Abusive Language Through the Lens of Online Communities

Haji Mohammad Saleem

Doctor of Philosophy

School of Computer Science
McGill University
Montreal, Quebec, Canada

December 15, 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Ph.D. in Computer Science

DEDICATION

To my parents, whose hard work and support made this possible. My sister Vareesha, who is my biggest supporter. My love Rowan and the light of all of our lives - Noor.

ACKNOWLEDGEMENTS

I would like to thank my fellow researcher Jana Kurrek for her continued help through this research journey. My deepest gratitude to my supervisor, Derek Ruths, for his patience, guidance, and mentoring over the past decade. I would also like to thank my lab mates Aengus Bridgman, Caitrin Armstrong, Deven Parekh, Edward Newell, Matt Gittings, Meghan Keenan, Koustuv Sinha, Oleg Zhillin, and Sunyam Bagga for their support and camaraderie. Finally, I owe my thanks to a few others, whose presence has made my life better and brighter. My dear friends Kate Gilbert, Charana and Swetha Rajagopal, Chu Wang, and Faizy Ahsan for making this process much more pleasant. My cats Bunnu, Toby, and Meimei for all their cuddles and affection. And most important, my partner Rowan Scullion for their love and Noor for being so cute.

ABSTRACT

Abusive language is pervasive across online platforms with user-generated content. Such language can cause severe harm and its moderation is critical for maintaining safe spaces online. However, detection of such content is challenging - abusive language is extremely diverse and we lack precise terminology as well as representative training resources for broad analysis. In this thesis, I address these issues by developing a formal taxonomy and introducing community-driven research of such abuse. Focusing on online communities allows me to ensure diversity in abusive language resources and add context to abuse detection. Through the taxonomy, I engage with such diversity and perform detailed evaluation of detection systems. Together, this taxonomy and online communities play a central and unifying role in my contributions. First, I present "polar" communities - self-identifying online communities that antagonize and support the same marginalized group - and establish them as a potential resource for training language models. Next, I construct a taxonomy of pejorative expressions for fine-grained assessment of abusive language and use it to annotate a large-scale corpus aggregated by sampling polar communities on Reddit. Through the taxonomy, I further analyze key elements that define abusive language corpora and find that both keyword filters and source platforms affect the corpus outcomes in different ways. Thereafter, I investigate the role of community context within abusive language detection by evaluating the performance of contextualized models across the taxonomy labels. I observe that the additional information helps reduce false-positive errors for language models as well as humans. Finally, I conduct a case-study on the banning of an antagonistic community as a platform moderation tool and discover this action reduces user-engagement but requires collective effort from platform administrators, moderators, and users. Overall, this thesis highlights the importance of integrating online communities and a detailed taxonomy in all facets of abusive language research.

ABRÉGÉ

Le langage abusif est omniprésent sur les plateformes en ligne avec du contenu généré par les utilisateurs. Ce langage peut causer de graves préjudices et sa modération est essentielle pour maintenir des espaces sûrs en ligne. Cependant, la détection de ce type de contenu est un défi - le langage abusif est extrêmement diversifié et nous manquons de terminologie précise ainsi que de ressources de formation représentatives pour une analyse large. Dans cette thèse, j'aborde ces questions en développant une taxonomie formelle et en introduisant la recherche de ces abus par les communautés. Le fait de me concentrer sur les communautés en ligne me permet de garantir la diversité des ressources en matière de langage abusif et d'ajouter un contexte à la détection des abus. Grâce à la taxonomie, je m'engage dans une telle diversité et effectue une évaluation détaillée des systèmes de détection. Ensemble, cette taxonomie et les communautés en ligne jouent un rôle central et unificateur dans mes contributions. Tout d'abord, je présente les communautés "polaires" - des communautés en ligne qui s'identifient elles-mêmes comme antagonistes et soutenant le même groupe marginalisé - et les considère comme une ressource potentielle pour l'entraînement des modèles de langage. Ensuite, je construis une taxonomie d'expressions péjoratives pour une évaluation fine du langage abusif et l'utilise pour annoter un corpus à grande échelle agrégé par échantillonnage des communautés polaires sur Reddit. Grâce à la taxonomie, je poursuis l'analyse des éléments clés qui définissent les corpus de langage abusif et je constate que les filtres de mots-clés et les plateformes sources affectent les résultats du corpus de différentes manières. Par la suite, j'étudie le rôle du contexte communautaire dans la détection du langage abusif en évaluant la performance des modèles contextualisés à travers les étiquettes de la taxonomie. J'observe que les informations supplémentaires permettent de réduire les erreurs fausses-positives pour les modèles de langage ainsi que pour les humains. Enfin, je réalise une étude de cas sur le bannissement d'une communauté antagoniste en tant qu'outil de modération de la plateforme et découvre que cette action réduit l'engagement des utilisateurs mais nécessite un effort collectif de la part des administrateurs de la plateforme, des modérateurs et des utilisateurs. Dans l'ensemble, cette thèse souligne l'importance d'intégrer les communautés en ligne et une taxonomie détaillée dans toutes les facettes de la recherche sur le langage abusif.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Online social media platforms have witnessed mass adoption around the globe. For example, Facebook user base has grown from 431 million to 2.8 billion monthly active users[1] over the past decade - an increase of almost 550%. Online platforms have led to increased connectivity, but they rely on high user-engagement and user-generated content to drive platform growth. Handling such high volumes of user-generated content is not a trivial task and brings forth a unique set of challenges - one of which is the prevalence of abusive language.

Abusive language is rampant across a wide variety of online spaces where user content is the norm. It is not limited to traditional popular platforms such as Facebook [50, 16], Twitter [32, 231], Reddit [156, 31], and Youtube [68, 69] but is also a problem in online video and their in-game chats [53, 28] as well as the comments section of news websites [244, 73].

Exposure to abusive language can alienate users and reduce their overall engagement with the platform [4], harming the health of the social space. On an individual level its effects are grim and can lead to a wide range of psychological harms, including degradation of mental health, depression, reduced self-esteem, and enhanced

[1] `investