonstrated the different routes by which motivated agents, i.e., bots, can exploit these aspects of information diffusion to manipulate the information environment. The simulations showed that highly active Replier bots representing minority stances are capable of overturning the dominance of a non-bot-backed majority stance, and that a naive saturation tactic can not only keep a minority theme, e.g., "conspiracy",

alive but also further its adoption over time without compromising stance propagation. Neutrality is also shown to be non-neutral in its threat to the continued dominance of the majority stance. Overall, our model shows that info ops originating from agents embedded in the network who are not counterbalanced by a similarly motivated oppositional force can easily pollute the information environment and induce different outcomes, such as increased polarization.

Although much of our paper has been steeped in the language of war, the model we introduced here can be used towards beneficial, pacific ends. Reducing polarization on social media through the introduction of bots that stubbornly advocate a neutral stance, for instance, can be studied through *Diluvion*, in the tradition of the numerous works that have been published on depolarization, e.g., [117, 345, 444, 479]. The assumption here is that polarization reduction is prosocial, an assumption that some may not agree with as they may believe that there is no peaceful coexistence with certain opinions or ideas, e.g., Nazism. On the other hand, there are real-life examples — the US, for one — of comfortable coexistence being possible between “extreme” ideologies and the ideals of liberal democracy, the best governing system ever and which Churchill, the widely venerated British prime minister who presided over the Bengal famine [267, 343, 413], wryly described as “the worst form of Government except for all those other forms that have been tried from time to time”. The US never lost its global status (global as in encompassing only the “garden” [313]) as the shining beacon of democracy despite being the direct progenitor of Nazism via cherished American ideals such as Manifest Destiny [340, 499] and despite fascism resurging in modern-day America [59, 124, 402, 405].

Or, perhaps, certain states, systems, ideas, etc., maintaining an untarnished reputation as a force for good are not due to their inherent merits but due to having waged successful info ops campaigns, and maintaining that reputation in the age of social media requires active effort in maintaining an edge in tools for wargaming info ops on social media, in tools such as *Diluvion*.

Chapter 6

Final Conclusion

6.1 Summary

This thesis shows how our research objective (Section 1.4) of studying saturation attacks in the information sphere was achieved by four separate yet interlinked research papers.

The first paper (Chapter 2) accomplishes the research sub-objective of improving defenses against saturation attacks targeting information systems by presenting a framework for entity matching, i.e., identifying duplicate entities, across databases, which can help us guard against unconventional attacks that employ duplicates to flood databases. Our proposed framework, *AttendEM*, also addresses the challenge of ensembling memory-intensive large language models by not using multiple architectures in the ensembles and instead relying on creating diversity during the text preprocessing and text tokenization stages to generate multiple unique models from a single transformer architecture. The *AttendEM* framework obtained better performance over the previous SOTA framework, Ditto, ranging from 2.99% to 3.43% depending on the set of Ditto replication results used. Although the improvement may seem slight, they are comparable to the improvements reported by the competing solutions, HierGAT and DAEM, over Ditto which ranged from 2.46% to 3.42%. In summary, the outcome of this study is an improved entity matcher that can be employed to mitigate saturation attacks targeting e-commerce databases via duplicate product listings.

The second paper (Chapter 3) accomplishes the research sub-objective of mitigating unintended damage to healthy public discourse due to overly aggressive defensive measures against saturation attacks by presenting a framework for separating satirical news from fake news, which prevents satirical news from ending up as collateral damage due to efforts in stemming the flood of fake news. This framework for the classification task of distinguishing between fake and satirical news features an integrated DistilBERT transformer model. The

transformer's tokenization process was modified to use multiple specialized aggregator tokens instead of a single token. Another method by which we elicited better performance from the transformer was to pre-train and fine-tune the model instead of just fine-tuning. Against a dataset of satirical and fake news, our model showed an improvement of 5.2% in F1 and 6.4% in accuracy over the then-SOTA model. Our model is also robust against the challenge of distinguishing different subcategories of dis/misinformation (propaganda, clickbait, satire, hoax), opening the avenue for an even more tailored approach to combating undesirable information. In summary, the outcome of this study is an improved fake versus satirical news discriminator that can be used to prevent satirical news from being unintentionally removed alongside fake news (due to fake news' high similarity to satirical news) when stakeholders attempt to reduce the saturation of fake news in information channels.

The third paper (Chapter 4) accomplishes the research sub-objective of gaining an understanding on information propagation dynamics of contentious issues by studying floods of messages bearing competing stances on a nascent non-pharmaceutical intervention, universal mask-wearing, vie for dominance (saturation) on a social media platform, Twitter. We found that the pro-mask stance, though dominating the conversation early, have their dominance slowly but surely eroded by the anti-mask stance. From the perspective of themes, the debates appear to grow increasingly fierce as more and more tweets are critical of the decision to wear or not to wear masks. Other notable findings were effective and potentially generalizable tactics of furthering the spread of one's stance: hijacking popular conversational threads and deliberate politicization of an issue. Conversely, to keep the spread of a competing stance at bay, one should prevent the emergence of a few centralized discussion threads whose influence can be greater than large numbers of small, scattered threads (incidentally, the formation of centralized threads being disrupted prompted us to suspect a censorship attempt from the platform owners). There is also commonality in tactics used across different stances, e.g., accusing members belonging to a different stance of authoritarianism. In summary, the outcome of the study is the successful development of an understanding of how different sides of real-world discourse can attempt to gain, erode, or maintain saturation, i.e., dominance, on an information exchange platform.

The fourth paper (Chapter 5) accomplishes the research sub-objective of constructing an information diffusion model based on our study of opinion dynamics on social media by introducing *DiluvSION*, an agent-based information diffusion model. The sub-objective of simulating interventions using the model is accomplished through our investigation of how discourse evolution on social media can be distorted by an engineered flood of messages and likes (expression of approval) creating an illusion of popular support. *DiluvSION*

demonstrates that a small group of staunch supporters for a stance, i.e., bots, can indeed effectively exploit humans who adhere to empirically supported herd information processing behaviors, namely bandwagoning and equating popularity with credibility. A number of information operations objectives can be accomplished with the bots' naive saturation attack, including maximizing the propagation of a stance. We also demonstrated that, should neutrality be backed by bots, neutrality as a stance is non-neutral in the context of stance propagation as it is positioned advantageously to erode support from the two polar stances. The model also shows that supporting multiple stances to camouflage a true goal, the dissemination of a theme, is an effective tactic. In summary, the outcome of the study is an information diffusion model, *Diluvision*, that incorporates a number of critical features, e.g., neutrality as an infection-capable stance, lacking in existing models and using *Diluvision* to simulate interventions against saturation attacks of undesirable information being spread by independent agents on social media, allowing us to identify viable interventions.

Taken as a whole, the four papers studied heterogeneous systems vulnerable to being overwhelmed by an information deluge and the papers also demonstrated solutions against such a flood, thereby accomplishing the research objective of deepening our knowledge of unorthodox saturation attacks and improving defensive measures against the attacks.

6.2 Discussion

As the storage, distribution, and consumption of information are increasingly cyber-mediated (see Section 1.1), algorithms, i.e., automated librarians, are supplanting humans librarians as information intermediaries when people seek to have their information needs met. The control wielded by these algorithms over information visibility — deciding which information to surface or submerge when presenting results of information retrieval requests — makes them potent gatekeepers. The field of information studies must not neglect advancing the knowledge frontier on algorithmic gatekeeping, because if gatekeeping is “done on a rational basis, rather than one that becomes simply another form of censorship, the role of gatekeeper may become one of the information profession’s most important roles within an information society” [379]. Our thesis focuses on one aspect of gatekeeping, which is using automated librarians as gatekeepers to safeguard information seekers from floods of undesirable information spread by adversarial forces engaged in information warfare.

There are many pathways towards algorithmically addressing informational saturation attacks and our research chapters represent several different pathways, which is why we will frame our discussion of our research chapters in terms of their position along three axes.

This framing provides a cohesive and informative picture of the entire research program. The axes have been briefly introduced at the beginning of our thesis in Section 1.4; the progression of our research chapters matches the progression along three axes: from tactical to strategic, from reactive to proactive, and from discriminative to non-discriminative.

The transition from tactical to strategic fosters the conceptualization of complex, multi-faceted defense strategies for social cybersecurity instead of fostering the development of solutions that lack consideration of how they fit into an integrated whole. As noted in a 2021 review of influence operations countermeasures, a disconnect exists between academic research and practical implementation, with fact-checking being “overstudied relative to its centrality in platforms’ regulatory toolkits” [114]. The shift to strategic thinking also encourages a broad, multi-disciplinary outlook of social cyberwarfare. Such an outlook can more effectively present solutions to fast evolving social cyberwarfare threats. We are not alone in expressing such a sentiment. Others have echoed it, stating a need for “[n]ew empirical results, *transdisciplinary* theories, models and ways of measuring resilience and mitigation” (emphasis ours) in making communities resilient against social cybersecurity attacks [87]. Therefore, instead of a monotonic pursuit of ever higher classification accuracies as found in the earlier research chapters of this thesis, our later research chapters sought to understand how agents can pursue a variety of influence maximization tactics. We relied on transdisciplinary theories, best exemplified by Sections 5.2.1.3 and 5.3.3.2.

The transition from reactive to proactive, from passively dampening saturation attacks to actively using saturation attacks against saturation attacks, is a lateral move in terms of the dangers involved. Just as an overly aggressive defensive measure (e.g., removing satirical content due to their similarity to fake news despite satirical content raising political awareness and stimulating political engagement) can harm the defender (e.g., damaging civil discourse), so too can taking proactive instead of reactive measures unintentionally inflict self-harm. One of the earliest conceptualizations of information warfare, published in 1995, recognized the danger of proactive measures: “information attacks have stochastic effects and that unless these are considered and evaluated in advance, an information attack may not have the effect ultimately desired” and “it may have consequences that are so undesirable that the attacker will rue that an attack was made in the first place” [457]. An example would be the introduction of a theme, (e.g., authority), for the purpose of inoculating a population against a competing opinion, a concept we explored in Section 5.6.4 under the theme conservation scenario. Widespread adoption of the theme may, over time, make the entire populace view every problem through that frame/theme, even when the problem perhaps should not be viewed through that frame/theme (e.g., scientific

consensus on vaccine safety and climate change viewed as tyrannical and authoritarian).

The transition from discriminative to non-discriminative frees researchers from the ethical concern of making value judgments. When training classifiers with datasets as we did in Chapters 2 and 3, the values (e.g., what constitutes true or false, fake or satire) embedded into the datasets by their creators becomes a part of the classifiers. These values can be biased, sometimes severely so. This problem is magnified by classifiers not being perfectly accurate. While error tolerance can be high when classifiers are applied to less sensitive tasks, we cannot be as accommodating when using classifiers to identify if a piece of content is fake news or to identify if a user is a “bot”, as misclassification that results in removal of the content or the user is tantamount to restricting free speech. There have been documented instances of humans misclassified as bots by bot-detection algorithms [398], most famously in Twitter’s witch hunt for Russian bots in 2018 in which Americans wound up on the bot list despite the list having supposedly undergone manual human examination prior to being submitted to the US Congress [84]. The misclassification rates are unlikely to improve as AI increases in sophistication and the number of users joining social media services grows. By not embedding value judgments into proposed solutions for saturation attacks (Chapter 5), the weight of ethical concerns is mostly shifted onto decision makers who have power over the deployment of the solutions.

6.3 Limitations and future work

Limitations and future research directions have also been discussed extensively in the individual chapters (Sections 2.5, 3.4.6, 4.7, and 5.7), and a common thread among all the proposed future research directions is that they seek to improve the generalizability of our models and findings because the research outcomes are derived from a restricted context, e.g., using fake and satirical news dataset that is English-only and is focused only on politics. For the entity matching framework, *AttendEM*, we seek to evaluate it against datasets outside the ER-Magellan benchmark set that is composed of commercial product listings and scientific bibliographic data. For the satire and fake news classifier, we seek to evaluate it against datasets from different languages and regions to avoid overfitting to an Anglo-American centric idea of satire and fake news. Our study on mask-wearing discourse on Twitter seek to confirm that similar trends can be observed for mask-wearing discourse for other languages and for different emerging contentious issues. As for the information diffusion model, *Diluvion*, we seek to incorporate other empirically validated cognitive heuristics such as the spiral of silence to present a more accurate picture of contested diffusion.

There are also individualized recommendations. Of note is the need for the satires versus fakes discriminator to offer better interpretability of how it arrives at its decisions to help its users develop a better understanding of its inner workings and to offer transparency to entities affected by its judgments; at present, the discriminator’s utility is limited by being a black box as there are application scenarios where interpretability is a pre-requisite for a discriminator’s adoption. The entity matching *AttendEM* that has been reliant on a brute force ensemble could benefit from a thorough approach to text cleaning; de-emphasizing text cleaning was a deliberate choice to distinguish *AttendEM* from competing solutions, so removing this artificially imposed limitation could result in improved discriminative ability. The information diffusion model *Diluvision* could introduce adaptive bots that self-improve via reinforcement learning instead of ones that follow fixed heuristics; to be explicit, the limitation we seek to overcome in this case is the bots being shackled to fixed heuristics.

Some of the future research directions have been partially accomplished: adding pre-training to a discriminator framework as proposed in Section 2.5 has been successfully tested in Chapter 3, and swapping transformer architectures as suggested in Section 3.4.6 has been implemented in Chapter 2, although the models differ significantly in that *AttendEM* uses ensembles with attention blocks but the satire versus fake news discriminator does not.

Beyond the proposals tailored to the particular sub-objectives laid out in Section 1.4, we propose holistic future research directions that advance our understanding of informational saturation attacks as a whole. One direction is to pursue the construction of a grand model that would account for all forms of informational saturation attacks at a societal level. Our present model, *Diluvision*, is limited by being only able to account for diffusion in a single information space (specifically, a single social media platform). The envisioned grand model would be able to, for instance, effectively simulate the decision making process of human agents being impacted by the interaction between a saturation attack of disguised advertisements on social media containing specific keywords and a saturation attack on e-commerce websites flooding them with product listings containing the aforementioned keywords. There have been tentative steps in the direction of a grand model, as evinced by multilayer network models [91, 523] of information diffusion that attempt to model the diffusion on one layer (e.g., mainstream mass media) impacting the diffusion on a different layer (e.g., social media).

Another direction is to be truly unconstrained and boldly imaginative in conceptualizing novel saturation attack vectors. Inundating the information space with duplicate product listing, though unconventional, is not an entirely never-before-imagined attack vector, as demonstrated by our ability to cite preceding works — the absence of an entirely novel saturation attack vector is one of our research program’s limitations. A truly unprecedented

saturation attack vector, if proposed and implemented early enough that no effective defenses against the attack exists, can help a state eliminate rivals and secure hegemonic status for many years to come.

Bibliography

- [1] Alexander Abad-Santos. This 13-Year-Old Is Scared When the Sky Is Blue Because of Our Drones. *The Atlantic*, October 2013. URL <https://www.theatlantic.com/politics/archive/2013/10/saddest-words-congress-briefing-drone-strikes/354548/>.
- [2] Daron Acemoğlu, Giacomo Como, Fabio Fagnani, and Asuman Ozdaglar. Opinion Fluctuations and Disagreement in Social Networks. *Mathematics of Operations Research*, 38(1):1–27, November 2012. ISSN 0364-765X. doi: 10.1287/moor.1120.0570. URL <https://pubsonline.informs-org.proxy3.library.mcgill.ca/doi/abs/10.1287/moor.1120.0570>.
- [3] Christian Adler, Stephan Eichler, Timo Kosch, Christoph Schroth, and Markus Strassberger. Self-organized and Context-Adaptive Information Diffusion in Vehicular Ad Hoc Networks. In *2006 3rd International Symposium on Wireless Communication Systems*, pages 307–311, September 2006. doi: 10.1109/ISWCS.2006.4362309.
- [4] Akshay Aggarwal, Aniruddha Chauhan, Deepika Kumar, Mamta Mittal, and Sharad Verma. Classification of Fake News by Fine-tuning Deep Bidirectional Transformers based Language Model. *EAI Endorsed Transactions on Scalable Information Systems*, 7 (27), April 2020. ISSN 2032-9407. URL <https://eudl.eu/doi/10.4108/eai.13-7-2018.163973>.
- [5] Naser Ahmadi, Hansjörg Sand, and Paolo Papotti. Unsupervised Matching of Data and Text. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1058–1070, May 2022. doi: 10.1109/ICDE53745.2022.00084.
- [6] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9, 2018. ISSN 2475-6725. doi: 10.1002/spy2.9. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9>.

-
- [7] Sadaf Ahmed, Monica Gentili, Daniel Sierra-Sosa, and Adel S. Elmaghraby. Multi-layer data integration technique for combining heterogeneous crime data. *Information Processing & Management*, 59(3):102879, May 2022. ISSN 0306-4573. doi: 10.1016/j.ipm.2022.102879. URL <https://www.sciencedirect.com/science/article/pii/S0306457322000115>.
- [8] Wasim Ahmed, Peter A. Bath, Laura Sbaffi, and Gianluca Demartini. Novel insights into views towards H1N1 during the 2009 Pandemic: A thematic analysis of Twitter data. *Health Information & Libraries Journal*, 36(1):60–72, 2019. ISSN 1471-1842. doi: 10.1111/hir.12247. URL <http://onlinelibrary.wiley.com/doi/abs/10.1111/hir.12247>.
- [9] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. COVID-19 and the 5G Conspiracy Theory: Social Network Analysis of Twitter Data. *Journal of Medical Internet Research*, 22(5):e19458, 2020. doi: 10.2196/19458. URL <https://www.jmir.org/2020/5/e19458/>.
- [10] Raphael Ahren. US funds aided 2015 campaign to oust Netanyahu, Senate probe finds. *Times of Israel*, 12 July 2016, 10:26 pm. URL <http://www.timesofisrael.com/us-funds-aided-2015-campaign-to-oust-netanyahu-senate-probe-finds/>.
- [11] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. People Are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):10–17, 2012. ISSN 2334-0770. doi: 10.1609/icwsm.v6i1.14236. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14236>.
- [12] Tim Aistrophe. Social media and counterterrorism strategy. *Australian Journal of International Affairs*, 70(2):121–138, March 2016. ISSN 1035-7718. doi: 10.1080/10357718.2015.1113230. URL <https://doi.org/10.1080/10357718.2015.1113230>.
- [13] AJ. Russia used all major social media platforms to aid Trump: Report. *Al Jazeera*, December 2018. URL <https://www.aljazeera.com/news/2018/12/17/russia-used-all-major-social-media-platforms-to-aid-trump-report>.
- [14] AJ. Amazon to suspend Parler after deadly Capitol Hill riot. *Al Jazeera*, January 2021. URL <https://www.aljazeera.com/news/2021/1/10/amazon-to-suspend-parler-after-deadly-capitol-hill-riot>.

-
- [15] Syeda Zainab Akbar, Anmol Panda, Divyanshu Kukreti, Azhagu Meena, and Joyojeet Pal. Misinformation as a Window into Prejudice: COVID-19 and the Information Environment in India. *Proceedings of the ACM on Human-Computer Interaction*, 4 (CSCW3):249:1–249:28, January 2021. doi: 10.1145/3432948. URL <https://doi.org/10.1145/3432948>.
- [16] Yaman Akdeniz. Anonymity, Democracy, and Cyberspace. *Social Research: An International Quarterly*, 69(1):223–237, March 2002. ISSN 1944-768X. doi: 10.1353/sor.2002.0010. URL <https://muse.jhu.edu/article/557276>.
- [17] Safinah Ali, Tyler Moroso, and Cynthia Breazeal. Can Children Learn Creativity from a Social Robot? In *Proceedings of the 2019 on Creativity and Cognition*, pages 359–368, San Diego CA USA, June 2019. ACM. ISBN 978-1-4503-5917-7. doi: 10.1145/3325480.3325499. URL <https://dl.acm.org/doi/10.1145/3325480.3325499>.
- [18] S. A. Alkhodair, S. H. H. Ding, B. C. M. Fung, and J. Liu. Detecting breaking news rumors of emerging topics in social media. *Information Processing and Management (IP&M)*, 57(2.102018):1–13, March 2020. doi: 10.1016/j.ipm.2019.02.016.
- [19] S. A. Alkhodair, B. C. M. Fung, S. H. H. Ding, W. K. Cheung, and S. C. Huang. Detecting high-engaging breaking news rumors in social media. *ACM Transactions on Management Information Systems (TMIS)*, 12(1):8.1–8.16, December 2020. doi: 10.1145/3416703.
- [20] Hunt Allcott and Matthew Gentzkow. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236, May 2017. ISSN 0895-3309. doi: 10.1257/jep.31.2.211. URL <http://www.aeaweb.org/articles?id=10.1257/jep.31.2.211>.
- [21] Sacha Altay, Anne-Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. Information delivered by a chatbot has a positive impact on COVID-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*, 29:52–62, 2023. ISSN 1939-2192. doi: 10.1037/xap0000400.
- [22] Mary Beth Altier, Christian N Thoroughgood, and John G Horgan. Turning away from terrorism: Lessons from psychology, sociology, and criminology. *Journal of Peace Research*, 51(5):647–661, September 2014. ISSN 0022-3433, 1460-3578. doi: 10.1177/0022343314535946. URL <http://journals.sagepub.com/doi/10.1177/0022343314535946>.

-
- [23] Amazon. 2022 Amazon Brand Protection Report. Technical report, Amazon, June 2022. URL <https://brandservices.amazon.com/progressreport>.
- [24] Monica Anderson and Nikolaos Papanikolopoulos. Implicit Cooperation Strategies for Multi-robot Search of Unknown Areas. *Journal of Intelligent and Robotic Systems*, 53(4):381–397, December 2008. ISSN 0921-0296, 1573-0409. doi: 10.1007/s10846-008-9242-5. URL <http://link.springer.com/10.1007/s10846-008-9242-5>.
- [25] Alessia Antelmi, Delfina Malandrino, and Vittorio Scarano. Characterizing the Behavioral Evolution of Twitter Users and The Truth Behind the 90-9-1 Rule. In *Companion Proceedings of The 2019 World Wide Web Conference, WWW '19*, pages 1035–1038, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3316705. URL <https://doi.org/10.1145/3308560.3316705>.
- [26] AP. US secretly created 'Cuban Twitter' to stir unrest and undermine government. *The Guardian*, April 2014. ISSN 0261-3077. URL <https://www.theguardian.com/world/2014/apr/03/us-cuban-twitter-zunzuneo-stir-unrest>.
- [27] Mohamed Arafa and Crystal Armstrong. "Facebook to Mobilize, Twitter to Coordinate Protests, and YouTube to Tell the World": New Media, Cyberactivism, and the Arab Spring. *Journal of Global Initiatives: Policy, Pedagogy, Perspective*, 10(1), January 2016. ISSN 1930-3009. URL <https://digitalcommons.kennesaw.edu/jgi/vol10/iss1/6>.
- [28] Sinan Aral and Paramveer S. Dhillon. Social influence maximization under empirical influence models. *Nature Human Behaviour*, 2(6):375–382, June 2018. ISSN 2397-3374. doi: 10.1038/s41562-018-0346-z. URL <http://www.nature.com/articles/s41562-018-0346-z>.
- [29] Patrick Arnold, Christian Wartner, and Erhard Rahm. Semi-Automatic Identification of Counterfeit Offers in Online Shopping Platforms. *Journal of Internet Commerce*, 15(1):59–75, January 2016. ISSN 1533-2861, 1533-287X. doi: 10.1080/15332861.2015.1121459. URL <http://www.tandfonline.com/doi/full/10.1080/15332861.2015.1121459>.
- [30] Ross Arnold, Kevin Carey, Benjamin Abruzzo, and Christopher Korpela. What is A Robot Swarm: A Definition for Swarming Robotics. In *2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0074–0081, October 2019. doi: 10.1109/UEMCON47517.2019.8993024.

-
- [31] Brooke Auxier and Monica Anderson. Social Media Use in 2021. Technical report, Pew Research Center, April 2021. URL <https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/>.
- [32] Aldo Averza, Khaled Slhoub, and Siddhartha Bhattacharyya. Evaluating the Influence of Twitter Bots via Agent-Based Social Simulation. *IEEE Access*, 10:129394–129407, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3228258.
- [33] Imran Awan. Cyber-Extremism: Isis and the Power of Social Media. *Society*, 54(2): 138–149, April 2017. ISSN 1936-4725. doi: 10.1007/s12115-017-0114-0. URL <https://doi.org/10.1007/s12115-017-0114-0>.
- [34] Robert Axelrod. The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *Journal of Conflict Resolution*, 41(2):203–226, April 1997. ISSN 0022-0027. doi: 10.1177/0022002797041002001. URL <https://doi.org/10.1177/0022002797041002001>.
- [35] Mehdi Azaouzi, Wassim Mnasri, and Lotfi Ben Romdhane. New trends in influence maximization models. *Computer Science Review*, 40:100393, May 2021. ISSN 1574-0137. doi: 10.1016/j.cosrev.2021.100393. URL <https://www.sciencedirect.com/science/article/pii/S1574013721000332>.
- [36] Jon Azpiri. B.C. satirical news website founder takes on fake news sites that ‘copied my articles’ — Globalnews.ca. *Global News*, December 13, 2016 10:16 pm. URL <https://globalnews.ca/news/3123282/b-c-satirical-news-website-founder-takes-on-fake-news-sites-that-copied-my-articles/>.
- [37] Jorge Bahamonde, Johan Bollen, Erick Elejalde, Leo Ferres, and Barbara Poblete. Power structure in Chilean news media. *PLOS ONE*, 13(6):e0197150, June 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0197150. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0197150>.
- [38] Vian Bakir. Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica’s Psychographic Profiling and Targeting. *Frontiers in Communication*, 5, 2020. ISSN 2297-900X. URL <https://www.frontiersin.org/articles/10.3389/fcomm.2020.00067>.
- [39] Anna Bałazy, Mika Toivola, Atin Adhikari, Satheesh K. Sivasubramani, Tiina Reponen, and Sergey A. Grinshpun. Do N95 respirators provide 95% protection level against airborne viruses, and how adequate are surgical masks? *American Journal of Infection*

- Control*, 34(2):51–57, March 2006. ISSN 01966553. doi: 10.1016/j.ajic.2005.08.018. URL <https://linkinghub.elsevier.com/retrieve/pii/S0196655305009119>.
- [40] James Bamford. The Trump Campaign’s Collusion With Israel. *The Nation*, March 2023. ISSN 0027-8378. URL <https://www.thenation.com/article/world/trump-israel-collusion/>.
- [41] Jack Bandy and Nicholas Diakopoulos. More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–28, April 2021. ISSN 2573-0142. doi: 10.1145/3449152. URL <https://dl.acm.org/doi/10.1145/3449152>.
- [42] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439):509–512, October 1999. doi: 10.1126/science.286.5439.509. URL <https://www.science.org/doi/10.1126/science.286.5439.509>.
- [43] Rodrigo Barbado, Oscar Araque, and Carlos A. Iglesias. A framework for fake review detection in online consumer electronics retailers. *Information Processing & Management*, 56(4):1234–1244, July 2019. ISSN 03064573. doi: 10.1016/j.ipm.2019.03.002. URL <http://arxiv.org/abs/1903.12452>.
- [44] Matthew Barnfield. Think Twice before Jumping on the Bandwagon: Clarifying Concepts in Research on the Bandwagon Effect. *Political Studies Review*, 18(4): 553–574, November 2020. ISSN 1478-9299, 1478-9302. doi: 10.1177/1478929919870691. URL <http://journals.sagepub.com/doi/10.1177/1478929919870691>.
- [45] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5460–5469, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00539. URL <https://ieeexplore.ieee.org/document/9878829/>.
- [46] Larry M. Bartels. *Unequal Democracy: The Political Economy of the New Gilded Age - Second Edition*. Princeton University Press, September 2016. ISBN 978-1-4008-8336-3. doi: 10.1515/9781400883363. URL <https://www.degruyter.com/document/doi/10.1515/9781400883363/html>.
- [47] Anomadarshi Barua and Mohammad Abdullah Al Faruque. PreMSat: Preventing Magnetic Saturation Attack on Hall Sensors. *IACR Transactions on Cryptographic*

- Hardware and Embedded Systems*, pages 438–462, August 2022. ISSN 2569-2925. doi: 10.46586/tches.v2022.i4.438-462. URL <https://tches.iacr.org/index.php/TCHES/article/view/9827>.
- [48] Veronika Batzdorfer, Holger Steinmetz, Marco Biella, and Meysam Alizadeh. Conspiracy theories on Twitter: Emerging motifs and temporal dynamics during the COVID-19 pandemic. *International Journal of Data Science and Analytics*, December 2021. ISSN 2364-4168. doi: 10.1007/s41060-021-00298-6. URL <https://doi.org/10.1007/s41060-021-00298-6>.
- [49] Asef Bayat. *Revolution without Revolutionaries : Making Sense of the Arab Spring*. Stanford Studies in Middle Eastern and Islamic Societies and Cultures. Stanford University Press, Stanford, CA, 2020. ISBN 978-1-5036-0307-3 1-5036-0307-5. doi: 10.1515/9781503603073. URL <https://doi.org/10.1515/9781503603073>.
- [50] BBC. Cambridge Analytica 'not involved' in Brexit referendum, says watchdog. *BBC News*, October 2020. URL <https://www.bbc.com/news/uk-politics-54457407>.
- [51] Amy B. Becker and Leticia Bode. Satire as a source for learning? The differential impact of news versus satire exposure on net neutrality knowledge gain. *Information, Communication & Society*, 21(4):612–625, April 2018. ISSN 1369-118X. doi: 10.1080/1369118X.2017.1301517. URL <https://doi.org/10.1080/1369118X.2017.1301517>.
- [52] Nicholas J. Belkin and Stephen E. Robertson. Information science and the phenomenon of information. *Journal of the American Society for Information Science*, 27(4):197–204, 1976. ISSN 1097-4571. doi: 10.1002/asi.4630270402. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.4630270402>.
- [53] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL <https://www.aclweb.org/anthology/D19-1371>.
- [54] Omer Benjakob and Gur Megiddo. Hacking, Extortion, Election Interference: These Are the Tools Used by Israel's Agents of Chaos and Manipulation. *Haaretz*, February 2023. URL <https://www.haaretz.com/israel-news/security-aviation/2023>

- 02-15/ty-article-magazine/.premium/hacking-extortion-election-interference-the-toolkit-of-israels-agents-of-chaos/00000186-4aa6-d933-af9e-cbe7aa9c0000.
- [55] Daniel Bennett. Exploring the impact of an evolving war and terror blogosphere on traditional media coverage of conflict. *Media, War & Conflict*, 6(1):37–53, April 2013. ISSN 1750-6352, 1750-6360. doi: 10.1177/1750635212469907. URL <http://journals.sagepub.com/doi/10.1177/1750635212469907>.
- [56] Besedo. Duplicate Real-Estate Listings. Why Should You Care?, July 2016. URL <https://besedo.com/resources/blog/duplicate-real-estate-listings-why-should-you-care/>.
- [57] David M. Beskow and Kathleen M. Carley. Agent Based Simulation of Bot Disinformation Maneuvers in Twitter. In *2019 Winter Simulation Conference (WSC)*, pages 750–761, December 2019. doi: 10.1109/WSC40007.2019.9004942.
- [58] Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive Influence Maximization in Social Networks. In Xiaotie Deng and Fan Chung Graham, editors, *Internet and Network Economics*, Lecture Notes in Computer Science, pages 306–311, Berlin, Heidelberg, 2007. Springer. ISBN 978-3-540-77105-0. doi: 10.1007/978-3-540-77105-0_31.
- [59] Chetan Bhatt. White Extinction: Metaphysical Elements of Contemporary Western Fascism. *Theory, Culture & Society*, 38(1):27–52, January 2021. ISSN 0263-2764, 1460-3616. doi: 10.1177/0263276420925523. URL <http://journals.sagepub.com/doi/10.1177/0263276420925523>.
- [60] Constance Bitso, Ina Fourie, and Theo J. D. Bothma. Trends in transition from classical censorship to Internet censorship : Selected country overviews. *Innovation : journal of appropriate librarianship and information work in Southern Africa*, 2013 (46):166–191, June 2013. doi: 10.10520/EJC148135. URL <https://journals.co.za/doi/abs/10.10520/EJC148135>.
- [61] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models, April 2023. URL <http://arxiv.org/abs/2304.08818>.
- [62] Daniel Boguslaw, Sam Biddle, and Ken Klippenstein. Army Unit for Protecting

- Generals Spies on Social Media. *The Intercept*, June 2023. URL <https://archive.is/uHzLa>.
- [63] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, June 2017. ISSN 2307-387X. doi: 10.1162/tac1_a_00051. URL https://doi.org/10.1162/tac1_a_00051.
- [64] Göran Bolin, Paul Jordan, and Per Ståhlberg. From Nation Branding to Information Warfare: Management of Information in the Ukraine-Russia Conflict. In Mervi Pantti, editor, *Media and the Ukraine Crisis: Hybrid Media Practices and Narratives of Conflict*, pages 3–18. Peter Lang, New York, 2016.
- [65] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '03, pages 132–139, USA, January 2003. Society for Industrial and Applied Mathematics. ISBN 978-0-89871-538-5.
- [66] Enrique Bonsón, David Perea, and Michaela Bednárová. Twitter as a tool for citizen engagement: An empirical study of the Andalusian municipalities. *Government Information Quarterly*, 36(3):480–489, July 2019. ISSN 0740-624X. doi: 10.1016/j.giq.2019.03.001. URL <http://www.sciencedirect.com/science/article/pii/S0740624X18304684>.
- [67] Allan Borodin, Yuval Filmus, and Joel Oren. Threshold Models for Competitive Influence in Social Networks. In Amin Saberi, editor, *Internet and Network Economics*, Lecture Notes in Computer Science, pages 539–550, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-17572-5. doi: 10.1007/978-3-642-17572-5_48.
- [68] Angela K. Bourne and Fernando Casal Bértoa. Mapping ‘Militant Democracy’: Variation in Party Ban Practices in European Democracies (1945-2015). *European Constitutional Law Review*, 13(2):221–247, June 2017. ISSN 1574-0196, 1744-5515. doi: 10.1017/S1574019617000098. URL https://www.cambridge.org/core/product/identifier/S1574019617000098/type/journal_article.
- [69] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard. Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation. Working Paper 2021.1, Oxford Internet Institute, Oxford, UK, 2021. URL <https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2021/02/CyberTroop-Report20-Draft9.pdf>.

- [70] David R. Brake. The Invisible Hand of the Unaccountable Algorithm: How Google, Facebook and Other Tech Companies Are Changing Journalism. In Jingrong Tong and Shih-Hung Lo, editors, *Digital Technology and Journalism*, pages 25–46. Springer International Publishing, Cham, 2017. ISBN 978-3-319-55025-1 978-3-319-55026-8. doi: 10.1007/978-3-319-55026-8_2. URL http://link.springer.com/10.1007/978-3-319-55026-8_2.
- [71] Paul R. Brewer, Dannagal Goldthwaite Young, and Michelle Morreale. The Impact of Real News about “Fake News”: Intertextual Processes and Political Satire. *International Journal of Public Opinion Research*, 25(3):323–343, September 2013. ISSN 0954-2892. doi: 10.1093/ijpor/edt015. URL <https://academic-oup-com.proxy3.library.mcgill.ca/ijpor/article/25/3/323/786961>.
- [72] Mayumi Brewster. Annual Retail Trade Survey Shows Impact of Online Shopping on Retail Sales During COVID-19 Pandemic. Technical report, United States Census Bureau, April 2022. URL <https://www.census.gov/library/stories/2022/04/e-commerce-sales-surged-during-pandemic.html>.
- [73] Anna D. Broido and Aaron Clauset. Scale-free networks are rare. *Nature Communications*, 10(1):1017, March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08746-5. URL <https://www.nature.com/articles/s41467-019-08746-5>.
- [74] John T. Brooks and Jay C. Butler. Effectiveness of Mask Wearing to Control Community Spread of SARS-CoV-2. *JAMA*, 325(10):998, March 2021. ISSN 0098-7484. doi: 10.1001/jama.2021.1505. URL <https://jamanetwork.com/journals/jama/fullarticle/2776536>.
- [75] Rogers Brubaker. Digital hyperconnectivity and the self. *Theory and Society*, 49(5-6): 771–801, October 2020. ISSN 0304-2421, 1573-7853. doi: 10.1007/s11186-020-09405-1. URL <https://link.springer.com/10.1007/s11186-020-09405-1>.
- [76] Lachlan Brumley, Carlo Kopp, and Kevin B. Korb. Cutting Through the Tangled Web: An Information-Theoretic Perspective on Information Warfare. *Air Power Australia Analyses*, IX(2):1, October 2012. ISSN 1832-2433. URL <http://www.ausairpower.net/APA-2012-02.html>.
- [77] Ursin Brunner and Kurt Stockinger. Entity matching with transformer architectures - a step forward in data integration. In *Proceedings of the 23rd International Conference on Extending Database Technology (EDBT)*, pages 463–473. OpenProceedings, March

2020. ISBN 978-3-89318-083-7. doi: 10.21256/zhaw-19637. URL <https://digitalcollection.zhaw.ch/handle/11475/19637>.
- [78] John M. Budd. Phenomenology and information studies. *Journal of Documentation*, 61(1):44–59, February 2005. ISSN 0022-0418. doi: 10.1108/00220410510578005. URL <https://www.emerald.com/insight/content/doi/10.1108/00220410510578005/full/html>.
- [79] Clint Burfoot and Timothy Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*, page 161, Suntec, Singapore, 2009. Association for Computational Linguistics. doi: 10.3115/1667583.1667633. URL <http://portal.acm.org/citation.cfm?doid=1667583.1667633>.
- [80] Kelly P. Burke and Andrea L. Cox. Hepatitis C virus evasion of adaptive immune responses: A model for viral persistence. *Immunologic Research*, 47(1):216–227, July 2010. ISSN 1559-0755. doi: 10.1007/s12026-009-8152-3. URL <https://doi.org/10.1007/s12026-009-8152-3>.
- [81] Alex Burns and Ben Eltham. Twitter free Iran: An evaluation of Twitter’s role in public diplomacy and information operations in Iran’s 2009 election crisis. In *Record of the Communications Policy and Research Forum 2009*, Sydney, Australia, November 2009.
- [82] Desmond Butler, Jack Gillum, and Alberto Arce. US secretly built ‘Cuban Twitter’ to stir unrest. *Associated Press*, April 2014. URL <https://apnews.com/article/technology-cuba-united-states-government-904a9a6a1bcd46cebfc14bea2ee30fdf>.
- [83] Matthew Byrne. Drone Skies. *Current Affairs*, March/April 2022, June 2022. ISSN 2471-2647. URL <https://www.currentaffairs.org/2022/06/drone-skies>.
- [84] Alex Calderwood, Erin Riglin, and Shreya Vaidyanathan. How Americans Wound Up on Twitter’s List of Russian Bots. *Wired*, July 2018. ISSN 1059-1028. URL <https://www.wired.com/story/how-americans-wound-up-on-twitters-list-of-russian-bots/>.
- [85] Roy Campbell, Jalal Al-Muhtadi, Prasad Naldurg, Geetanjali Sampemane, and M. Dennis Mickunas. Towards Security and Privacy for Pervasive Computing. In Gerhard Goos, Juris Hartmanis, Jan van Leeuwen, Mitsuhiro Okada, Benjamin C. Pierce, Andre Scedrov, Hideyuki Tokuda, and Akinori Yonezawa, editors, *Software*

- Security — Theories and Systems*, volume 2609, pages 1–15. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003. ISBN 978-3-540-00708-1 978-3-540-36532-7. doi: 10.1007/3-540-36532-X_1. URL http://link.springer.com/10.1007/3-540-36532-X_1.
- [86] Longbing Cao and Qing Liu. COVID-19 Modeling: A Review. *arXiv:2104.12556 [cs]*, August 2021. URL <http://arxiv.org/abs/2104.12556>.
- [87] Kathleen M. Carley. Social cybersecurity: An emerging science. *Computational and Mathematical Organization Theory*, 26(4):365–381, December 2020. ISSN 1572-9346. doi: 10.1007/s10588-020-09322-9. URL <https://doi.org/10.1007/s10588-020-09322-9>.
- [88] Tim Carnes, Chandrashekhara Nagarajan, Stefan M. Wild, and Anke van Zuylen. Maximizing influence in a competitive social network: A follower’s perspective. In *Proceedings of the Ninth International Conference on Electronic Commerce, ICEC ’07*, pages 351–360, New York, NY, USA, August 2007. Association for Computing Machinery. ISBN 978-1-59593-700-1. doi: 10.1145/1282100.1282167. URL <https://dl.acm.org/doi/10.1145/1282100.1282167>.
- [89] Vincenzo Carrieri, Leonardo Madio, and Francesco Principe. Vaccine hesitancy and (fake) news: Quasi-experimental evidence from Italy. *Health Economics*, 28(11):1377–1382, 2019. ISSN 1099-1050. doi: 10.1002/hec.3937. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hec.3937>.
- [90] David J. Caspi, Joshua D. Freilich, and Steven M. Chermak. Worst of the bad: Violent white supremacist groups and lethality. *Dynamics of Asymmetric Conflict*, 5(1):1–17, March 2012. ISSN 1746-7586. doi: 10.1080/17467586.2012.679664. URL <https://doi.org/10.1080/17467586.2012.679664>.
- [91] Ho-Chun Herbert Chang and Feng Fu. Co-diffusion of social contagions. *New Journal of Physics*, 20(9):095001, September 2018. ISSN 1367-2630. doi: 10.1088/1367-2630/aadce7. URL <https://iopscience.iop.org/article/10.1088/1367-2630/aadce7>.
- [92] Sarah Chayes. The Strategies Are Foreign, but the Corruption Is American: A Response to “The Rise of Strategic Corruption”. *Foreign Affairs*, 99(6):167–170, November/December 2020. ISSN 0015-7120. URL <https://www.foreignaffairs.com/articles/united-states/2020-10-13/strategies-are-foreign-corruption-american>.

- [93] Hsuan-Ting Chen, Chen Gan, and Ping Sun. How Does Political Satire Influence Political Participation? Examining the Role of Counter- and Pro-Attitudinal Exposure, Anger, and Personal Issue Importance. *International Journal of Communication*, 11(0):19, July 2017. ISSN 1932-8036. URL <https://ijoc.org/index.php/ijoc/article/view/6158>.
- [94] Li Chen and Fengxia Zhu. Seller information sharing in online marketplaces. *Journal of Marketing Management*, 37(11-12):1191–1218, July 2021. ISSN 0267-257X. doi: 10.1080/0267257X.2021.1900895. URL <https://doi.org/10.1080/0267257X.2021.1900895>.
- [95] P. Chen and S. Redner. Majority rule dynamics in finite dimensions. *Physical Review E*, 71(3):036101, March 2005. doi: 10.1103/PhysRevE.71.036101. URL <https://link.aps.org/doi/10.1103/PhysRevE.71.036101>.
- [96] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1053. URL <https://www.aclweb.org/anthology/N19-1053>.
- [97] Vincent Chi-Chung Cheng, Shuk-Ching Wong, Vivien Wai-Man Chuang, Simon Yung-Chun So, Jonathan Hon-Kwan Chen, Siddharth Sridhar, Kelvin Kai-Wang To, Jasper Fuk-Woo Chan, Ivan Fan-Ngai Hung, Pak-Leung Ho, and Kwok-Yung Yuen. The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2. *Journal of Infection*, 81(1):107–114, July 2020. ISSN 0163-4453. doi: 10.1016/j.jinf.2020.04.024. URL <http://www.sciencedirect.com/science/article/pii/S0163445320302358>.
- [98] Cynthia Chew and Gunther Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLOS ONE*, 5(11):e14118, November 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0014118. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014118>.
- [99] Joohn Choe. Let me explain something about NAFO as simply as I can: NAFO is a LEADERLESS MOVEMENT that does only two things: 1. fights disinfo 2. fundraise for Ukraine #1 means we ratio people like this guy. That’s low for us. It’ll hit probably

- 1.5 - 2:1 when all is said and done. 1/3 <https://t.co/d4m1ydqCvp>, October 2022. URL <https://twitter.com/JoohnChoe/status/1585099669851746304>.
- [100] Joohn Choe. JChoe (@JoohnChoe) / Twitter, January 2023. URL <https://twitter.com/JoohnChoe>.
- [101] Joohn Choe. Joohn Choe - Disinformation & extremism researcher and consultant - Freelance — LinkedIn, April 2023. URL <https://www.linkedin.com/in/joohnchoe>.
- [102] Noam Chomsky. Noam Chomsky: The New York Times is pure propaganda, May 2015. URL https://www.salon.com/2015/05/25/noam_chomsky_the_new_york_times_is_pure_proganda_partner/.
- [103] Dennis Chong and James N. Druckman. Framing Theory. *Annual Review of Political Science*, 10(1):103–126, June 2007. ISSN 1094-2939, 1545-1577. doi: 10.1146/annurev.polisci.10.072805.103054. URL <https://www.annualreviews.org/doi/10.1146/annurev.polisci.10.072805.103054>.
- [104] Peter Christen. Febrl -: An open source data cleaning, deduplication and record linkage system with a graphical user interface. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 1065–1068, New York, NY, USA, August 2008. Association for Computing Machinery. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1402020. URL <https://doi.org/10.1145/1401890.1402020>.
- [105] Derek K. Chu, Elie A. Akl, Stephanie Duda, Karla Solo, Sally Yaacoub, Holger J. Schünemann, Derek K. Chu, Elie A. Akl, Amena El-harakeh, Antonio Bognanni, Tamara Lotfi, Mark Loeb, Anisa Hajizadeh, Anna Bak, Ariel Izcovich, Carlos A. Cuello-Garcia, Chen Chen, David J. Harris, Ewa Borowiack, Fatimah Chamseddine, Finn Schünemann, Gian Paolo Morgano, Giovanna E. U. Muti Schünemann, Guang Chen, Hong Zhao, Ignacio Neumann, Jeffrey Chan, Joanne Khabsa, Layal Hneiny, Leila Harrison, Maureen Smith, Nesrine Rizk, Paolo Giorgi Rossi, Pierre AbiHanna, Rayane El-khoury, Rosa Stalteri, Tejan Baldeh, Thomas Piggott, Yuan Zhang, Zahra Saad, Assem Khamis, Marge Reinap, Stephanie Duda, Karla Solo, Sally Yaacoub, and Holger J. Schünemann. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: A systematic review and meta-analysis. *The Lancet*, 395(10242):1973–1987, June 2020. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(20)31142-9. URL [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(20\)31142-9/abstract](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31142-9/abstract).

- [106] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *Scientific Reports*, 10(1):16598, October 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-73510-5. URL <https://www.nature.com/articles/s41598-020-73510-5>.
- [107] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, March 2021. doi: 10.1073/pnas.2023301118. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2023301118>.
- [108] William R. Clark. *Petrodollar Warfare : Oil, Iraq and the Future of the Dollar*. NSP: New Society Publishers, 2005. ISBN 978-1-55092-335-3 978-0-86571-514-1. URL <https://canadacommons.ca/artifacts/1870198/petrodollar-warfare/2619327/>.
- [109] Jean Claude Ameisen and Andre Capron. Cell dysfunction and depletion in AIDS: The programmed cell death hypothesis. *Immunology Today*, 12(4):102–105, April 1991. ISSN 0167-5699. doi: 10.1016/0167-5699(91)90092-8. URL <https://www.sciencedirect.com/science/article/pii/0167569991900928>.
- [110] Peter Clifford and Aidan Sudbury. A Model for Spatial Conflict. *Biometrika*, 60(3):581–588, 1973. ISSN 0006-3444. doi: 10.2307/2335008. URL <https://www.jstor.org/stable/2335008>.
- [111] Ian Cobain and Nick Fielding. Revealed: US spy operation that manipulates social media. *The Guardian*, March 2011. ISSN 0261-3077. URL <https://www.theguardian.com/technology/2011/mar/17/us-spy-operation-social-networks>.
- [112] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and Measuring the Geometry of BERT. *arXiv:1906.02715 [cs, stat]*, October 2019. URL <http://arxiv.org/abs/1906.02715>.
- [113] B. Collard and F. X. Standaert. A Statistical Saturation Attack against the Block Cipher PRESENT. In Marc Fischlin, editor, *Topics in Cryptology – CT-RSA 2009*, Lecture Notes in Computer Science, pages 195–210, Berlin, Heidelberg, 2009. Springer. ISBN 978-3-642-00862-7. doi: 10.1007/978-3-642-00862-7_13.

- [114] Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro. Review of social science research on the impact of countermeasures against influence operations. *Harvard Kennedy School Misinformation Review*, September 2021. doi: 10.37016/mr-2020-79. URL <https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/>.
- [115] Dan Craigen, Nadia Diakun-Thibault, and Randy Purse. Defining Cybersecurity. *Technology Innovation Management Review*, 4(10):13–21, October 2014. URL <https://www.proquest.com/docview/1638205509/abstract/27EF00BA55AC44F7PQ/1>.
- [116] Matthieu Crozet and Julian Hinz. Friendly fire: The trade impact of the Russia sanctions and counter-sanctions. *Economic Policy*, 35(101):97–146, January 2020. ISSN 0266-4658. doi: 10.1093/epolic/eiaa006. URL <https://doi.org/10.1093/epolic/eiaa006>.
- [117] Christopher Brian Currin, Sebastián Vallejo Vera, and Ali Khaledi-Nasab. Depolarization of echo chambers by random dynamical nudge. *Scientific Reports*, 12(1): 9234, June 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-12494-w. URL <https://www.nature.com/articles/s41598-022-12494-w>.
- [118] D. J. Daley and D. G. Kendall. Epidemics and Rumours. *Nature*, 204(4963):1118–1118, December 1964. ISSN 1476-4687. doi: 10.1038/2041118a0. URL <https://www.nature.com/articles/2041118a0>.
- [119] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised User Stance Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:141–152, May 2020. ISSN 2334-0770. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/7286>.
- [120] Abhimanyu Das, Sreenivas Gollapudi, and Kamesh Munagala. Modeling opinion dynamics in social networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14*, pages 403–412, New York, NY, USA, February 2014. Association for Computing Machinery. ISBN 978-1-4503-2351-2. doi: 10.1145/2556195.2559896. URL <https://doi.org/10.1145/2556195.2559896>.
- [121] Dipto Das and Anthony J. Clark. Satire vs Fake News: You Can Tell by the Way They Say It. In *2019 First International Conference on Transdisciplinary AI (TransAI)*, pages 22–26, September 2019. doi: 10.1109/TransAI46475.2019.00012.

- [122] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P. Gummadi. When the Umpire is also a Player: Bias in Private Label Product Recommendations on E-commerce Marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 873–884, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445944. URL <https://dl.acm.org/doi/10.1145/3442188.3445944>.
- [123] Lynn Davidman and Arthur L. Greil. Characters in Search of a Script: The Exit Narratives of Formerly Ultra-Orthodox Jews. *Journal for the Scientific Study of Religion*, 46(2):201–216, 2007. ISSN 1468-5906. doi: 10.1111/j.1468-5906.2007.00351.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-5906.2007.00351.x>.
- [124] Carl Davidson and Jerry Harris. Globalisation, theocracy and the new fascism: The US Right's rise to power. *Race & Class*, 47(3):47–67, January 2006. ISSN 0306-3968, 1741-3125. doi: 10.1177/0306396806061086. URL <http://journals.sagepub.com/doi/10.1177/0306396806061086>.
- [125] Rick Davies and Logan Mudge. Is it all downhill from here for NZ's e-commerce giant? Hell no, says Trade Me. *The Spinoff*, September 2018. URL <https://thespinoff.co.nz/business/29-09-2018/is-it-all-downhill-from-here-for-trademe/>.
- [126] Maira A de C Gatti, Ana Paula Appel, Cicero Nogueira dos Santos, Claudio Santos Pinhanetz, Paulo Rodrigo Cavalin, and Samuel Barbosa Neto. A simulation-based approach to analyze the information diffusion in Microblogging Online Social Network. In *2013 Winter Simulations Conference (WSC)*, pages 1685–1696, December 2013. doi: 10.1109/WSC.2013.6721550.
- [127] Janaína Ignácio de Moraes, Hugo Queiroz Abonizio, Gabriel Marques Tavares, André Azevedo da Fonseca, and Sylvio Barbon. Deciding among Fake, Satirical, Objective and Legitimate news: A multi-label classification system. In *Proceedings of the XV Brazilian Symposium on Information Systems, SBSI'19*, pages 1–8, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-7237-4. doi: 10.1145/3330204.3330231. URL <https://doi.org/10.1145/3330204.3330231>.
- [128] Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. Attending Sentences to detect Satirical Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3371–3380, Santa Fe, New Mexico, USA, August 2018.

- Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/C18-1285>.
- [129] Isabel Debre. Israeli company targeted Nigerian election in Facebook disinformation campaign. *Times of Israel*, 18 May 2019, 4:16am. URL <https://www.timesofisrael.com/israeli-company-targeted-nigerian-election-in-facebook-disinformation-campaign/>.
- [130] Bernd Debusmann. United States lags in propaganda war against ISIS — Bernd Debusmann. *The Arab Weekly*, June 2015. URL <https://thearabweekly.com/unit-ed-states-lags-propaganda-war-against-isis>.
- [131] Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. *Personality and Social Psychology Review*, 14(2):238–257, May 2010. ISSN 1088-8683, 1532-7957. doi: 10.1177/1088868309352251. URL <http://journals.sagepub.com/doi/10.1177/1088868309352251>.
- [132] Julia R. DeCook and Mi Hyun Yoon. Kung Flu and Roof Koreans: Asian/Americans as the Hated Other and Proxies of Hating in the White Imaginary. *Journal of Hate Studies*, 17(1):119–132, October 2021. ISSN 2169-7442. doi: 10.33972/jhs.199. URL https://jhs.press.gonzaga.edu/articles/10.33972/jhs.199?utm_source=TrendMD&utm_medium=cpc&utm_campaign=Journal_of_Hate_Studies_TrendMD_1.
- [133] Morris H. DeGroot. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345):118–121, March 1974. ISSN 0162-1459. doi: 10.1080/01621459.1974.10480137. URL <https://www.tandfonline.com/doi/abs/10.1080/01621459.1974.10480137>.
- [134] Michael S Deiner, Cherie Fathy, Jessica Kim, Katherine Niemeyer, David Ramirez, Sarah F Ackley, Fengchen Liu, Thomas M Lietman, and Travis C Porco. Facebook and Twitter vaccine sentiment in response to measles outbreaks. *Health Informatics Journal*, 25(3):1116–1132, September 2019. ISSN 1460-4582. doi: 10.1177/1460458217740723. URL <https://doi.org/10.1177/1460458217740723>.
- [135] L. DeNardis and A.M. Hackl. Internet governance by social media platforms. *Telecommunications Policy*, 39(9):761–770, October 2015. ISSN 03085961. doi: 10.1016/j.telpol.2015.04.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0308596115000592>.

- [136] Carlos Andres Devia and Giulia Giordano. A framework to analyze opinion formation models. *Scientific Reports*, 12(1):13441, August 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-17348-z. URL <https://www.nature.com/articles/s41598-022-17348-z>.
- [137] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota, 2019.
- [138] Dictionary.com. Ratio, May 2021. URL <https://www.dictionary.com/browse/ratio>.
- [139] Dominic DiFranzo and Kristine Gloria-Garcia. Filter Bubbles and Fake News. *XRDS: Crossroads, The ACM Magazine for Students*, 23(3):32–35, April 2017. ISSN 1528-4972. doi: 10.1145/3055153. URL <http://doi.acm.org/10.1145/3055153>.
- [140] Dimitar Dimitrov, Erdal Baran, Pavlos Fafalios, Ran Yu, Xiaofei Zhu, Matthäus Zloch, and Stefan Dietze. TweetsCOVID - A Knowledge Base of Semantically Annotated Tweets about the COVID-19 Pandemic. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2991–2998, Virtual Event, Ireland, October 2020. Association for Computing Machinery. ISBN 978-1-4503-6859-9. URL <https://doi.org/10.1145/3340531.3412765>.
- [141] Ly Dinh and Nikolaus Parulian. COVID-19 pandemic and information diffusion analysis on Twitter. *Proceedings of the Association for Information Science and Technology*, 57(1):e252, 2020. ISSN 2373-9231. doi: 10.1002/pra2.252. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.252>.
- [142] Chenxiao Dou, Yi Cui, Daniel Sun, Raymond Wong, Muhammad Atif, Guoqiang Li, and Rajiv Ranjan. Unsupervised blocking and probabilistic parallelisation for record matching of distributed big data. *The Journal of Supercomputing*, 75(2): 623–645, February 2019. ISSN 1573-0484. doi: 10.1007/s11227-017-2008-8. URL <https://doi.org/10.1007/s11227-017-2008-8>.
- [143] Melissa-Ellen Dowling. Cyber information operations: Cambridge Analytica’s challenge to democratic legitimacy. *Journal of Cyber Policy*, 7(2):230–248, May 2022. ISSN 2373-8871. doi: 10.1080/23738871.2022.2081089. URL <https://doi.org/10.1080/23738871.2022.2081089>.

- [144] Clare Duffy and Brian Fung. Elon Musk commissioned this bot analysis in his fight with Twitter. Now it shows what he could face if he takes over the platform — CNN Business. *CNN*, October 2022. URL <https://www.cnn.com/2022/10/10/tech/elon-musk-twitter-bot-analysis-cyabra/index.html>.
- [145] Gregory Eady, Tom Paskhalis, Jan Zilinsky, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature Communications*, 14(1):62, January 2023. ISSN 2041-1723. doi: 10.1038/s41467-022-35576-9. URL <https://www.nature.com/articles/s41467-022-35576-9>.
- [146] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq Joty, Mourad Ouzani, and Nan Tang. Distributed Representations of Tuples for Entity Resolution. *Proc. VLDB Endow.*, 11(11):1454–1467, July 2018. ISSN 2150-8097. doi: 10.14778/3236187.3269461. URL <https://doi.org/10.14778/3236187.3269461>.
- [147] Chad Edwards, Autumn Edwards, Patric R. Spence, and Ashleigh K. Shelton. Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior*, 33:372–376, April 2014. ISSN 07475632. doi: 10.1016/j.chb.2013.08.013. URL <https://linkinghub.elsevier.com/retrieve/pii/S0747563213003129>.
- [148] EIU. The Economist Intelligence Unit’s Democracy Index 2016. Technical report, The Economist Intelligence Unit, 2017. URL <https://impact.economist.com/perspectives/economic-development/revenge-deplorables/white-paper/eius-2016-democracy-index>.
- [149] EIU. The Economist Intelligence Unit’s Democracy Index 2022. Technical report, The Economist Intelligence Unit, 2023. URL <https://www.eiu.com/n/campaigns/democracy-index-2022/>.
- [150] Norman E. Emery, Robert S. Earl, and Raymond Buettner. Terrorist Use of Information Operations. *Journal of Information Warfare*, 13(2):14, 2014. ISSN 14453312. URL <https://www.proquest.com/docview/2164442205/abstract/B4766C3DE2FE464FPQ/2>.
- [151] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-*

- Computer Interaction*, 1(CSCW):1–22, December 2017. ISSN 2573-0142. doi: 10.1145/3134677. URL <https://dl.acm.org/doi/10.1145/3134677>.
- [152] Ziv Epstein, Hause Lin, Gordon Pennycook, and David Rand. How many others have shared this? Experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media, July 2022. URL <http://arxiv.org/abs/2207.07562>.
- [153] Liana Ermakova, Diana Nurbakova, and Irina Ovchinnikova. Analysis of Users Engaged in Online Discussions about Controversial Covid-19 Treatments. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 162–166, Utrecht, Netherlands, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8367-7. URL <https://doi.org/10.1145/3450614.3462239>.
- [154] Mohammad Eslami. Iran’s Drone Supply to Russia and Changing Dynamics of the Ukraine War. *Journal for Peace and Nuclear Disarmament*, 5(2):507–518, July 2022. ISSN null. doi: 10.1080/25751654.2022.2149077. URL <https://doi.org/10.1080/25751654.2022.2149077>.
- [155] Richard Falk and Virginia Tilley. Israeli Practices towards the Palestinian People and the Question of Apartheid, Executive Summary. *Middle East Policy*, 24(2):33–40, 2017. ISSN 1475-4967. doi: 10.1111/mepo.12265. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/mepo.12265>.
- [156] Patrizia Farci, Harvey J. Alter, Sugantha Govindarajan, Doris C. Wong, Ronald Engle, Richard R. Lesniewski, Isa K. Mushahwar, Suresh M. Desai, Roger H. Miller, Norio Ogata, and Robert H. Purcell. Lack of Protective Immunity Against Reinfection with Hepatitis C Virus. *Science*, 258(5079):135–140, October 1992. doi: 10.1126/science.1279801. URL <https://www.science.org/doi/abs/10.1126/science.1279801>.
- [157] Stephen Farrall, Stuart Wilks-Heeg, Robert Struthers, and Emily Gray. Who are the victims of electoral fraud in Great Britain? Evidence from survey research. *British Politics*, 17(3):333–352, September 2022. ISSN 1746918X. doi: 10.1057/s41293-021-00189-1. URL <https://www.proquest.com/docview/2701342210/24F35A44DDA34416PQ/1>.
- [158] Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144:993–1002, 2015. ISSN 1939-2222. doi: 10.1037/xge0000098.

- [159] Leonor Fernández, Alan Fossa, Zhiyong Dong, Tom Delbanco, Joann Elmore, Patricia Fitzgerald, Kendall Harcourt, Jocelyn Perez, Jan Walker, and Catherine DesRoches. Words Matter: What Do Patients Find Judgmental or Offensive in Outpatient Notes? *Journal of General Internal Medicine*, 36(9):2571–2578, September 2021. ISSN 1525-1497. doi: 10.1007/s11606-020-06432-7. URL <https://doi.org/10.1007/s11606-020-06432-7>.
- [160] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina, and Christos Faloutsos. RSC: Mining and Modeling Temporal Activity in Social Media. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 269–278, New York, NY, USA, August 2015. Association for Computing Machinery. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783294. URL <https://dl.acm.org/doi/10.1145/2783258.2783294>.
- [161] Sebastian Fiedler, Monika A. Piziorska, Viola Denninger, Alexey S. Morgunov, Alison Ilsley, Anisa Y. Malik, Matthias M. Schneider, Sean R. A. Devenish, Georg Meisl, Vasilis Kosmoliaptsis, Adriano Aguzzi, Heike Fiegler, and Tuomas P. J. Knowles. Antibody Affinity Governs the Inhibition of SARS-CoV-2 Spike/ACE2 Binding in Patient Serum. *ACS Infectious Diseases*, 7(8):2362–2369, August 2021. doi: 10.1021/acsinfecdis.1c00047. URL <https://doi.org/10.1021/acsinfecdis.1c00047>.
- [162] Sam Fogel. I Got Investigated by the Secret Service. Here’s How to Not Be Me. *Wired*, March 2023. ISSN 1059-1028. URL <https://www.wired.com/story/secret-service-investigated-me-twitter/>.
- [163] Kathy Frankovic. How Trump and Biden supporters differ on face masks — YouGov, October 03, 2020, 9:00 AM GMT-4. URL <https://today.yougov.com/topics/politics/articles-reports/2020/10/03/how-trump-and-biden-supporters-differ-face-masks>.
- [164] Seth J. Frantzman. Israel’s Iron Dome Won’t Last Forever. *Foreign Policy*, June 2021. URL <https://foreignpolicy.com/2021/06/03/israels-iron-dome-wont-last-forever/>.
- [165] Deen Freelon and Tetyana Lokot. Russian Twitter disinformation campaigns reach across the American political spectrum. *Harvard Kennedy School Misinformation Review*, 1(1), January 2020. doi: 10.37016/mr-2020-003. URL <https://misinformationreview.hks.harvard.edu/article/russian-disinformation-campaigns-on-twitter/>.

- [166] Chas W. Freeman. A Diplomatic Taxonomy for the New World Disorder. In Paul Webster Hare, Juan Luis Manfredi-Sánchez, and Kenneth Weisbrode, editors, *The Palgrave Handbook of Diplomatic Reform and Innovation*, Studies in Diplomacy and International Relations, pages 41–58. Springer International Publishing, Cham, 2023. ISBN 978-3-031-10971-3. doi: 10.1007/978-3-031-10971-3_3. URL https://doi.org/10.1007/978-3-031-10971-3_3.
- [167] John R. P. French Jr. A formal theory of social power. *Psychological Review*, 63:181–194, 1956. ISSN 1939-1471(Electronic),0033-295X(Print). doi: 10.1037/h0046123.
- [168] Feng Fu and Long Wang. Coevolutionary dynamics of opinions and networks: From diversity to uniformity. *Physical Review E*, 78(1):016104, July 2008. doi: 10.1103/PhysRevE.78.016104. URL <https://link.aps.org/doi/10.1103/PhysRevE.78.016104>.
- [169] Christian Fuchs. *Propaganda 2.0: Herman and Chomsky’s Propaganda Model in the Age of the Internet, Big Data and Social Media in the Age of the Internet, Big Data and Social Media*, pages 71–92. University of Westminster Press, October 2018. ISBN 978-1-912656-16-5. doi: 10.16997/book27.f. URL <https://www.uwestminsterpress.co.uk/site/chapters/10.16997/book27.f/>.
- [170] Vijaya Gadde and Keyvon Beykpour. Setting the record straight on shadow banning, July 2018. URL https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html.
- [171] Monica Gandhi and George W. Rutherford. Facial Masking for Covid-19 — Potential for “Variolation” as We Await a Vaccine. *New England Journal of Medicine*, 383(18):e101, October 2020. ISSN 0028-4793. doi: 10.1056/NEJMp2026913. URL <https://doi.org/10.1056/NEJMp2026913>.
- [172] Raphael Tsavkko Garcia. Harmful tweets from high places: Why Is Twitter acting now? *Salon*, April 2020. URL https://www.salon.com/2020/04/25/harmful-tweets-from-high-places-why-is-twitter-acting-now_partner/.
- [173] Andrew Garfield. The U.S. Counter-propaganda Failure in Iraq. *Middle East Quarterly*, 14(4):23–32, September 2007. URL <https://www.meforum.org/1753/the-us-counter-propaganda-failure-in-iraq>.
- [174] Bob Garfield. Revolution, Inc. — On the Media, November 2005. URL <https://www.wnycstudios.org/podcasts/otm/segments/129044-revolution-inc>.

- [175] Ruth Garland. The Surrender to the 24/7 News Cycle. In Ruth Garland, editor, *Government Communications and the Crisis of Trust: From Political Spin to Post-truth*, pages 107–121. Springer International Publishing, Cham, 2021. ISBN 978-3-030-77576-6. doi: 10.1007/978-3-030-77576-6_7. URL https://doi.org/10.1007/978-3-030-77576-6_7.
- [176] Maíra Gatti, Paulo Cavalin, Samuel Barbosa Neto, Claudio Pinhanez, Cícero dos Santos, Daniel Gribel, and Ana Paula Appel. Large-Scale Multi-agent-Based Modeling and Simulation of Microblogging-Based Online Social Network. In Shah Jamal Alam and H. Van Dyke Parunak, editors, *Multi-Agent-Based Simulation XIV*, Lecture Notes in Computer Science, pages 17–33, Berlin, Heidelberg, 2014. Springer. ISBN 978-3-642-54783-6. doi: 10.1007/978-3-642-54783-6_2.
- [177] Daniel Gayo-Avello. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" – A Balanced Survey on Election Prediction using Twitter Data. *arXiv:1204.6441 [physics]*, April 2012. URL <http://arxiv.org/abs/1204.6441>.
- [178] Dylan Gerstel. ISIS and Innovative Propaganda: Confronting Extremism in the Digital Age. *Swarthmore International Relations Journal*, 1(1):1–9, 2016. ISSN 25740113. doi: 10.24968/2574-0113.1.5. URL <http://works.swarthmore.edu/swarthmoreirjournal/vol1/iss1/5/>.
- [179] Souvick Ghosh and Chirag Shah. Towards automatic fake news classification. *Proceedings of the Association for Information Science and Technology*, 55(1):805–807, 2018. ISSN 2373-9231. doi: 10.1002/pra2.2018.14505501125. URL <https://asis.tdl.onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2018.14505501125>.
- [180] A. Giachanou, G. Zhang, and P. Rosso. Multimodal Multi-image Fake News Detection. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 647–654, October 2020. doi: 10.1109/DSAA49011.2020.00091.
- [181] Martin Gilens. Inequality and Democratic Responsiveness. *Public Opinion Quarterly*, 69(5):778–796, January 2005. ISSN 0033-362X. doi: 10.1093/poq/nfi058. URL <https://doi.org/10.1093/poq/nfi058>.
- [182] Martin Gilens and Benjamin I. Page. Testing Theories of American Politics: Elites, Interest Groups, and Average Citizens. *Perspectives on Politics*, 12(3):564–581, September 2014. ISSN 1537-5927, 1541-0986. doi: 10.1017/S1537592714001595. URL <https://www.cambridge.org/core/journals/perspectives-on-politics/>

- article/testing-theories-of-american-politics-elites-interest-groups-and-average-citizens/62327F513959D0A304D4893B382B992B.
- [183] Robert Gilpin. The Theory of Hegemonic War. *The Journal of Interdisciplinary History*, 18(4):591–613, 1988. ISSN 0022-1953. doi: 10.2307/204816. URL <https://www.jstor.org/stable/204816>.
- [184] Henry A. Giroux. Totalitarian Paranoia in the Post-Orwellian Surveillance State. *Cultural Studies*, 29(2):108–140, March 2015. ISSN 0950-2386, 1466-4348. doi: 10.1080/09502386.2014.917118. URL <http://www.tandfonline.com/doi/abs/10.1080/09502386.2014.917118>.
- [185] Vindu Goel. Facebook Tinkers With Users’ Emotions in News Feed Experiment, Stirring... *New York Times*, September 2014. URL http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html?_r=0.
- [186] Chaitanya Gokhale, Sanjib Das, AnHai Doan, Jeffrey F. Naughton, Narasimhan Rampalli, Jude Shavlik, and Xiaojin Zhu. Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’14, pages 601–612, Snowbird, Utah, USA, June 2014. Association for Computing Machinery. ISBN 978-1-4503-2376-5. doi: 10.1145/2588555.2588576. URL <https://doi.org/10.1145/2588555.2588576>.
- [187] Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. Fake News vs Satire: A Dataset and Analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci ’18, pages 17–21, Amsterdam, Netherlands, May 2018. Association for Computing Machinery. ISBN 978-1-4503-5563-6. doi: 10.1145/3201064.3201100. URL <https://doi.org/10.1145/3201064.3201100>.
- [188] Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12(3):

- 211–223, August 2001. ISSN 1573-059X. doi: 10.1023/A:1011122126881. URL <https://doi.org/10.1023/A:1011122126881>.
- [189] Josh A Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, and Katerina Sedova. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Technical report, Stanford Internet Observatory, OpenAI, and Georgetown University’s Center for Security and Emerging Technology, January 2023.
- [190] Yevgeniy Golovchenko. Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media. *The Journal of Politics*, 84(2):639–654, April 2022. ISSN 0022-3816, 1468-2508. doi: 10.1086/716949. URL <https://www.journals.uchicago.edu/doi/10.1086/716949>.
- [191] Mark Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, May 1978. ISSN 0002-9602. doi: 10.1086/226707. URL <https://www.journals.uchicago.edu/doi/abs/10.1086/226707>.
- [192] Graphika and SIO. Unheard Voice: Evaluating five years of pro-Western covert influence operations. Technical report, Graphika, Stanford Internet Observatory, August 2022.
- [193] Hannes Grassegger. Facebook says its ‘voter button’ is good for turnout. But should the tech giant be nudging us at all? *The Guardian*, April 2018. ISSN 0029-7712. URL <https://www.theguardian.com/technology/2018/apr/15/facebook-says-it-voter-button-is-good-for-turn-but-should-the-tech-giant-be-nudging-us-at-all>.
- [194] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting Diurnal Patterns of Real World Activity from Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 7(1):205–214, 2013. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v7i1.14409. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14409>.
- [195] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. Zighed. Information Diffusion in Online Social Networks: A Survey. *SIGMOD Rec.*, 42(2):17–28, July 2013. ISSN 0163-5808. doi: 10.1145/2503792.2503797. URL <http://doi.acm.org/10.1145/2503792.2503797>.

- [196] Sangwook Ha and JoongHo Ahn. Why Are You Sharing Others' Tweets?: The Impact of Argument Quality and Source Credibility on Information Sharing Behavior. *ICIS 2011 Proceedings*, December 2011. URL <https://aisel.aisnet.org/icis2011/proceedings/humanbehavior/4>.
- [197] Haaretz. Israeli Hackers Reportedly Offered Trump-linked Firm Cambridge Analytica Private Info on Two Heads of State. *Haaretz*, March 2018. URL <https://www.haaretz.com/israel-news/2018-03-21/ty-article/israeli-hackers-offered-trump-linked-firm-info-on-2-heads-of-state/0000017f-dbc2-d3a5-af7f-fbee1a830000>.
- [198] Mark Hachman. The best AI art generators: Bring your wildest dreams to life. *PCWorld*, Apr 4, 2023 3:45 am PDT. URL <https://www.pcworld.com/article/1672975/the-best-ai-art-generators-for-you-midjourney-bing-and-more.html>.
- [199] Andrew H. Hales and Kipling D. Williams. Marginalized Individuals and Extremism: The Role of Ostracism in Openness to Extreme Groups. *Journal of Social Issues*, 74(1):75–92, 2018. ISSN 1540-4560. doi: 10.1111/josi.12257. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/josi.12257>.
- [200] Rob Hall, Charles Sutton, and Andrew McCallum. Unsupervised Deduplication Using Cross-field Dependencies. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 310–317, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401931. URL <http://doi.acm.org/10.1145/1401890.1401931>.
- [201] Kristin Haltinner, Dilshani Sarathchandra, and Thomas Ptak. How Believing That Climate Change Is a Conspiracy Affects Skeptics' Environmental Attitudes. *Environment: Science and Policy for Sustainable Development*, 63(3):25–33, May 2021. ISSN 0013-9157, 1939-9154. doi: 10.1080/00139157.2021.1898900. URL <https://www.tandfonline.com/doi/full/10.1080/00139157.2021.1898900>.
- [202] Charles R. Haruna, MengShu Hou, Moses J. Eghan, Michael Y. Kpiebaareh, and Lawrence Tandoh. An Effective and Cost-Based Framework for a Qualitative Hybrid Data Deduplication. In Sanjiv K. Bhatia, Shailesh Tiwari, Krishn K. Mishra, and Munesh C. Trivedi, editors, *Advances in Computer Communication and Computational Sciences*, volume 924, pages 511–520. Springer Singapore, Singapore, 2019. ISBN 9789811368608 9789811368615. doi: 10.1007/978-981-13-6861-5_44. URL http://link.springer.com/10.1007/978-981-13-6861-5_44.

- [203] Aumyo Hassan and Sarah J. Barber. The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6(1):38, May 2021. ISSN 2365-7464. doi: 10.1186/s41235-021-00301-5. URL <https://doi.org/10.1186/s41235-021-00301-5>.
- [204] Fuad Mire Hassan and Mark Lee. Multi-stage News-Stance Classification Based on Lexical and Neural Features. In Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor Quintián, and Emilio Corchado, editors, *13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020)*, Advances in Intelligent Systems and Computing, pages 218–228, Cham, 2021. Springer International Publishing. ISBN 978-3-030-57805-3. doi: 10.1007/978-3-030-57805-3_21.
- [205] Hossein Hassani and Emmanuel Sirmal Silva. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing*, 7(2):62, June 2023. ISSN 2504-2289. doi: 10.3390/bdcc7020062. URL <https://www.mdpi.com/2504-2289/7/2/62>.
- [206] Kai He. Undermining Adversaries: Unipolarity, Threat Perception, and Negative Balancing Strategies after the Cold War. *Security Studies*, 21(2):154–191, April 2012. ISSN 0963-6412. doi: 10.1080/09636412.2012.679201. URL <https://doi.org/10.1080/09636412.2012.679201>.
- [207] Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, December 2021. URL <http://arxiv.org/abs/2111.09543>.
- [208] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.
- [209] Edward S. Herman. The Propaganda Model: A retrospective. *Journalism Studies*, 1(1):101–112, January 2000. ISSN 1461-670X, 1469-9699. doi: 10.1080/14616700361195. URL <http://www.tandfonline.com/doi/abs/10.1080/146167000361195>.
- [210] Edward S. Herman and Noam Chomsky. *Manufacturing Consent: The Political Economy of the Mass Media*. Random House, 2010.

- [211] Kelly Hernandez. I'm a Doctor and Here's Why Masks are Better Than Vaccines. *Yahoo Life*, September 2020. URL <https://www.yahoo.com/lifestyle/im-doctor-heres-why-masks-145348019.html>.
- [212] Linda Herrera. Youth and Citizenship in the Digital Age: A View from Egypt. *Harvard Educational Review*, 82(3):333–352, September 2012. ISSN 0017-8055, 1943-5045. doi: 10.17763/haer.82.3.88267r117u710300. URL <https://meridian.allenpress.com/her/article/82/3/333/32085/Youth-and-Citizenship-in-the-Digital-Age-A-View>.
- [213] Linda Herrera. *Revolution in the Age of Social Media : The Egyptian Popular Insurrection and the Internet*. Verso, London, 2014. ISBN 978-1-78168-275-3 1-78168-275-5 978-1-78168-276-0 1-78168-276-3. URL [9781781682753.jpg](https://www.versobooks.com/9781781682753.jpg).
- [214] Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4): 599–653, January 2000. ISSN 0036-1445. doi: 10.1137/S0036144500371907. URL <https://epubs.siam.org/doi/abs/10.1137/s0036144500371907>.
- [215] Francis Heylighen and Cliff Joslyn. Cybernetics and Second-Order Cybernetics. In *Encyclopedia of Physical Science and Technology*, pages 155–169. Elsevier, 2003. ISBN 978-0-12-227410-7. doi: 10.1016/B0-12-227410-5/00161-7. URL <https://linkinghub.elsevier.com/retrieve/pii/B0122274105001617>.
- [216] Richard A. Holley and Thomas M. Liggett. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. *The Annals of Probability*, 3(4):643–663, August 1975. ISSN 0091-1798, 2168-894X. doi: 10.1214/aop/1176996306. URL <http://projecteuclid.org/euclid.aop/1176996306>.
- [217] Duncan Hollis. The Influence of War; the War for Influence. *Temple International & Comparative Law Journal*, 32(1):31–46, 2018.
- [218] Petter Holme. Rare and everywhere: Perspectives on scale-free networks. *Nature Communications*, 10(1):1016, March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09038-8. URL <https://www.nature.com/articles/s41467-019-09038-8>.
- [219] Glen E. Holt. Democracy or Republic: Answering a Frequent Reference Question. *Public Library Quarterly*, 35(3):243–253, July 2016. ISSN 0161-6846. doi: 10.1080/01616846.2016.1210454. URL <https://doi.org/10.1080/01616846.2016.1210454>.

- [220] Benjamin D. Horne and Sibel Adali. This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News. In *Eleventh International AAAI Conference on Web and Social Media*, pages 759–766, May 2017. URL <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772>.
- [221] Philip N. Howard, Aiden Duffy, Deen Freelon, M. M. Hussain, Will Mari, and Marwa Maziad. Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?, 2011. URL <https://papers.ssrn.com/abstract=2595096>.
- [222] Haibo Hu. Competing opinion diffusion on social networks. *Royal Society Open Science*, 4(11), November 2017. ISSN 2054-5703. doi: 10.1098/rsos.171160. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5717675/>.
- [223] Jiacheng Huang, Wei Hu, Zhifeng Bao, Qijin Chen, and Yuzhong Qu. Deep entity matching with adversarial active learning. *The VLDB Journal*, April 2022. ISSN 0949-877X. doi: 10.1007/s00778-022-00745-1. URL <https://doi.org/10.1007/s00778-022-00745-1>.
- [224] Ethan Hunt, Ritvik Janamsetty, Chanana Kinares, Chanel Koh, Alexis Sanchez, Felix Zhan, Murat Ozdemir, Shabnam Waseem, Osman Yolcu, Binay Dahal, Justin Zhan, Laxmi Gewali, and Paul Oh. Machine Learning Models for Paraphrase Identification and its Applications on Plagiarism Detection. In *2019 IEEE International Conference on Big Knowledge (ICBK)*, pages 97–104, November 2019. doi: 10.1109/ICBK.2019.00021.
- [225] Keisuke Ikeda, Yoshiyuki Okada, Fujio Toriumi, Takeshi Sakaki, Kazuhiro Kazama, Itsuki Noda, Kosuke Shinoda, Hirohiko Suwa, and Satoshi Kurihara. Multi-agent Information Diffusion Model for Twitter. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 21–26, Warsaw, Poland, August 2014. IEEE. ISBN 978-1-4799-4143-8. doi: 10.1109/WI-IAT.2014.12. URL <http://ieeexplore.ieee.org/document/6927521/>.
- [226] Benjamin F Jackson. Censorship and Freedom of Expression in the Age of Facebook. *New Mexico Law Review*, 44:121–167, 2014.
- [227] Peter J. Jacques and Claire Connolly Knox. Hurricanes and hegemony: A qualitative analysis of micro-level climate change denial discourses. *Environmental Politics*, 25(5):831–852, September 2016. ISSN 0964-4016. doi: 10.1080/09644016.2016.1189233. URL <https://doi.org/10.1080/09644016.2016.1189233>.

- [228] Nishant Jain, Abhay Kumar, Shekhar Singh, Chirag Singh, and Suraj Tripathi. Deceptive Reviews Detection Using Deep Learning Techniques. In Elisabeth Métais, Farid Meziane, Sunil Vadera, Vijayan Sugumaran, and Mohamad Saraee, editors, *Natural Language Processing and Information Systems*, Lecture Notes in Computer Science, pages 79–91, Salford, UK, June 2019. Springer International Publishing. ISBN 978-3-030-23281-8. doi: 10.1007/978-3-030-23281-8_7.
- [229] Saeed Jamalzadeh, Kash Barker, Andrés D. González, and Sridhar Radhakrishnan. Protecting infrastructure performance from disinformation attacks. *Scientific Reports*, 12(1):12707, July 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-16832-w. URL <https://www.nature.com/articles/s41598-022-16832-w>.
- [230] Amelia M. Jamison, David A. Broniatowski, Mark Dredze, Anu Sangraula, Michael C. Smith, and Sandra C. Quinn. Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 ‘infodemic’ on Twitter. *Harvard Kennedy School Misinformation Review*, 1(3), September 2020. doi: 10.37016/mr-2020-38. URL <https://misinforeview.hks.harvard.edu/article/not-just-conspiracy-theories-vaccine-opponents-and-proponents-add-to-the-covid-19-infodemic-on-twitter/>.
- [231] Birgit Jennen and Ania Nussbaum. Germany and France Oppose Trump’s Twitter Exile. *Bloomberg.com*, January 2021. URL <https://www.bloomberg.com/news/articles/2021-01-11/merkel-sees-closing-trump-s-social-media-accounts-problematic>.
- [232] Euisol Jeong and Jieun Lee. We take the red pill, we confront the DickTrix: Online feminist activism and the augmentation of gendered realities in South Korea. *Feminist Media Studies*, 18(4):705–717, July 2018. ISSN 1468-0777. doi: 10.1080/14680777.2018.1447354. URL <https://doi.org/10.1080/14680777.2018.1447354>.
- [233] Caio Libanio Melo Jeronimo, Leandro Balby Marinho, Claudio E. C. Campelo, Adriano Veloso, and Allan Sales da Costa Melo. Fake News Classification Based on Subjective Language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*, iiWAS2019, pages 15–24, New York, NY, USA, December 2019. Association for Computing Machinery. ISBN 978-1-4503-7179-7. doi: 10.1145/3366030.3366039. URL <https://doi.org/10.1145/3366030.3366039>.

- [234] Yichuan Jiang and J. C. Jiang. Diffusion in Social Networks: A Multiagent Perspective. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(2):198–213, February 2015. ISSN 2168-2232. doi: 10.1109/TSMC.2014.2339198.
- [235] Zi-Hang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. ConvBERT: Improving BERT with Span-based Dynamic Convolution. In *Advances in Neural Information Processing Systems*, volume 33, pages 12837–12848. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/96da2f590cd7246bbde0051047b0d6f7-Abstract.html>.
- [236] Neil F. Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro- and anti-vaccination views. *Nature*, 582(7811):230–233, June 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2281-1. URL <https://www.nature.com/articles/s41586-020-2281-1>.
- [237] Marc Owen Jones. Satire, social media and revolutionary cultural production in the Bahrain uprising: From utopian fiction to political satire. *Communication and the Public*, 2(2):136–153, June 2017. ISSN 2057-0473. doi: 10.1177/2057047317706372. URL <https://doi.org/10.1177/2057047317706372>.
- [238] Heejung Jwa, Dongsuk Oh, Kinam Park, Jang Mook Kang, and Heuseok Lim. exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences*, 9(19):4062, January 2019. doi: 10.3390/app9194062. URL <https://www.mdpi.com/2076-3417/9/19/4062>.
- [239] Daniel Kahneman. *Thinking, Fast and Slow*. Doubleday Canada, November 2011. ISBN 978-0-385-67652-6.
- [240] André Kaiser. House of Lords and Monarchy: British Majoritarian Democracy and the Current Reform Debate about its Pre-Democratic Institutions. In Peter Catterall, Wolfram Kaiser, and Ulrike Walton-Jordan, editors, *Reforming the Constitution*, pages 97–128. Routledge, 1 edition, May 2014. ISBN 978-1-315-03978-7. doi: 10.4324/9781315039787-4. URL <https://www.taylorfrancis.com/books/9781135305628/chapters/10.4324/9781315039787-4>.
- [241] Chaitanya Kaligotla, Enver Yücesan, and Stephen E. Chick. Diffusion of competing rumours on social media. *Journal of Simulation*, 0(0):1–21, August 2020. ISSN 1747-7778. doi: 10.1080/17477778.2020.1785345. URL <https://doi.org/10.1080/17477778.2020.1785345>.

- [242] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, January 2021. ISSN 1573-7721. doi: 10.1007/s11042-020-10183-2. URL <https://doi.org/10.1007/s11042-020-10183-2>.
- [243] Amina Kammoun, Rim Slama, Hedi Tabia, Tarek Ouni, and Mohmed Abid. Generative Adversarial Networks for Face Generation: A Survey. *ACM Computing Surveys*, 55(5):94:1–94:37, December 2022. ISSN 0360-0300. doi: 10.1145/3527850. URL <https://dl.acm.org/doi/10.1145/3527850>.
- [244] Cho Ryok Kang and Sae Kyung Joo. Internet cafe and karaoke could be a hotbed for explosive spreading of COVID-19 into community. *Public Health Affairs*, 4(1), December 2020. ISSN 2586-2901, 2586-5609. doi: 10.29339/pha.20.10. URL <http://www.pha.or.kr/journal/view.php?number=74>.
- [245] Jeffrey Kaplan. ‘Leaderless resistance’. *Terrorism and Political Violence*, 9(3):80–95, September 1997. ISSN 0954-6553, 1556-1836. doi: 10.1080/09546559708427417. URL <http://www.tandfonline.com/doi/abs/10.1080/09546559708427417>.
- [246] Ali Kashefi and Tapan Mukerji. ChatGPT for Programming Numerical Methods, April 2023. URL <http://arxiv.org/abs/2303.12093>.
- [247] M. Kejriwal and D. P. Miranker. An Unsupervised Algorithm for Learning Blocking Schemes. In *2013 IEEE 13th International Conference on Data Mining*, pages 340–349, December 2013. doi: 10.1109/ICDM.2013.60.
- [248] Suzanne Rowan Kelleher. Why Some Cities And Counties Are Banning Face Masks With Valves. *Forbes*, May 2020. URL <https://www.forbes.com/sites/suzanne-rowankelleher/2020/05/26/why-some-cities-and-counties-are-banning-face-masks-with-valves/>.
- [249] Simon Kemp. WhatsApp is the World’s Favorite Social Platform (And Other Facts), April 2021. URL <https://blog.hootsuite.com/simon-kemp-social-media/>.
- [250] W. O. Kermack and A. G. McKendrick. A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927. ISSN 0950-1207. URL <http://www.jstor.org/stable/94815>.

- [251] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. Contributions to the mathematical theory of epidemics. II. —The problem of endemicity. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 138(834):55–83, October 1932. doi: 10.1098/rspa.1932.0171. URL <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1932.0171>.
- [252] Daniel Keyes. Jeff Bezos says third parties are besting Amazon — here’s why that’s good for the firm. *Business Insider*, April 2019. URL <https://www.businessinsider.com/jeff-bezos-notes-third-party-sellers-are-besting-amazon-2019-4>.
- [253] Samer Y Khamaiseh, Izzat Alsmadi, and Abdullah Al-Alaj. Deceiving Machine Learning-Based Saturation Attack Detection Systems in SDN. In *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, pages 44–50, November 2020. doi: 10.1109/NFV-SDN50289.2020.9289908.
- [254] Zahra Khanjani, Gabrielle Watson, and Vandana P. Janeja. Audio deepfakes: A survey. *Frontiers in Big Data*, 5:1001063, January 2023. ISSN 2624-909X. doi: 10.3389/fdata.2022.1001063. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9869423/>.
- [255] Ben Kiernan. The First Genocide: Carthage, 146 BC. *Diogenes*, 51(3):27–39, August 2004. ISSN 0392-1921, 1467-7695. doi: 10.1177/0392192104043648. URL <http://journals.sagepub.com/doi/10.1177/0392192104043648>.
- [256] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy, and Tim C. Kietzmann. Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146, March 2020. ISSN 0007-6813. doi: 10.1016/j.bushor.2019.11.006. URL <https://www.sciencedirect.com/science/article/pii/S0007681319301600>.
- [257] Jaechun Kim. Democratic Peace and Covert War: A Case Study of the U.S. Covert War in Chile. *Journal of International and Area Studies*, 12(1):25–47, June 2005.
- [258] Stephanie Kirchgaessner, Manisha Ganguly, David Pegg, and Jason Burke. Revealed: The hacking and disinformation team meddling in elections — Technology — The Guardian. *The Guardian*, 15 Feb 2023 04.00 GMT. URL <https://www.theguardian.com/world/2023/feb/15/revealed-disinformation-team-jorge-claim-meddling-elections-tal-hanan>.

- [259] Nishadi Kirielle, Peter Christen, and Thilina Ranbaduge. Unsupervised Graph-Based Entity Resolution for Complex Entities. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–30, February 2023. ISSN 1556-4681, 1556-472X. doi: 10.1145/3533016. URL <https://dl.acm.org/doi/10.1145/3533016>.
- [260] Andrew Kohut, Carroll Doherty, Michael Dimock, and Scott Keeter. Further Decline in Credibility Ratings for Most News Organizations. Technical report, Pew Research Center, August 2012. URL <https://www.pewresearch.org/politics/2012/08/16/further-decline-in-credibility-ratings-for-most-news-organizations/>.
- [261] Abhiteja Konda, Abhinav Prakash, Gregory A. Moss, Michael Schmoldt, Gregory D. Grant, and Supratik Guha. Aerosol Filtration Efficiency of Common Fabrics Used in Respiratory Cloth Masks. *ACS Nano*, 14(5):6339–6347, May 2020. ISSN 1936-0851. doi: 10.1021/acsnano.0c03252. URL <https://doi.org/10.1021/acsnano.0c03252>.
- [262] Carlo Kopp. Considerations on Deception Techniques Used in Political and Product Marketing. *Proceedings of 7th Australian Information Warfare and Security Conference*, 2006. doi: 10.4225/75/57A80FDFAA0CC. URL <http://ro.ecu.edu.au/isw/13>.
- [263] Carlo Kopp, Kevin B. Korb, and Bruce I. Mills. Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to “fake news”. *PLOS ONE*, 13(11):e0207383, November 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0207383. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0207383>.
- [264] Jan Kostka, Yvonne Anne Oswald, and Roger Wattenhofer. Word of Mouth: Rumor Dissemination in Social Networks. In Alexander A. Shvartsman and Pascal Felber, editors, *Structural Information and Communication Complexity*, Lecture Notes in Computer Science, pages 185–196, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-69355-0. doi: 10.1007/978-3-540-69355-0_16.
- [265] Stephanie Kramer. More Americans say they are regularly wearing masks in stores and other businesses, August 2020. URL <https://www.pewresearch.org/short-reads/2020/08/27/more-americans-say-they-are-regularly-wearing-masks-in-stores-and-other-businesses/>.
- [266] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. Advanced social engineering attacks. *Journal of Information Security and Applications*, 22:

- 113–122, June 2015. ISSN 2214-2126. doi: 10.1016/j.jisa.2014.09.005. URL <https://www.sciencedirect.com/science/article/pii/S2214212614001343>.
- [267] Bilal Kuchay. Churchill’s policies to blame for 1943 Bengal famine: Study. *Al Jazeera*, April 2019. URL <https://www.aljazeera.com/news/2019/4/1/churchills-policies-to-blame-for-1943-bengal-famine-study>.
- [268] Dominika Kunertova. The war in Ukraine shows the game-changing effect of drones depends on the game. *Bulletin of the Atomic Scientists*, 79(2):95–102, March 2023. ISSN 0096-3402. doi: 10.1080/00963402.2023.2178180. URL <https://doi.org/10.1080/00963402.2023.2178180>.
- [269] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, Raleigh North Carolina USA, April 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772751. URL <https://dl.acm.org/doi/10.1145/1772690.1772751>.
- [270] Haewoon Kwak, Hyunwoo Chun, and Sue Moon. Fragile online relationship: A first look at unfollow dynamics in twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, pages 1091–1100, New York, NY, USA, May 2011. Association for Computing Machinery. ISBN 978-1-4503-0228-9. doi: 10.1145/1978942.1979104. URL <https://dl.acm.org/doi/10.1145/1978942.1979104>.
- [271] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised Language Model Pre-training for French. *arXiv:1912.05372 [cs]*, March 2020. URL <http://arxiv.org/abs/1912.05372>.
- [272] Min Jung Lee and Dan Merica. Hillary Clinton: US should ‘take out’ Assad’s air fields — CNN Politics. *CNN*, April 2017. URL <https://www.cnn.com/2017/04/06/politics/hillary-clinton-syria-assad/index.html>.
- [273] Piroska Lendvai and Uwe Reichel. Contradiction Detection for Rumorous Claims. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics (ExProM)*, pages 31–40, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

- [274] Manuel Leone, Stefano Huber, Akhil Arora, Alberto García-Durán, and Robert West. A critical re-evaluation of neural methods for entity alignment. *Proceedings of the VLDB Endowment*, 15(8):1712–1725, April 2022. ISSN 2150-8097. doi: 10.14778/3529337.3529355. URL <https://doi.org/10.14778/3529337.3529355>.
- [275] Irwin P. Levin, Sandra L. Schneider, and Gary J. Gaeth. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organizational Behavior and Human Decision Processes*, 76(2):149–188, November 1998. ISSN 07495978. doi: 10.1006/obhd.1998.2804. URL <https://linkinghub.elsevier.com/retrieve/pii/S0749597898928047>.
- [276] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *arXiv:1910.13461 [cs, stat]*, October 2019. URL <http://arxiv.org/abs/1910.13461>.
- [277] Jiacheng Li, Qianwen Ma, Chunyuan Yuan, Wei Zhou, Jizhong Han, and Songlin Hu. Fusion Convolutional Attention Network for Opinion Spam Detection. In Tom Gedeon, Kok Wai Wong, and Minhoo Lee, editors, *Neural Information Processing*, Lecture Notes in Computer Science, pages 223–235, Cham, 2019. Springer International Publishing. ISBN 978-3-030-36708-4. doi: 10.1007/978-3-030-36708-4_19.
- [278] Jiwei Li, Myle Ott, and Claire Cardie. Identifying Manipulated Offerings on Review Portals. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Process*, pages 1933–1942, Seattle, Washington, USA, 18.
- [279] Weihua Li, Quan Bai, Minjie Zhang, and Tung Doan Nguyen. Modelling Multiple Influences Diffusion in On-line Social Networks. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS ’18, pages 1053–1061, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems. URL <http://dl.acm.org/citation.cfm?id=3237383.3237854>.
- [280] Weihua Li, Quan Bai, and Minjie Zhang. A Multi-agent System for Modelling Preference-Based Complex Influence Diffusion in Social Networks. *The Computer Journal*, 62(3):430–447, March 2019. ISSN 0010-4620. doi: 10.1093/comjnl/bxy078. URL <http://academic.oup.com/comjnl/article/62/3/430/5067534>.
- [281] Yuchen Li, Dongxiang Zhang, and Kian-Lee Tan. Real-time Targeted Influence Maximization for Online Advertisements. *Proc. VLDB Endow.*, 8(10):1070–1081,

- June 2015. ISSN 2150-8097. doi: 10.14778/2794367.2794376. URL <https://doi.org/10.14778/2794367.2794376>.
- [282] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. Deep entity matching with pre-trained language models. *Proceedings of the VLDB Endowment*, 14(1):50–60, September 2020. ISSN 2150-8097. doi: 10.14778/3421424.3421431. URL <http://doi.org/10.14778/3421424.3421431>.
- [283] Yuliang Li, Jinfeng Li, Yoshi Suhara, AnHai Doan, and Wang-Chiew Tan. Effective entity matching with transformers. *The VLDB Journal*, January 2023. ISSN 1066-8888, 0949-877X. doi: 10.1007/s00778-023-00779-z. URL <https://link.springer.com/10.1007/s00778-023-00779-z>.
- [284] Mingming Liang, Liang Gao, Ce Cheng, Qin Zhou, John Patrick Uy, Kurt Heiner, and Chenyu Sun. Efficacy of face mask in preventing respiratory virus transmission: A systematic review and meta-analysis. *Travel Medicine and Infectious Disease*, 36: 101751, July 2020. ISSN 1477-8939. doi: 10.1016/j.tmaid.2020.101751. URL <http://www.sciencedirect.com/science/article/pii/S1477893920302301>.
- [285] Xialing Lin, Patric R. Spence, and Kenneth A. Lachlan. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, 63:264–271, October 2016. ISSN 07475632. doi: 10.1016/j.chb.2016.05.002. URL <https://linkinghub.elsevier.com/retrieve/pii/S074756321630320X>.
- [286] Mai Skjott Linneberg and Steffen Korsgaard. Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal*, 19(3):259–270, March 2019. doi: 10.1108/QRJ-12-2018-0012. URL <https://www-proquest-com.proxy3.library.mcgill.ca/docview/2234410819?accountid=12339>.
- [287] Travis Linnemann and Corina Medley. Black sites, “dark sides”: War power, police power, and the violence of the (un)known. *Crime, Media, Culture: An International Journal*, 15(2):341–358, August 2019. ISSN 1741-6590, 1741-6604. doi: 10.1177/1741659018777779. URL <http://journals.sagepub.com/doi/10.1177/1741659018777779>.
- [288] Chao Liu, Xinghua Wu, Min Yu, Gang Li, Jianguo Jiang, Weiqing Huang, and Xiang Lu. A Two-Stage Model Based on BERT for Short Fake News Detection. In Christos Douligieris, Dimitris Karagiannis, and Dimitris Apostolou, editors, *Knowledge Science, Engineering and Management*, Lecture Notes in Computer Science, pages 172–183,

- Cham, 2019. Springer International Publishing. ISBN 978-3-030-29563-9. doi: 10.1007/978-3-030-29563-9_17.
- [289] Jinliang Liu, Tingting Yin, Mouquan Shen, Xiangpeng Xie, and Jie Cao. State estimation for cyber–physical systems with limited communication resources, sensor saturation and denial-of-service attacks. *ISA Transactions*, 104:101–114, September 2020. ISSN 0019-0578. doi: 10.1016/j.isatra.2018.12.032. URL <https://www.sciencedirect.com/science/article/pii/S0019057818305275>.
- [290] Yang Liu and Mirella Lapata. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL <https://www.aclweb.org/anthology/D19-1387>.
- [291] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. URL <http://arxiv.org/abs/1907.11692>.
- [292] Yun Liu, Su-Meng Diao, Yi-Xiang Zhu, and Qing Liu. SHIR competitive information diffusion model for online social media. *Physica A: Statistical Mechanics and its Applications*, 461:543–553, November 2016. ISSN 0378-4371. doi: 10.1016/j.physa.2016.06.080. URL <http://www.sciencedirect.com/science/article/pii/S037843711630348X>.
- [293] Zhan Liu, Shaban Shabani, Nicole Glassey Balet, and Maria Sokhn. Detection of Satiric News on Social Media: Analysis of the Phenomenon with a French Dataset. In *2019 28th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6, July 2019. doi: 10.1109/ICCCN.2019.8847041.
- [294] Elena Lloret, María Teresa Romá-Ferri, and Manuel Palomar. COMPENDIUM: A text summarization system for generating abstracts of research papers. *Data & Knowledge Engineering*, 88:164–175, November 2013. ISSN 0169-023X. doi: 10.1016/j.datak.2013.08.005. URL <https://www.sciencedirect.com/science/article/pii/S0169023X13000815>.
- [295] David Lloyd. Settler Colonialism and the State of Exception: The Example of Palestine/Israel. *Settler Colonial Studies*, 2(1):59–80, January 2012. ISSN 2201-

- 473X. doi: 10.1080/2201473X.2012.10648826. URL <https://doi.org/10.1080/2201473X.2012.10648826>.
- [296] Paul A Longley, Muhammad Adnan, and Guy Lansley. The Geotemporal Demographics of Twitter Usage. *Environment and Planning A: Economy and Space*, 47(2): 465–484, February 2015. ISSN 0308-518X, 1472-3409. doi: 10.1068/a130122p. URL <http://journals.sagepub.com/doi/10.1068/a130122p>.
- [297] Jamie Lopez Bernal, Nick Andrews, Charlotte Gower, Eileen Gallagher, Ruth Simons, Simon Thelwall, Julia Stowe, Elise Tessier, Natalie Groves, Gavin Dabrera, Richard Myers, Colin N.J. Campbell, Gayatri Amirthalingam, Matt Edmunds, Maria Zambon, Kevin E. Brown, Susan Hopkins, Meera Chand, and Mary Ramsay. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. *New England Journal of Medicine*, 385(7):585–594, August 2021. ISSN 0028-4793. doi: 10.1056/NEJMoa2108891. URL <https://doi.org/10.1056/NEJMoa2108891>.
- [298] Jwen Fai Low, Benjamin C. M. Fung, Farkhund Iqbal, and Shih-Chia Huang. Distinguishing between fake news and satire with transformers. *Expert Systems with Applications*, 187:115824, January 2022. ISSN 0957-4174. doi: 10.1016/j.eswa.2021.115824. URL <https://www.sciencedirect.com/science/article/pii/S0957417421011891>.
- [299] Jwen Fai Low, Benjamin C. M. Fung, and Farkhund Iqbal. Of Stances, Themes, and Anomalies in COVID-19 Mask-Wearing Tweets. *IEEE Access*, 11:87009–87033, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3298557.
- [300] Jwen Fai Low, Benjamin C.M. Fung, Farkhund Iqbal, and Claude Fachkha. Carthago Delenda Est: Co-opetitive Indirect Information Diffusion Model for Influence Operations on Online Social Media. *IEEE Transactions on Cybernetics*, 2023.
- [301] Jwen Fai Low, Benjamin CM Fung, Farkhund Iqbal, and Pulei Xiong. Better Entity Matching with Transformers through Ensembles. *Electronic Commerce Research and Applications*, 2023.
- [302] Zhicong Lu, Yue Jiang, Chenxinran Shen, Margaret C. Jack, Daniel Wigdor, and Mor Naaman. "Positive Energy": Perceptions and Attitudes Towards COVID-19 Information on Social Media in China. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):177:1–177:25, April 2021. doi: 10.1145/3449251. URL <https://doi.org/10.1145/3449251>.

- [303] Bill Lucas and Janet Hanson. Thinking Like an Engineer: Using Engineering Habits of Mind and Signature Pedagogies to Redesign Engineering Education. *International Journal of Engineering Pedagogy (IJEP)*, 6(2):4–13, May 2016. URL <https://doi.org/10.3991/ijep.v6i2.5366>.
- [304] Eric A. M. Luijck. Information assurance and the information Society. In Urs E. Gattiker, Pia Pedersen, and Karsten Petersen, editors, *EICAR Best Paper Proceedings*, 1999.
- [305] Dorian Lyons, Ceyda Kiyak, Deniz Cetinkaya, Sarah Hodge, and John McAlaney. Design and Development of a Mobile Application to Combat Digital Addiction and Dissociative States During Phone Usage. In *2022 IEEE International Conference on E-Business Engineering (ICEBE)*, pages 204–209, Bournemouth, United Kingdom, October 2022. IEEE. ISBN 978-1-66549-244-7. doi: 10.1109/ICEBE55470.2022.00043. URL <https://ieeexplore.ieee.org/document/10035050/>.
- [306] Kim Lyons. Twitter censored tweets critical of India’s handling of the pandemic at its government’s request, April 2021. URL <https://www.theverge.com/2021/4/24/22400976/twitter-removed-tweets-critical-india-censor-coronavirus>.
- [307] Wei Lyu and George L. Wehby. Community Use Of Face Masks And COVID-19: Evidence From A Natural Experiment Of State Mandates In The US. *Health Affairs*, 39(8):1419–1425, June 2020. ISSN 0278-2715. doi: 10.1377/hlthaff.2020.00818. URL <https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.2020.00818>.
- [308] C. Raina MacIntyre and Abrar Ahmad Chughtai. Facemasks for the prevention of infection in healthcare and community settings. *BMJ*, 350, April 2015. ISSN 1756-1833. doi: 10.1136/bmj.h694. URL <https://www.bmj.com/content/350/bmj.h694>.
- [309] C. Raina MacIntyre and Abrar Ahmad Chughtai. A rapid systematic review of the efficacy of face masks and respirators against coronaviruses and other respiratory transmissible viruses for the community, healthcare workers and sick patients. *International Journal of Nursing Studies*, 108:103629, August 2020. ISSN 0020-7489. doi: 10.1016/j.ijnurstu.2020.103629. URL <http://www.sciencedirect.com/science/article/pii/S0020748920301139>.
- [310] Alan MacLeod. Manufacturing Consent in Venezuela: Media Misreporting of a Country, 1998–2014 - Alan MacLeod, 2020. *Critical Sociology*, December 2018. URL <https://journals-sagepub-com.proxy3.library.mcgill.ca/doi/10.1177/0896920518820934>.

- [311] Elisabeth Mahase. Covid-19: What is the evidence for cloth masks? *BMJ*, 369, April 2020. ISSN 1756-1833. doi: 10.1136/bmj.m1422. URL <https://www.bmj.com/content/369/bmj.m1422>.
- [312] Elisabeth Mahase. Covid-19: Novavax vaccine efficacy is 86% against UK variant and 60% against South African variant. *BMJ*, 372:n296, February 2021. ISSN 1756-1833. doi: 10.1136/bmj.n296. URL <https://www.bmj.com/content/372/bmj.n296>.
- [313] Roshni Majumdar. EU top diplomat denies 'jungle' remark was racist. *Deutsche Welle*, October 2022. URL <https://www.dw.com/en/eu-top-diplomat-denies-jungle-remark-was-racist/a-63484246>.
- [314] Daniel P. Maki and Maynard Thompson. *Mathematical Models and Applications : With Emphasis on the Social, Life, and Management Sciences*. Prentice-Hall, Englewood Cliffs, N.J., 1973. ISBN 0-13-561670-0 978-0-13-561670-3. URL <http://www.gbv.de/dms/hbz/toc/ht004611076.PDF>.
- [315] Larissa G. Malagoli, Julia Stancioli, Carlos H. G. Ferreira, Marisa Vasconcelos, Ana Paula Couto da Silva, and Jussara M. Almeida. A Look into COVID-19 Vaccination Debate on Twitter. In *WebSci '21: WebSci '21 13th ACM Web Science Conference 2021*, pages 225–233, Virtual Event, United Kingdom, June 2021. Association for Computing Machinery. ISBN 978-1-4503-8330-1. URL <https://doi.org/10.1145/3447535.3462498>.
- [316] Philip Marcelo. FACT FOCUS: Fake image of Pentagon explosion briefly sends jitters through stock market. *Associated Press*, May 2023. URL <https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4>.
- [317] Vukosi Marivate, Avashlin Moodley, and Athandiwe Saba. Extracting and categorising the reactions to COVID-19 by the South African public – A social media study. *arXiv:2006.06336 [cs]*, June 2020. URL <http://arxiv.org/abs/2006.06336>.
- [318] VenkataSwamy Martha, Weizhong Zhao, and Xiaowei Xu. A study on Twitter user-follower network: A network based analysis. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '13*, pages 1405–1409, New York, NY, USA, August 2013. Association for Computing Machinery. ISBN 978-1-4503-2240-9. doi: 10.1145/2492517.2500298. URL <https://dl.acm.org/doi/10.1145/2492517.2500298>.

- [319] Alice E. Marwick and Danah Boyd. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1): 114–133, February 2011. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444810365313. URL <http://journals.sagepub.com/doi/10.1177/1461444810365313>.
- [320] Michael Mastanduno. System Maker and Privilege Taker: U.S. Power and the International Political Economy. *World Politics*, 61(1):121–154, January 2009. ISSN 0043-8871, 1086-3338. doi: 10.1017/S0043887109000057. URL <https://muse.jhu.edu/article/260515>.
- [321] Michele Mazza, Guglielmo Cola, and Maurizio Tesconi. Ready-to-(ab)use: From fake account trafficking to coordinated inauthentic behavior on Twitter. *Online Social Networks and Media*, 31:100224, September 2022. ISSN 24686964. doi: 10.1016/j.osnem.2022.100224. URL <https://linkinghub.elsevier.com/retrieve/pii/S2468696422000271>.
- [322] Mark Mazzetti, Ronen Bergman, David D. Kirkpatrick, and Maggie Haberman. Rick Gates Sought Online Manipulation Plans From Israeli Intelligence Firm for Trump Campaign. *The New York Times*, October 2018. ISSN 0362-4331. URL <https://www.nytimes.com/2018/10/08/us/politics/rick-gates-psy-group-trump.html>.
- [323] J. Lee McCullough and Thomas M. Ostrom. Repetition of highly similar messages and attitude change. *Journal of Applied Psychology*, 59(3):395–397, 1974. URL <https://doi.org/10.1037/h0036658>.
- [324] Jenna McLaughlin. Why the US government’s counter-terrorism tweeters are finding it tough to fight ISIS Online, September 2014. URL <https://www.motherjones.com/politics/2014/09/isis-social-media-state-department/>.
- [325] Karen McVeigh. Drone strikes: Tears in Congress as Pakistani family tells of mother’s death. *The Guardian*, October 2013. ISSN 0261-3077. URL <https://www.theguardian.com/world/2013/oct/29/pakistan-family-drone-victim-testimony-congress>.
- [326] Richard J Medford, Sameh N Saleh, Andrew Sumarsono, Trish M Perl, and Christoph U Lehmann. An “Infodemic”: Leveraging High-Volume Twitter Data to Understand Early Public Sentiment for the Coronavirus Disease 2019 Outbreak. *Open Forum Infectious Diseases*, 7(ofaa258), July 2020. ISSN 2328-8957. doi: 10.1093/ofid/ofaa258. URL <https://doi.org/10.1093/ofid/ofaa258>.

- [327] Richard Medina and George Hepner. Geospatial Analysis of Dynamic Terrorist Networks. In Ibrahim A. Karawan, Wayne McCormack, and Stephen E. Reynolds, editors, *Values and Violence*, volume 4, pages 151–167. Springer Netherlands, Dordrecht, 2009. ISBN 978-1-4020-8659-5 978-1-4020-8660-1. doi: 10.1007/978-1-4020-8660-1_10. URL http://link.springer.com/10.1007/978-1-4020-8660-1_10.
- [328] Gur Megiddo and Omer Benjakob. The Israelis Destabilizing Democracy and Disrupting Elections Worldwide. *Haaretz*, November 2022. URL <https://www.haaretz.com/israel-news/security-aviation/2022-11-16/ty-article-static-ext/the-israelis-destabilizing-democracy-and-disrupting-elections-worldwide/0000186-461e-d80f-abff-6e9e08b10000>.
- [329] Gur Megiddo and Omer Benjakob. Antisemitic Influencers, Fake Journalists: Inside an Israeli Factory for Online Deception. *Haaretz*, February 2023. URL <https://www.haaretz.com/israel-news/security-aviation/2023-02-16/ty-article-magazine/.premium/antisemitic-influencers-fake-journalists-inside-an-israeli-factory-for-online-deception/00000186-4b7c-da6c-a3f6-cb7f03cf0000>.
- [330] Gur Megiddo and Omer Benjakob. Cambridge Analytica’s Israeli Black Ops Team – Exposed at Last. *Haaretz*, February 2023. URL <https://www.haaretz.com/israel-news/security-aviation/2023-02-15/ty-article-magazine/.premium/cambridge-analyticas-israeli-black-ops-team-exposed-at-last/00000186-4b78-da04-a186-7bfa41fe0000>.
- [331] Ulises A Mejias and Nikolai E Vokuev. Disinformation and the media: The case of Russia and Ukraine. *Media, Culture & Society*, 39(7):1027–1042, October 2017. ISSN 0163-4437. doi: 10.1177/0163443716686672. URL <https://doi.org/10.1177/0163443716686672>.
- [332] Alimuddin Melleng, Anna-Jurek Loughrey, and Deepak P. Sentiment and Emotion Based Text Representation for Fake Reviews Detection. In *Proceedings - Natural Language Processing in a Deep Learning World*, pages 750–757. Incoma Ltd., Shoumen, Bulgaria, October 2019. ISBN 978-954-452-056-4. doi: 10.26615/978-954-452-056-4_087. URL <https://acl-bg.org/proceedings/2019/RANLP2019/pdf/RANLP087.pdf>.
- [333] Erwan Le Merrer, Benoit Morgan, and Gilles Trédan. Setting the Record Straighter on Shadow Banning. *arXiv:2012.05101 [cs]*, April 2021. URL <http://arxiv.org/abs/2012.05101>.

- [334] Merriam-Webster. Words We're Watching: 'Ratioed', 2017. URL <https://www.merriam-webster.com/words-at-play/words-were-watching-ratio-ratioed-ratioing>.
- [335] Panagiotis Metaxas, Eni Mustafaraj, Kily Wong, Laura Zeng, Megan O'Keefe, and Samantha Finn. What Do Retweets Indicate? Results from User Survey and Meta-Review of Research. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 9(1):658–661, 2015. ISSN 2334-0770, 2162-3449. doi: 10.1609/icwsm.v9i1.14661. URL <https://ojs.aaai.org/index.php/ICWSM/article/view/14661>.
- [336] Miriam J. Metzger and Andrew J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59:210–220, December 2013. ISSN 0378-2166. doi: 10.1016/j.pragma.2013.07.012. URL <https://www.sciencedirect.com/science/article/pii/S0378216613001768>.
- [337] Samantha B. Meyer, Stephanie K. Lu, Laurie Hoffman-Goetz, Bryan Smale, Heather MacDougall, and Alex R. Pearce. A Content Analysis of Newspaper Coverage of the Seasonal Flu Vaccine in Ontario, Canada, October 2001 to March 2011. *Journal of Health Communication*, 21(10):1088–1097, October 2016. ISSN 1081-0730. doi: 10.1080/10810730.2016.1222038. URL <https://doi.org/10.1080/10810730.2016.1222038>.
- [338] Christopher Meyers. Partisan News, the Myth of Objectivity, and the Standards of Responsible Journalism. *Journal of Media Ethics*, 35(3):180–194, July 2020. ISSN 2373-6992, 2373-700X. doi: 10.1080/23736992.2020.1780131. URL <https://www.tandfonline.com/doi/full/10.1080/23736992.2020.1780131>.
- [339] Daniel Michaels. Ukraine's Internet Army of 'NAFO Fellas' Fights Russian Trolls and Re.... *Wall Street Journal*, September 2022. URL <https://www.wsj.com/amp/articles/ukraines-internet-army-of-nafo-fellas-fights-russian-trolls-and-rewards-donors-with-dogs-11664271002>.
- [340] Robert J. Miller. Nazi Germany's Race Laws, the United States, and American Indians. *St. John's Law Review*, 94(3):751–818, 2020. URL <https://heinonline.org/HOL/P?h=hein.journals/stjohn94&i=769>.
- [341] Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Christopher M. Danforth, and Peter Sheridan Dodds. Ratioing the President: An exploration of public engagement

- with Obama and Trump on Twitter. *arXiv:2006.03526 [physics]*, June 2020. URL <http://arxiv.org/abs/2006.03526>.
- [342] Joshua R. Minot, Michael V. Arnold, Thayer Alshaabi, Christopher M. Danforth, and Peter Sheridan Dodds. Ratioing the President: An exploration of public engagement with Obama and Trump on Twitter. *PLOS ONE*, 16(4):e0248880, April 2021. ISSN 1932-6203. doi: 10.1371/journal.pone.0248880. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0248880>.
- [343] Vimal Mishra, Amar Deep Tiwari, Saran Aadhar, Reepal Shah, Mu Xiao, D. S. Pai, and Dennis Lettenmaier. Drought and Famine in India, 1870–2016. *Geophysical Research Letters*, 46(4):2075–2083, 2019. ISSN 1944-8007. doi: 10.1029/2018GL081477. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2018GL081477>.
- [344] Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology*, 17(3):1–23, July 2017. ISSN 1533-5399, 1557-6051. doi: 10.1145/3003433. URL <https://dl.acm.org/doi/10.1145/3003433>.
- [345] Samantha L. Moore-Berg, Jacob M. Parelman, Yphtach Lelkes, and Emily B. Falk. Neural polarization and routes to depolarization. *Proceedings of the National Academy of Sciences*, 117(46):28552–28554, November 2020. doi: 10.1073/pnas.2020107117. URL <https://www.pnas.org/doi/full/10.1073/pnas.2020107117>.
- [346] David Morin. Information Influence Operations: The Future of Information Dominance. *The Cyber Defense Review*, 6(1):133–140, 2023. URL <https://www.jstor.org/stable/26994117>.
- [347] Sidharth Mudgal, Han Li, Theodoros Rekatsinas, and AnHai Doan. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data*, page 16, Houston, TX, USA, June 2018. doi: 10.1145/3183713.3196926.
- [348] Michael Muhlmeyer, Shaurya Agarwal, and Archie J. Huang. Modeling Social Contagion and Information Diffusion in Complex Socio-Technical Systems. *IEEE Systems Journal*, 14(4):5187–5198, December 2020. ISSN 1937-9234. doi: 10.1109/JSYST.2020.2993542.
- [349] Tawanda Mukurunge and Senate Rapitse. Comic Relief, Telling Unwanted Truth to Power and Ethical Journalism: Satire Writing in Print Media in Lesotho. *Journal*

- of Psychology Research*, 9(9), September 2019. ISSN 21595542, 21595550. doi: 10.17265/2159-5542/2019.09.004. URL <http://www.davidpublisher.org/index.php/Home/Article/index?id=41968.html>.
- [350] Andrew Mullen and Jeffery Klaehn. The Herman–Chomsky Propaganda Model: A Critical Approach to Analysing Mass Media Behaviour. *Sociology Compass*, 4(4):215–229, 2010. ISSN 1751-9020. doi: 10.1111/j.1751-9020.2010.00275.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-9020.2010.00275.x>.
- [351] mumbling. Ratioed, December 2017. URL <https://www.urbandictionary.com/define.php?term=Ratioed>.
- [352] Hanna Paalgard Munden. Extremist group exits: What autobiographies by male right-wing formers reveal about identity transformation. *Behavioral Sciences of Terrorism and Political Aggression*, 0(0):1–22, April 2023. ISSN 1943-4472. doi: 10.1080/19434472.2023.2192771. URL <https://doi.org/10.1080/19434472.2023.2192771>.
- [353] S. A. Myers and J. Leskovec. Clash of the Contagions: Cooperation and Competition in Information Diffusion. In *2012 IEEE 12th International Conference on Data Mining*, pages 539–548, December 2012. doi: 10.1109/ICDM.2012.159.
- [354] Seth A. Myers and Jure Leskovec. The bursty dynamics of the Twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW ’14, pages 913–924, New York, NY, USA, April 2014. Association for Computing Machinery. ISBN 978-1-4503-2744-2. doi: 10.1145/2566486.2568043. URL <https://doi.org/10.1145/2566486.2568043>.
- [355] Richard Nadeau, Edouard Cloutier, and J.-H. Guay. New Evidence About the Existence of a Bandwagon Effect in the Opinion Formation Process. *International Political Science Review*, 14(2):203–213, April 1993. ISSN 0192-5121, 1460-373X. doi: 10.1177/019251219301400204. URL <http://journals.sagepub.com/doi/10.1177/019251219301400204>.
- [356] Alexander Nazaryan. CDC chief says masks better at stopping coronavirus than a vaccine. *Yahoo News*, September 2020. URL <https://news.yahoo.com/cdc-chief-says-masks-better-at-stopping-coronavirus-than-a-vaccine-173526486.html>.

- [357] M. Lynne Neufeld and Martha Cornog. Database history: From dinosaurs to compact discs. *Journal of the American Society for Information Science*, 37(4):183–190, July 1986. ISSN 00028231, 10974571. doi: 10.1002/(SICI)1097-4571(198607)37:4<183::AID-ASI2>3.0.CO;2-W. URL [https://onlinelibrary.wiley.com/doi/10.1002/\(SICI\)1097-4571\(198607\)37:4<183::AID-ASI2>3.0.CO;2-W](https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(198607)37:4<183::AID-ASI2>3.0.CO;2-W).
- [358] M. E. J. Newman. The Structure and Function of Complex Networks. *SIAM Review*, 45(2):167–256, January 2003. ISSN 0036-1445. doi: 10.1137/S003614450342480. URL <https://epubs.siam.org/doi/abs/10.1137/s003614450342480>.
- [359] First Draft News. About, March 2020. URL <https://firstdraftnews.org:443/about/>.
- [360] H. T. Nguyen, T. N. Dinh, and M. T. Thai $\pm$. Cost-aware Targeted Viral Marketing in billion-scale networks. In *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, pages 1–9, April 2016. doi: 10.1109/INFOCOM.2016.7524377.
- [361] Hao Nie, Xianpei Han, Ben He, Le Sun, Bo Chen, Wei Zhang, Suhui Wu, and Hao Kong. Deep Sequence-to-Sequence Entity Matching for Heterogeneous Entity Resolution. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 629–638, Beijing China, November 2019. ACM. ISBN 978-1-4503-6976-3. doi: 10.1145/3357384.3358018. URL <https://dl.acm.org/doi/10.1145/3357384.3358018>.
- [362] Beryl Noë, Liam D. Turner, David E. J. Linden, Stuart M. Allen, Bjorn Winkens, and Roger M. Whitaker. Identifying Indicators of Smartphone Addiction Through User-App Interaction. *Computers in Human Behavior*, 99:56–65, October 2019. ISSN 0747-5632. doi: 10.1016/j.chb.2019.04.023. URL <https://www.sciencedirect.com/science/article/pii/S0747563219301712>.
- [363] Elisabeth Noelle-Neumann. The Spiral of Silence A Theory of Public Opinion. *Journal of Communication*, 24(2):43–51, 1974. ISSN 1460-2466. doi: 10.1111/j.1460-2466.1974.tb00367.x. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.1974.tb00367.x>.
- [364] Jeppe Nørregaard, Benjamin D. Horne, and Sibel Adalı. NELA-GT-2018: A Large Multi-Labelled News Dataset for the Study of Misinformation in News Articles.

- Proceedings of the International AAI Conference on Web and Social Media*, 13:630–638, July 2019. ISSN 2334-0770. URL <https://aaai.org/ojs/index.php/ICWSM/article/view/3261>.
- [365] Abby Ohlheiser. How K-pop fans became celebrated online vigilantes, June 2020. URL <https://www.technologyreview.com/2020/06/05/1002781/kpop-fans-and-black-lives-matter/>.
- [366] Rodrigo Olivares, Francisco Muñoz, and Fabián Riquelme. A multi-objective linear threshold influence spread model solved by swarm intelligence-based methods. *Knowledge-Based Systems*, 212:106623, January 2021. ISSN 0950-7051. doi: 10.1016/j.knosys.2020.106623. URL <https://www.sciencedirect.com/science/article/pii/S0950705120307528>.
- [367] Daniela Oliveira, Marissa Rosenthal, Nicole Morin, Kuo-Chuan Yeh, Justin Cappos, and Yanyan Zhuang. It’s the psychology stupid: How heuristics explain software vulnerabilities and how priming can illuminate developer’s blind spots. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 296–305, New Orleans Louisiana USA, December 2014. ACM. ISBN 978-1-4503-3005-3. doi: 10.1145/2664243.2664254. URL <https://dl.acm.org/doi/10.1145/2664243.2664254>.
- [368] Stefan Ollinger, Lorik Dumani, Premtim Sahitaj, Ralph Bergmann, and Ralf Schenkel. Same Side Stance Classification Task: Facilitating Argument Stance Classification by Fine-tuning a BERT Model. *arXiv:2004.11163 [cs]*, April 2020. URL <http://arxiv.org/abs/2004.11163>.
- [369] Amarta O’Neil. Right wing terrorism in the west: Radicalization and decentralization. *Pathways to Peace and Security*, 1(1):125–137, 2017. ISSN 23071494, 23115238. doi: 10.20542/2307-1494-2017-1-125-137. URL http://www.imemo.ru/files/File/magazines/puty_miru/2017/01/080Neil.pdf.
- [370] Agnieszka Onuchowska and Donald J Berndt. Using Agent-Based Modelling to Address Malicious Behavior on Social Media. In *ICIS 2019 Proceedings*, Munich, 2019.
- [371] World Health Organization. Advice on the use of masks in the context of COVID-19: Interim guidance, 6 April 2020. Technical Report WHO/2019-nCov/IPC-Masks/2020.3, World Health Organization, 2020. URL <https://apps.who.int/iris/handle/10665/331693>.

- [372] World Health Organization. Advice on the use of masks in the context of COVID-19: Interim guidance, 5 June 2020. Technical Report WHO/2019-nCov/IPC-Masks/2020.4, World Health Organization, 2020. URL <https://apps.who.int/iris/handle/10665/332293>.
- [373] Matteo Paganelli, Francesco Del Buono, Andrea Baraldi, and Francesco Guerra. Analyzing how BERT performs entity matching. *Proceedings of the VLDB Endowment*, 15(8):1726–1738, April 2022. ISSN 2150-8097. doi: 10.14778/3529337.3529356. URL <https://dl.acm.org/doi/10.14778/3529337.3529356>.
- [374] Michael 1933 Parenti. *Inventing Reality : The Politics of News Media*. St. Martin's Press, New York, 2nd ed. edition, 1993. ISBN 0-312-02013-9 978-0-312-02013-2 0-312-08629-6 978-0-312-08629-9.
- [375] Jaram Park, Meeyoung Cha, Hoh Kim, and Jaeseung Jeong. Managing Bad News in Social Media: A Case Study on Domino's Pizza Crisis. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, page 8, 2012.
- [376] Sungkyu Park, Sungwon Han, Jeongwook Kim, Mir Majid Molaie, Hoang Dieu Vu, Karandeep Singh, Jiyoung Han, Wonjae Lee, and Meeyoung Cha. COVID-19 Discourse on Twitter in Four Asian Countries: Case Study of Risk Communication. *Journal of Medical Internet Research*, 23(3):e23272, March 2021. doi: 10.2196/23272. URL <https://www.jmir.org/2021/3/e23272>.
- [377] Jon Passantino and Donie O'Sullivan. 'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion — CNN Business. *CNN*, May 2023. URL <https://www.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html>.
- [378] Michael Patty. Social Media and Censorship: Rethinking State Action Once Again. *Mitchell Hamline Law Journal of Public Policy and Practice*, 40:99–136, 2019.
- [379] J. Michael Pemberton and Christine R. Nugent. Information Studies: Emergent Field, Convergent Curriculum. *Journal of Education for Library and Information Science*, 36(2):126, 1995. ISSN 07485786. doi: 10.2307/40322913. URL <https://www.jstor.org/stable/10.2307/40322913?origin=crossref>.
- [380] Qiwei Peng, David Weir, Julie Weeds, and Yekun Chai. Predicate-Argument Based Bi-Encoder for Paraphrase Identification. In *Proceedings of the 60th Annual Meeting*

- of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5579–5589, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.382. URL <https://aclanthology.org/2022.acl-long.382>.
- [381] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL <http://aclweb.org/anthology/D14-1162>.
- [382] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7):770–780, June 2020.
- [383] Jay Peters. Twitter will now make it easier to see quote tweets on iOS, May 2020. URL <https://www.theverge.com/2020/5/12/21256586/twitter-ios-quote-tweets-retweets-with-comments>.
- [384] Chansavath Phetsouphanh, David R. Darley, Daniel B. Wilson, Annett Howe, C. Mee Ling Munier, Sheila K. Patel, Jennifer A. Juno, Louise M. Burrell, Stephen J. Kent, Gregory J. Dore, Anthony D. Kelleher, and Gail V. Matthews. Immunological dysfunction persists for 8 months following initial mild-to-moderate SARS-CoV-2 infection. *Nature Immunology*, 23(2):210–216, February 2022. ISSN 1529-2916. doi: 10.1038/s41590-021-01113-x. URL <https://www.nature.com/articles/s41590-021-01113-x>.
- [385] José R.C. Piqueira, Mauro Zilbovicius, and Cristiane M. Batistela. Daley–Kendal models in fake-news scenario. *Physica A: Statistical Mechanics and its Applications*, 548:123406, June 2020. ISSN 03784371. doi: 10.1016/j.physa.2019.123406. URL <https://linkinghub.elsevier.com/retrieve/pii/S0378437119319028>.
- [386] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. STANCY: Stance Classification Based on Consistency Cues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6412–6417, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1675. URL <https://www.aclweb.org/anthology/D19-1675>.

- [387] L Portnoff, J Schall, J Brannen, N Suhon, K Strickland, and J Meyers. Filtering facepiece respirators with an exhalation valve: Measurements of filtration efficiency to evaluate their potential for source control. Technical Report DHHS (NIOSH) Publication No. 2021-107, U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health, December 2020. URL <https://www.cdc.gov/niosh/docs/2021-107/>.
- [388] B. Aditya Prakash, Alex Beutel, Roni Rosenfeld, and Christos Faloutsos. Winner takes all: Competing viruses or ideas on fair-play networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1037–1046, Lyon France, April 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187975. URL <https://dl.acm.org/doi/10.1145/2187836.2187975>.
- [389] Psy Group. Project "Rome": Campaign Intelligence & Influence Services Proposal. Technical report, Psy Group, April 2016. URL <https://int.nyt.com/data/documenthelper/360-trump-project-rome/574d679d1ff58a30836c/optimized/full.pdf>.
- [390] Juliet R. C. Pulliam, Cari van Schalkwyk, Nevashan Govender, Anne von Gottberg, Cheryl Cohen, Michelle J. Groome, Jonathan Dushoff, Koleka Mlisana, and Harry Moultrie. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science*, 376(6593):eabn4947, March 2022. doi: 10.1126/science.abn4947. URL <https://www.science.org/doi/full/10.1126/science.abn4947>.
- [391] Hao Qin, Rupesh Kumar, and Romain Alléaume. Quantum hacking: Saturation attack on practical continuous-variable quantum key distribution. *Physical Review A*, 94(1):012325, July 2016. doi: 10.1103/PhysRevA.94.012325. URL <https://link.aps.org/doi/10.1103/PhysRevA.94.012325>.
- [392] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. Tool Learning with Foundation Models, April 2023. URL <http://arxiv.org/abs/2304.08354>.

- [393] Walter Quattrociocchi, Guido Caldarelli, and Antonio Scala. Opinion dynamics on interacting networks: Media competition and social influence. *Scientific Reports*, 4:4938, May 2014. ISSN 2045-2322. doi: 10.1038/srep04938. URL <http://www.nature.com/articles/srep04938>.
- [394] Ignacio Ojea Quintana, Ritsaart Reimann, Marc Cheong, Mark Alfano, and Colin Klein. Polarization and trust in the evolution of vaccine discourse on Twitter during COVID-19. *PLOS ONE*, 17(12):e0277292, December 2022. ISSN 1932-6203. doi: 10.1371/journal.pone.0277292. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0277292>.
- [395] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [396] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1317. URL <https://www.aclweb.org/anthology/D17-1317>.
- [397] Bhavtosh Rath, Wei Gao, Jing Ma, and Jaideep Srivastava. Utilizing computational trust to identify rumor spreaders on Twitter. *Social Network Analysis and Mining*, 8(1): 64, December 2018. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-018-0540-z. URL <http://link.springer.com/10.1007/s13278-018-0540-z>.
- [398] Adrian Rauchfleisch and Jonas Kaiser. The False positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10):e0241045, October 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0241045. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241045>.
- [399] Emma Remy. How public and private Twitter users in the U.S. compare — and why it might matter for your research, July 2019. URL <https://www.pewresearch.org/decoded/2019/07/15/how-public-and-private-twitter-users-in-the-u-s-compare-and-why-it-might-matter-for-your-research/>.
- [400] Yafeng Ren and Donghong Ji. Neural networks for deceptive opinion spam detection: An empirical study. *Information Sciences*, 385–386:213–224, April 2017. ISSN

- 0020-0255. doi: 10.1016/j.ins.2017.01.015. URL <http://www.sciencedirect.com/science/article/pii/S0020025517300166>.
- [401] Reuters Staff. What are the links between Cambridge Analytica and a Brexit campaign group? *Reuters*, March 2018. URL <https://www.reuters.com/article/us-facebook-cambridge-analytica-leave-eu-idUSKBN1GX2IO>.
- [402] Imogen Richards, Callum Jones, and Gearóid Brinn. Eco-Fascism Online: Conceptualizing Far-Right Actors' Response to Climate Change on Stormfront. *Studies in Conflict & Terrorism*, 0(0):1–27, December 2022. ISSN 1057-610X. doi: 10.1080/1057610X.2022.2156036. URL <https://doi.org/10.1080/1057610X.2022.2156036>.
- [403] Martin J. Riedl, Kelsey N. Whipple, and Ryan Wallace. Antecedents of support for social media content moderation and platform regulation: The role of presumed effects on self and others. *Information, Communication & Society*, 25(11):1632–1649, August 2022. ISSN 1369-118X. doi: 10.1080/1369118X.2021.1874040. URL <https://doi.org/10.1080/1369118X.2021.1874040>.
- [404] Michael Robinson, Kevin Jones, and Helge Janicke. Cyber warfare: Issues and challenges. *Computers & Security*, 49:70–94, March 2015. ISSN 0167-4048. doi: 10.1016/j.cose.2014.11.007. URL <https://www.sciencedirect.com/science/article/pii/S0167404814001692>.
- [405] William I. Robinson and Mario Barrera. Global capitalism and twenty-first century fascism: A US case study. *Race & Class*, 53(3):4–29, January 2012. ISSN 0306-3968. doi: 10.1177/0306396811425983. URL <https://doi.org/10.1177/0306396811425983>.
- [406] Allegra Rosenberg. Custom AI chatbots are quietly becoming the next big thing in fandom. *The Verge*, March 2023. URL <https://www.theverge.com/23627402/character-ai-fandom-chat-bots-fanfiction-role-playing>.
- [407] Allegra Rosenberg. AI-generated audio of Joe Biden and Donald Trump trashtalking while gaming is taking over TikTok. *Business Insider*, Mar 1, 2023, 2:06 PM. URL <https://www.businessinsider.com/voice-ai-audio-joe-biden-donald-trump-tiktok-2023-3>.
- [408] Astrid M. Rosenthal-von der Pütten, Matthias R. Hastall, Sören Köcher, Christian Meske, Timo Heinrich, Franziska Labrenz, and Sebastian Ocklenburg. “Likes” as social rewards: Their role in online social comparison and decisions to like other

- People's selfies. *Computers in Human Behavior*, 92:76–86, March 2019. ISSN 0747-5632. doi: 10.1016/j.chb.2018.10.017. URL <https://www.sciencedirect.com/science/article/pii/S0747563218305016>.
- [409] Björn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, July 2019. ISSN 0960-085X, 1476-9344. doi: 10.1080/0960085X.2018.1560920. URL <https://www.tandfonline.com/doi/full/10.1080/0960085X.2018.1560920>.
- [410] Wanying Ruan, Yongbin Sun, Yimin Deng, and Haibin Duan. Hawk-Pigeon Game Tactics for Unmanned Aerial Vehicle Swarm Target Defense. *IEEE Transactions on Industrial Informatics*, pages 1–11, 2023. ISSN 1941-0050. doi: 10.1109/TII.2023.3248075.
- [411] Victoria Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17, San Diego, California, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-0802. URL <http://aclweb.org/anthology/W16-0802>.
- [412] Phineas Rueckert. How a Journalist's Murder Led to a Global Investigation Into the Mercenaries Who Sell Disinformation and Destroy Democracies. *Haaretz*, February 2023. URL <https://www.haaretz.com/israel-news/security-aviation/2023-02-14/ty-article-magazine/.premium/how-a-journalists-murder-led-to-the-pursuit-of-cybermercenaries-who-destroy-democracies/00000186-4b01-da6c-a3f6-cb0bb4c60000>.
- [413] Michael Safi. Churchill's policies contributed to 1943 Bengal famine – study. *The Guardian*, March 2019. ISSN 0261-3077. URL <https://www.theguardian.com/world/2019/mar/29/winston-churchill-policies-contributed-to-1943-bengal-famine-study>.
- [414] David Saldaña, Amanda Prorok, Shreyas Sundaram, Mario F. M. Campos, and Vijay Kumar. Resilient consensus for time-varying networks of dynamic agents. In *2017 American Control Conference (ACC)*, pages 252–258, May 2017. doi: 10.23919/ACC.2017.7962962.

- [415] Emmanuelle Saliba. 'A real worry': How AI is making it harder to spot fake images, April 7, 2023, 6:08 AM. URL <https://abcnews.go.com/US/real-worry-ai-makin-g-harder-spot-fake-images/story?id=98188795>.
- [416] Malik Sallam. COVID-19 Vaccine Hesitancy Worldwide: A Concise Systematic Review of Vaccine Acceptance Rates. *Vaccines*, 9(2):160, February 2021. doi: 10.3390/vaccines9020160. URL <https://www.mdpi.com/2076-393X/9/2/160>.
- [417] Jean-Loup C. Samaan. Missiles, Drones, and the Houthis in Yemen. *Parameters*, 50(1):51–64, 2020. ISSN 00311723. URL <https://www.proquest.com/docview/2381628487/abstract/7DF74726B3034525PQ/1>.
- [418] Abraham Sanders, Rachael White, Lauren Severson, Rufeng Ma, Richard McQueen, Haniel Campos Alcanatara Paulo, Yucheng Zhang, John S. Erickson, and Kristin P. Bennett. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *medRxiv*, page 2020.08.28.20183863, September 2020. doi: 10.1101/2020.08.28.20183863. URL <https://www.medrxiv.org/content/10.1101/2020.08.28.20183863v2>.
- [419] Abraham C. Sanders, Rachael C. White, Lauren S. Severson, Rufeng Ma, Richard McQueen, Haniel C. Alcântara Paulo, Yucheng Zhang, John S. Erickson, and Kristin P. Bennett. Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of COVID-19 Twitter discourse. *AMIA Summits on Translational Science Proceedings*, 2021:555–564, May 2021. ISSN 2153-4063. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378598/>.
- [420] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, March 2020. URL <http://arxiv.org/abs/1910.01108>.
- [421] Kazutoshi Sasahara, Wen Chen, Hao Peng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402, May 2021. ISSN 2432-2725. doi: 10.1007/s42001-020-00084-7. URL <https://doi.org/10.1007/s42001-020-00084-7>.
- [422] Andreas Schedler. Electoral Authoritarianism. In *Emerging Trends in the Social and Behavioral Sciences*, pages 1–16. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-90077-2. doi: 10.1002/9781118900772.etrds0098. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118900772.etrds0098>.

- [423] Christoph Schweimer. Fast generation of simple directed social network graphs with reciprocal edges and high clustering. *Social Network Analysis and Mining*, 12(1):127, December 2022. ISSN 1869-5450, 1869-5469. doi: 10.1007/s13278-022-00963-z. URL <https://link.springer.com/10.1007/s13278-022-00963-z>.
- [424] Christoph Schweimer, Christine Gfrerer, Florian Lugstein, David Pape, Jan A. Velimsky, Robert Elsässer, and Bernhard C. Geiger. Generating Simple Directed Social Network Graphs for Information Spreading. In *Proceedings of the ACM Web Conference 2022*, pages 1475–1485, April 2022. doi: 10.1145/3485447.3512194. URL <http://arxiv.org/abs/2205.02485>.
- [425] Mark Scott. Digital Bridge: NAFO fellas — Aussie encryption vs. safety — UK lessons for California. *POLITICO*, September 2022. URL <https://www.politico.eu/newsletter/digital-bridge/nafo-fellas-aussie-encryption-vs-safety-uk-lessons-for-california/>.
- [426] Susan A. Seibert. Problem-based learning: A strategy to foster generation Z's critical thinking and perseverance. *Teaching and Learning in Nursing*, 16(1):85–88, January 2021. ISSN 1557-3087. doi: 10.1016/j.teln.2020.09.002. URL <https://www.sciencedirect.com/science/article/pii/S1557308720301049>.
- [427] Shaban Shabani and Maria Sokhn. Hybrid Machine-Crowd Approach for Fake News Detection. In *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*, pages 299–306, October 2018. doi: 10.1109/CIC.2018.00048.
- [428] Masaud Shah and Hyun Goo Woo. Omicron: A Heavily Mutated SARS-CoV-2 Variant Exhibits Stronger Binding to ACE2 and Potently Escapes Approved COVID-19 Therapeutic Antibodies. *Frontiers in Immunology*, 12:830527, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.830527.
- [429] Zubair Shah, Paige Martin, Enrico Coiera, Kenneth D. Mandl, and Adam G. Dunn. Modeling Spatiotemporal Factors Associated With Sentiment on Twitter: Synthesis and Suggestions for Improving the Identification of Localized Deviations. *Journal of Medical Internet Research*, 21(5):e12881, May 2019. doi: 10.2196/12881. URL <https://www.jmir.org/2019/5/e12881>.
- [430] Albin Shaji, R. V. Belfin, and E. Grace Mary Kanaga. An Innovated SIRS Model for Information Spreading. In Elijah Blessing Rajsingh, Jey Veerasamy, Amir H. Alavi, and J. Dinesh Peter, editors, *Advances in Big Data and Cloud Computing*, Advances

- in *Intelligent Systems and Computing*, pages 405–413, Singapore, 2018. Springer. ISBN 978-981-10-7200-0. doi: 10.1007/978-981-10-7200-0_37.
- [431] Jingyu Shao, Qing Wang, and Yu Lin. Skyblocking for entity resolution. *Information Systems*, 85:30–43, November 2019. ISSN 03064379. doi: 10.1016/j.is.2019.06.003. URL <https://linkinghub.elsevier.com/retrieve/pii/S0306437918304770>.
- [432] Jolin Shaynn-Ly Kwan and Kwan Hui Lim. Understanding Public Sentiments, Opinions and Topics about COVID-19 using Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 623–626, December 2020. doi: 10.1109/ASONAM49781.2020.9381384.
- [433] Lauren E. Sherman, Ashley A. Payton, Leanna M. Hernandez, Patricia M. Greenfield, and Mirella Dapretto. The Power of the Like in Adolescence: Effects of Peer Influence on Neural and Behavioral Responses to Social Media. *Psychological Science*, 27(7): 1027–1035, July 2016. ISSN 0956-7976. doi: 10.1177/0956797616645673. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5387999/>.
- [434] Hocheol Shin, Dohyun Kim, Yujin Kwon, and Yongdae Kim. Illusion and Dazzle: Adversarial Optical Channel Exploits Against Lidars for Automotive Applications. In Wieland Fischer and Naofumi Homma, editors, *Cryptographic Hardware and Embedded Systems – CHES 2017*, Lecture Notes in Computer Science, pages 445–467, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66787-4. doi: 10.1007/978-3-319-66787-4_22.
- [435] Inyoung Shin, Luxuan Wang, and Yi-Ta Lu. Twitter and Endorsed (Fake) News: The Influence of Endorsement by Strong Ties, Celebrities, and a User Majority on Credibility of Fake News During the COVID-19 Pandemic. *International Journal of Communication*, 16(0):23, April 2022. ISSN 1932-8036. URL <https://ijoc.org/index.php/ijoc/article/view/18187>.
- [436] Jieun Shin, Lian Jian, Kevin Driscoll, and François Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, June 2018. ISSN 0747-5632. doi: 10.1016/j.chb.2018.02.008. URL <https://www.sciencedirect.com/science/article/pii/S0747563218300669>.
- [437] James Shires and Max Smeets. Contesting ”Cyber”. Technical report, New America, December 2017. URL <https://d1y8sb8igg2f8e.cloudfront.net/documents/ContestingCyber12.12.17.pdf>.

- [438] Clay Shirky. The Political Power of Social Media: Technology, the Public Sphere, and Political Change. *Foreign Affairs*, 90(1):28–41, 2011. ISSN 0015-7120. URL <https://www.jstor.org/stable/25800379>.
- [439] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS ONE*, 6(5):e19467, May 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0019467. URL <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0019467>.
- [440] Sujoy Sikdar, Byungkyu Kang, John ODonovan, Tobias Höllerer, and Sibel Adah. Understanding Information Credibility on Twitter. In *2013 International Conference on Social Computing*, pages 19–24, September 2013. doi: 10.1109/SocialCom.2013.9.
- [441] Aybike Şimşek and Resul Kara. Using swarm intelligence algorithms to detect influential individuals for influence maximization in social networks. *Expert Systems with Applications*, 114:224–236, December 2018. ISSN 0957-4174. doi: 10.1016/j.eswa.2018.07.038. URL <https://www.sciencedirect.com/science/article/pii/S0957417418304561>.
- [442] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47, September 2019. doi: 10.1109/BigMM.2019.00-44.
- [443] Jason Jay Smart. Founder of NAFO Reveals Identity, Discusses Raison D’être. *Kyiv Post*, November 2022. URL <https://www.kyivpost.com/post/204>.
- [444] Paweł Sobkowicz. Social Depolarization and Diversity of Opinions—Unified ABM Framework. *Entropy*, 25(4):568, April 2023. ISSN 1099-4300. doi: 10.3390/e25040568. URL <https://www.mdpi.com/1099-4300/25/4/568>.
- [445] Paweł Sobkowicz and Antoni Sobkowicz. Agent Based Model of Anti-Vaccination Movements: Simulations and Comparison with Empirical Data. *Vaccines*, 9(8): 809, July 2021. ISSN 2076-393X. doi: 10.3390/vaccines9080809. URL <https://www.mdpi.com/2076-393X/9/8/809>.
- [446] Milena Sokolowska. Outsmarting SARS-CoV-2 by empowering a decoy ACE2. *Signal Transduction and Targeted Therapy*, 5(1):1–3, November 2020. ISSN 2059-3635. doi:

- 10.1038/s41392-020-00370-w. URL <https://www.nature.com/articles/s41392-020-00370-w>.
- [447] Yanhui Song, Kaiyang Wei, Siluo Yang, Fei Shu, and Junping Qiu. Analysis on the research progress of library and information science since the new century. *Library Hi Tech*, 41(4):1145–1157, August 2023. ISSN 0737-8831. doi: 10.1108/LHT-06-2020-0126. URL <https://www.emerald.com/insight/content/doi/10.1108/LHT-06-2020-0126/full/html>.
- [448] A. Srivastava, C. Chelmiss, and V. K. Prasanna. Social influence computation and maximization in signed networks with competing cascades. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 41–48, August 2015. doi: 10.1145/2808797.2809304.
- [449] Galen Stocking, Amy Mitchell, Katerina Eva Matsa, Regina Widjaya, Mark Jurkowitz, Shreenita Ghosh, Aaron Smith, Sarah Naseer, and Christopher St. Aubin. The Role of Alternative Social Media in the News and Information Environment. Technical report, Pew Research Center, October 2022. URL <https://www.pewresearch.org/journalism/2022/10/06/the-role-of-alternative-social-media-in-the-news-and-information-environment/>.
- [450] V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. The DARPA Twitter Bot Challenge. *Computer*, 49(6):38–46, June 2016. ISSN 0018-9162. doi: 10.1109/MC.2016.183.
- [451] Julia E. Sullivan and Dmitriy Kamensky. How cyber-attacks in Ukraine show the vulnerability of the U.S. power grid. *The Electricity Journal*, 30(3):30–35, April 2017. ISSN 1040-6190. doi: 10.1016/j.tej.2017.02.006. URL <https://www.sciencedirect.com/science/article/pii/S1040619017300507>.
- [452] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to Fine-Tune BERT for Text Classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 194–206, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32381-3. doi: 10.1007/978-3-030-32381-3_16.
- [453] Heshan Sun. Sellers’ Trust and Continued Use of Online Marketplaces. *Journal of the Association for Information Systems*, 11(4), April 2010. ISSN 1536-9323. doi: 10.17705/1jais.00226. URL <https://aisel.aisnet.org/jais/vol11/iss4/2>.

- [454] Gerald Sussman and Sascha Krader. Template Revolutions: Marketing U.S. Regime Change in Eastern Europe. *Westminster Papers in Communication and Culture*, 5 (3):91, August 2008. ISSN 1744-6716, 1744-6708. doi: 10.16997/wpcc.95. URL <http://www.westminsterpapers.org/article/10.16997/wpcc.95/>.
- [455] Nicolas P. Suzor, Sarah Myers West, Andrew Quodling, and Jillian York. What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. *International Journal of Communication*, 13(0):18, March 2019. ISSN 1932-8036. URL <https://ijoc.org/index.php/ijoc/article/view/9736>.
- [456] David Sylvan and Stephen Majeski. An Agent-Based Model of the Acquisition of U.S. Client States. In *Paper Prepared for Presentation at the 44th Annual Convention of the International Studies Association*, Portland, OR, 2003.
- [457] Richard Szafranski. A theory of information warfare: Preparing for 2020. *Airpower Journal*, 9(1):56–66, 1995. ISSN 08970823. URL <https://proxy.library.mcgill.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9504172674&scope=site>.
- [458] Anna Szolucha. Real Politics in Occupy: Transcending the Rules of the Day. *Globalizations*, 12(1):66–82, January 2015. ISSN 1474-7731. doi: 10.1080/14747731.2014.971532. URL <https://doi.org/10.1080/14747731.2014.971532>.
- [459] Daiki Takeuchi, Gouhei Tanaka, Ryo Fujie, and Hideyuki Suzuki. Public opinion formation with the spiral of silence on complex social networks. *Nonlinear Theory and Its Applications, IEICE*, 6(1):15–25, 2015. ISSN 2185-4106. doi: 10.1587/nolta.6.15. URL https://www.jstage.jst.go.jp/article/nolta/6/1/6_15/_article.
- [460] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153, February 2018. ISSN 2167-0811, 2167-082X. doi: 10.1080/21670811.2017.1360143. URL <https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143>.
- [461] Byron Tau and Rebecca Ballhaus. Israeli Intelligence Company Formed Venture With Trump Campaign Firm Cambridge Analytica. *Wall Street Journal*, May 2018. ISSN 0099-9660. URL <https://www.wsj.com/articles/israeli-intelligence-company-formed-venture-with-trump-campaign-firm-cambridge-analytica-1527030765>.

- [462] Tommaso Teofili, Donatella Firmani, Nick Koudas, Vincenzo Martello, Paolo Merialdo, and Divesh Srivastava. Effective Explanations for Entity Resolution Models. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2709–2721, May 2022. doi: 10.1109/ICDE53745.2022.00248.
- [463] TOI Staff and AP. Trump campaign mined Facebook user data using Israeli ‘intelligence gathering’. *Times of Israel*, 20 March 2018, 3:42 am. URL <https://www.timesofisrael.com/trump-campaign-mined-facebook-user-data-using-israeli-intelligence-gathering/>.
- [464] Cagri Toraman, Furkan Şahinuç, Eyup Halit Yilmaz, and Ibrahim Batuhan Akkaya. Understanding social engagements: A comparative analysis of user and text features in Twitter. *Social Network Analysis and Mining*, 12(1):47, March 2022. ISSN 1869-5469. doi: 10.1007/s13278-022-00872-1. URL <https://doi.org/10.1007/s13278-022-00872-1>.
- [465] Robyn Torok. Developing an explanatory model for the process of online radicalisation and terrorism. *Security Informatics*, 2(1):6, February 2013. ISSN 2190-8532. doi: 10.1186/2190-8532-2-6. URL <https://doi.org/10.1186/2190-8532-2-6>.
- [466] Cassandra Lynn-Collins Troy. Get Ratioed: Questioning the Fossil Fuel Industry’s Social License to Operate on Twitter. *The Journal of Public Interest Communications*, 6(1):4, June 2022. ISSN 2573-4342. doi: 10.32473/jpic.v6.i1.p4. URL <https://journals.flvc.org/jpic/article/view/129248>.
- [467] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Chengliang Chai, Guoliang Li, Ruixue Fan, and Xiaoyong Du. Domain Adaptation for Deep Entity Resolution. In *Proceedings of the 2022 International Conference on Management of Data, SIGMOD ’22*, pages 443–457, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9249-5. doi: 10.1145/3514221.3517870. URL <https://doi.org/10.1145/3514221.3517870>.
- [468] Zeynep Tufekci and Christopher Wilson. Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square. *Journal of Communication*, 62(2):363–379, April 2012. ISSN 0021-9916. doi: 10.1111/j.1460-2466.2012.01629.x. URL <https://doi.org/10.1111/j.1460-2466.2012.01629.x>.
- [469] Amos Tversky and Daniel Kahneman. The Framing of Decisions and the Psychology of Choice. *Science*, 211(4481):453–458, January 1981. doi: 10.1126/science.7455683. URL <https://www.science.org/doi/abs/10.1126/science.7455683>.

- [470] Twitter. Don't miss the Tweets about your Tweet. Now on iOS, you can see Retweets with comments all in one place. <https://t.co/oanjZfzC6y>, May 2020. URL <https://twitter.com/Twitter/status/1260294888811347969>.
- [471] Joshua Uyheng, Lynnette Hui Xian Ng, and Kathleen M. Carley. Active, aggressive, but to little avail: Characterizing bot activity during the 2020 Singaporean elections. *Computational and Mathematical Organization Theory*, 27(3):324–342, September 2021. ISSN 1572-9346. doi: 10.1007/s10588-021-09332-1. URL <https://doi.org/10.1007/s10588-021-09332-1>.
- [472] Vaibhav Vaibhav, Raghuram Mandyam, and Eduard Hovy. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 134–139, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5316. URL <https://www.aclweb.org/anthology/D19-5316>.
- [473] Jens van 't Klooster and Clément Fontan. The Myth of Market Neutrality: A Comparative Study of the European Central Bank's and the Swiss National Bank's Corporate Security Purchases. *New Political Economy*, 25(6):865–879, September 2020. ISSN 1356-3467. doi: 10.1080/13563467.2019.1657077. URL <https://doi.org/10.1080/13563467.2019.1657077>.
- [474] Lakshmi Varanasi. Twitter bots appear to be in line with the company's estimate of below 5% — but you wouldn't know it from how much they tweet, researchers say. *Business Insider*, Sep 8, 2022, 7:48 PM EDT. URL <https://www.businessinsider.com/twitter-bots-comprise-less-than-5-but-tweet-more-2022-9>.
- [475] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>.
- [476] Alessandro Vespignani. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39, January 2012. ISSN 1745-2481. doi: 10.1038/nphys2160. URL <https://www.nature.com/articles/nphys2160>.

- [477] Marco A. Vieira. Understanding Resilience in International Relations: The Non-Aligned Movement and Ontological Security. *International Studies Review*, 18(2): 290–311, June 2016. ISSN 1521-9488. doi: 10.1093/isr/viw002. URL <https://doi.org/10.1093/isr/viw002>.
- [478] Ramon Villa-Cox, Sumeet Kumar, Matthew Babcock, and Kathleen M. Carley. Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Twitter Conversations, June 2020. URL <http://arxiv.org/abs/2006.00691>.
- [479] Amiram Vinokur and Eugene Burnstein. Depolarization of attitudes in groups. *Journal of Personality and Social Psychology*, 36(8):872–885, 1978. ISSN 1939-1315. doi: 10.1037/0022-3514.36.8.872. URL <https://psycnet.apa.org/fulltext/1980-00965-001.pdf>.
- [480] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H. Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander S. Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom L. Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, November 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1724-z. URL <http://www.nature.com/articles/s41586-019-1724-z>.
- [481] Madalina Vlasceanu and Alin Coman. The impact of information sources on COVID-19 knowledge accumulation and vaccination intention. *International Journal of Data Science and Analytics*, pages 1–12, January 2022. ISSN 2364-415X. doi: 10.1007/s41060-021-00307-8. URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8751473/>.
- [482] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2102. URL <http://aclweb.org/anthology/P17-2102>.

- [483] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, March 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aap9559. URL <https://science.sciencemag.org/content/359/6380/1146>.
- [484] Claes H. Vreese. News framing: Theory and typology. *Information Design Journal*, 13(1):51–62, April 2005. ISSN 0142-5471, 1569-979X. doi: 10.1075/idjdd.13.1.06vre. URL <http://www.jbe-platform.com/content/journals/10.1075/idjdd.13.1.06vre>.
- [485] Chao Wang, Zong Xuan Tan, Ye Ye, Lu Wang, Kang Hao Cheong, and Neng-gang Xie. A rumor spreading model based on information entropy. *Scientific Reports*, 7(1):9615, August 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-09171-8. URL <https://www.nature.com/articles/s41598-017-09171-8>.
- [486] Eric Wang. Prediction of antibody binding to SARS-CoV-2 RBDs. *Bioinformatics Advances*, 3(1):vbac103, January 2023. ISSN 2635-0041. doi: 10.1093/bioadv/vbac103. URL <https://doi.org/10.1093/bioadv/vbac103>.
- [487] Guanghui Wang, Xuefeng Sun, Liping Zhang, and Chao Lv. Saturation attack based route planning and threat avoidance algorithm for cruise missiles. *Journal of Systems Engineering and Electronics*, 22(6):948–953, December 2011. ISSN 1004-4132. doi: 10.3969/j.issn.1004-4132.2011.06.011.
- [488] Haopei Wang, Lei Xu, and Guofei Gu. FloodGuard: A DoS Attack Prevention Extension in Software-Defined Networks. In *2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pages 239–250, June 2015. doi: 10.1109/DSN.2015.27.
- [489] Jun Wang, Shimin Cai, Wei Wang, and Tao Zhou. Link cooperation effect of cooperative epidemics on complex networks. *Applied Mathematics and Computation*, 437:127537, January 2023. ISSN 00963003. doi: 10.1016/j.amc.2022.127537. URL <https://linkinghub.elsevier.com/retrieve/pii/S0096300322006117>.
- [490] Qing Wang, Mingyuan Cui, and Huizhi Liang. Semantic-Aware Blocking for Entity Resolution. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):166–180, January 2016. ISSN 1041-4347. doi: 10.1109/TKDE.2015.2468711. URL <http://ieeexplore.ieee.org/document/7202898/>.

- [491] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-Shot Talking Face Generation from Single-Speaker Audio-Visual Correlation Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2531–2539, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i3.20154. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20154>.
- [492] Wei Wang, Quan-Hui Liu, Junhao Liang, Yanqing Hu, and Tao Zhou. Coevolution spreading in complex networks. *Physics Reports*, July 2019. ISSN 0370-1573. doi: 10.1016/j.physrep.2019.07.001. URL <http://www.sciencedirect.com/science/article/pii/S0370157319302583>.
- [493] Xiaohan Wang, Fakui Wang, Zisen Nie, Yuhui Ai, and Tianjiang Hu. optiSwarm: Optical Swarm Robots Using Implicit Cooperation. *IEEE Sensors Journal*, 22(24): 24380–24394, December 2022. ISSN 1558-1748. doi: 10.1109/JSEN.2022.3216445.
- [494] Claire Wardle. Fake news. It’s complicated., February 2017. URL <https://firstdraftnews.org:443/latest/fake-news-complicated/>.
- [495] Gunnar Weimann, Beatrice Berton, and Antonios Samouris. IS’s global digital arenas: Strategy and platforms. In *Salafi-Jihadism and Digital Media*. Routledge, 2022. ISBN 978-1-00-326131-5.
- [496] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2:335, March 2012. ISSN 2045-2322. doi: 10.1038/srep00335. URL <http://www.nature.com/articles/srep00335>.
- [497] Richard F. West, Russell J. Meserve, and Keith E. Stanovich. Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, 103(3):506–519, September 2012. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0028857. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/a0028857>.
- [498] Stephen White and Ian McAllister. Did Russia (Nearly) have a Facebook Revolution in 2011? Social Media’s Challenge to Authoritarianism. *Politics*, 34(1):72–84, February 2014. ISSN 0263-3957, 1467-9256. doi: 10.1111/1467-9256.12037. URL <http://journals.sagepub.com/doi/10.1111/1467-9256.12037>.
- [499] James Q. Whitman. Hitler’s American Model: The United States and the Making of Nazi Race Law. In *Hitler’s American Model*. Princeton University Press, February

2017. ISBN 978-1-4008-8463-6. doi: 10.1515/9781400884636. URL <https://www.degruyter.com/document/doi/10.1515/9781400884636/html>.
- [500] Aurélie Wiedemann, Emile Foucat, Hakim Hocini, Cécile Lefebvre, Boris P. Hejblum, Mélanie Durand, Miriam Krüger, Alpha Kabinet Keita, Ahidjo Ayoub, Stéphane Mély, José-Carlos Fernandez, Abdoulaye Touré, Slim Fourati, Claire Lévy-Marchal, Hervé Raoul, Eric Delaporte, Lamine Koivogui, Rodolphe Thiébaud, Christine Lacabaratz, and Yves Lévy. Long-lasting severe immune dysfunction in Ebola virus disease survivors. *Nature Communications*, 11(1):3730, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17489-7. URL <https://www.nature.com/articles/s41467-020-17489-7>.
- [501] Wayne A. Wiegand. Libraries and the Invention of Information. In *A Companion to the History of the Book*, chapter 55, pages 827–840. John Wiley & Sons, Ltd, 2019. ISBN 978-1-119-01819-3. doi: 10.1002/9781119018193.ch55. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119018193.ch55>.
- [502] Malith Wijesekara and Gamage Upeksha Ganegoda. Source credibility analysis on Twitter users. In *2020 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 96–102, September 2020. doi: 10.1109/SCSE49731.2020.9313064.
- [503] Richard Wike, Laura Silver, Janell Fetterolf, Christine Huang, Sarah Austin, Laura Clancy, and Sneha Gubbala. Social Media Seen as Mostly Good for Democracy Across Many Nations, But U.S. is a Major Outlier. Technical report, Pew Research Center, December 2022. URL https://www.pewresearch.org/global/wp-content/uploads/sites/2/2022/12/PG_2022.12.06_Online-Civic-Engagement_REPORT.pdf.
- [504] Richard Wilson. Cambridge Analytica, Facebook, and Influence Operations: A Case Study and Anticipatory Ethical Analysis. In *European Conference on Cyber Warfare and Security*, pages 587–595, XX, Reading, United Kingdom, July 2019. Academic Conferences International Limited. URL <https://www.proquest.com/docview/2261006731/abstract/7535B9A14F504BB7PQ/1>.
- [505] Donna Winslow. Rites of Passage and Group Bonding in the Canadian Airborne. *Armed Forces & Society*, 25(3):429–457, 1999. ISSN 0095-327X. URL <https://www.jstor.org/stable/45346316>.
- [506] Olivier J. Wouters, Kenneth C. Shadlen, Maximilian Salcher-Konrad, Andrew J. Pollard, Heidi J. Larson, Yot Teerawattananon, and Mark Jit. Challenges in ensur-

- ing global access to COVID-19 vaccines: Production, affordability, allocation, and deployment. *The Lancet*, 397(10278):1023–1034, March 2021. ISSN 0140-6736, 1474-547X. doi: 10.1016/S0140-6736(21)00306-8. URL [https://www.thelancet.com/journals/lancet/article/PIIS0140-6736\(21\)00306-8/abstract](https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00306-8/abstract).
- [507] Jinhua Wu, Peibei Ma, and Jun Ji. Research on Cooperative Control Method of Saturation Attack. In *2007 IEEE International Conference on Automation and Logistics*, pages 1409–1413, August 2007. doi: 10.1109/ICAL.2007.4338791.
- [508] Leyun Wu, Liping Zhou, Mengxia Mo, Tingting Liu, Chengkun Wu, Chunye Gong, Kai Lu, Likun Gong, Weiliang Zhu, and Zhijian Xu. SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2. *Signal Transduction and Targeted Therapy*, 7(1):1–3, January 2022. ISSN 2059-3635. doi: 10.1038/s41392-021-00863-2. URL <https://www.nature.com/articles/s41392-021-00863-2>.
- [509] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv:1609.08144 [cs]*, October 2016. URL <http://arxiv.org/abs/1609.08144>.
- [510] Bo Xie, Daqing He, Tim Mercer, Youfa Wang, Dan Wu, Kenneth R. Fleischmann, Yan Zhang, Linda H. Yoder, Keri K. Stephens, Michael Mackert, and Min Kyung Lee. Global health crises are also information crises: A call to action. *Journal of the Association for Information Science and Technology*, n/a(n/a):1–5, March 2020. ISSN 2330-1643. doi: 10.1002/asi.24357. URL <https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24357>.
- [511] Bo Xu, Yun Huang, Haewoon Kwak, and Noshir Contractor. Structures of broken ties: Exploring unfollow behavior on twitter. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 871–876, San Antonio Texas USA, February 2013. ACM. ISBN 978-1-4503-1331-5. doi: 10.1145/2441776.2441875. URL <https://dl.acm.org/doi/10.1145/2441776.2441875>.

- [512] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. MobileFaceSwap: A Lightweight Framework for Video Face Swapping. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2973–2981, June 2022. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v36i3.20203. URL <https://ojs.aaai.org/index.php/AAAI/article/view/20203>.
- [513] Fei Yang, Chaoyang Dong, and Qing Wang. Multi-Target Decision-Making of Saturation Attack for Anti-Ship Missile Weapon-Target Assignment. In *2009 International Workshop on Intelligent Systems and Applications*, pages 1–4, May 2009. doi: 10.1109/IWISA.2009.5072970.
- [514] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf>.
- [515] Dezhong Yao, Yuhong Gu, Gao Cong, Hai Jin, and Xinqiao Lv. Entity Resolution with Hierarchical Graph Attention Networks. In *Proceedings of the 2022 International Conference on Management of Data, SIGMOD ’22*, pages 429–442, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9249-5. doi: 10.1145/3514221.3517872. URL <https://doi.org/10.1145/3514221.3517872>.
- [516] Ercan Yildiz, Asuman Ozdaglar, Daron Acemoglu, Amin Saberi, and Anna Scaglione. Binary Opinion Dynamics with Stubborn Agents. *ACM Trans. Econ. Comput.*, 1(4): 19:1–19:30, December 2013. ISSN 2167-8375. doi: 10.1145/2538508. URL <http://doi.acm.org/10.1145/2538508>.
- [517] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4576–4585, June 2021. doi: 10.1109/CVPR46437.2021.00455.
- [518] Xinran Yu, Chao Zhong, Dandan Li, and Wei Xu. Sentiment analysis for news and social media in COVID-19. In *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Emergency Management Using GIS, EM-GIS ’20*, pages 1–4, Washington, Seattle, November 2020. Association for Computing Machinery. ISBN 978-1-4503-

- 8159-8. doi: 10.1145/3423333.3431794. URL <https://doi.org/10.1145/3423333.3431794>.
- [519] Xiaoyi Yuan, Ross J. Schuchard, and Andrew T. Crooks. Examining Emergent Communities and Social Bots Within the Polarized Online Vaccination Debate in Twitter. *Social Media + Society*, 5(3):2056305119865465, April 2019. ISSN 2056-3051. doi: 10.1177/2056305119865465. URL <https://doi.org/10.1177/2056305119865465>.
- [520] Savvas Zannettou. "I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv:2101.07183 [cs]*, April 2021. URL <http://arxiv.org/abs/2101.07183>.
- [521] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *Journal of Data and Information Quality*, 11(3):10:1–10:37, May 2019. URL <http://doi.org/10.1145/3309699>.
- [522] Dongxiang Zhang, Dongsheng Li, Long Guo, and Kian-Lee Tan. Unsupervised Entity Resolution with Blocking and Graph Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020. ISSN 1558-2191. doi: 10.1109/TKDE.2020.2991063.
- [523] Hegui Zhang, Xiaolong Chen, Yi Peng, Gang Kou, and Ruijie Wang. The interaction of multiple information on multiplex social networks. *Information Sciences*, 605: 366–380, August 2022. ISSN 00200255. doi: 10.1016/j.ins.2022.05.036. URL <https://linkinghub.elsevier.com/retrieve/pii/S002002552200456X>.
- [524] Tao Zhang, Jin Zhang, Chengfu Huo, and Weijun Ren. Automatic Generation of Pattern-controlled Product Description in E-commerce. In *The World Wide Web Conference, WWW '19*, pages 2355–2365, New York, NY, USA, May 2019. Association for Computing Machinery. ISBN 978-1-4503-6674-8. doi: 10.1145/3308558.3313407. URL <https://dl.acm.org/doi/10.1145/3308558.3313407>.
- [525] Zizhu Zhang, Weiliang Zhao, Jian Yang, Cecile Paris, and Surya Nepal. Learning Influence Probabilities and Modelling Influence Diffusion in Twitter. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1087–1094, San Francisco USA, May 2019. ACM. ISBN 978-1-4503-6675-5. doi: 10.1145/3308560.3316701. URL <https://dl.acm.org/doi/10.1145/3308560.3316701>.

- [526] Laijun Zhao, Jiajia Wang, Yucheng Chen, Qin Wang, Jingjing Cheng, and Hongxin Cui. SIHR rumor spreading model in social networks. *Physica A: Statistical Mechanics and its Applications*, 391(7):2444–2453, April 2012. ISSN 0378-4371. doi: 10.1016/j.physa.2011.12.008. URL <https://www.sciencedirect.com/science/article/pii/S0378437111009058>.
- [527] Zilong Zhao, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, Junjie Wu, and Shlomo Havlin. Fake news propagates differently from real news even at early stages of spreading. *EPJ Data Science*, 9(1):7, December 2020. ISSN 2193-1127. doi: 10.1140/epjds/s13688-020-00224-z. URL <https://epjdata.science.springeropen.com/articles/10.1140/epjds/s13688-020-00224-z>.
- [528] Daming Zhou, Wanwisa Dejnirattisai, Piyada Supasa, Chang Liu, Alexander J. Mentzer, Helen M. Ginn, Yuguang Zhao, Helen M. E. Duyvesteyn, Aekkachai Tuekprakhon, Rungtiwa Nutalai, Beibei Wang, Guido C. Paesen, Cesar Lopez-Camacho, Jose Slon-Campos, Bassam Hallis, Naomi Coombes, Kevin Bewley, Sue Charlton, Thomas S. Walter, Donal Skelly, Sheila F. Lumley, Christina Dold, Robert Levin, Tao Dong, Andrew J. Pollard, Julian C. Knight, Derrick Crook, Teresa Lambe, Elizabeth Clutterbuck, Sagida Bibi, Amy Flaxman, Mustapha Bittaye, Sandra Belij-Rammerstorfer, Sarah Gilbert, William James, Miles W. Carroll, Paul Klenerman, Eleanor Barnes, Susanna J. Dunachie, Elizabeth E. Fry, Juthathip Mongkolsapaya, Jingshan Ren, David I. Stuart, and Gavin R. Screaton. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell*, February 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.02.037. URL <https://www.sciencedirect.com/science/article/pii/S0092867421002269>.
- [529] Yanqiu Rachel Zhou and William D. Coleman. Accelerated Contagion and Response: Understanding the Relationships among Globalization, Time, and Disease. *Globalizations*, 13(3):285–299, May 2016. ISSN 1474-7731. doi: 10.1080/14747731.2015.1056498. URL <https://doi.org/10.1080/14747731.2015.1056498>.
- [530] Linhong Zhu, Majid Ghasemi-Gol, Pedro Szekely, Aram Galstyan, and Craig A. Knoblock. Unsupervised Entity Resolution on Multi-type Graphs. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, Lecture Notes in Computer Science, pages 649–667, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46523-4. doi: 10.1007/978-3-319-46523-4_39.

- [531] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: Dynamic Memory Generative Adversarial Networks for Text-To-Image Synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5795–5803, Long Beach, CA, USA, June 2019. IEEE. ISBN 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00595. URL <https://ieeexplore.ieee.org/document/8954283/>.
- [532] Melissa Zimdars. False, Misleading, Clickbait-y, and Satirical “News” Sources, 2016. URL https://docs.google.com/document/d/10eA5-mCZLSS4MQY5QGb5ewC3VAL6pLkT53V_81ZyitM/preview?usp=embed_facebook.
- [533] Melissa Zimdars. My ‘fake news list’ went viral. But made-up stories are only part of the problem. *Washington Post*, 2016. ISSN 0190-8286. URL <https://www.washingtonpost.com/posteverything/wp/2016/11/18/my-fake-news-list-went-viral-but-made-up-stories-are-only-part-of-the-problem/>.
- [534] Florian Zollmann. Bringing Propaganda Back into News Media Studies. *Critical Sociology*, 45(3):329–345, May 2019. ISSN 0896-9205, 1569-1632. doi: 10.1177/0896920517731134. URL <http://journals.sagepub.com/doi/10.1177/0896920517731134>.
- [535] Diana Zulli, Miao Liu, and Robert Gehl. Rethinking the “social” in “social media”: Insights into topology, abstraction, and scale on the Mastodon social network. *New Media & Society*, 22(7):1188–1205, July 2020. ISSN 1461-4448, 1461-7315. doi: 10.1177/1461444820912533. URL <http://journals.sagepub.com/doi/10.1177/1461444820912533>.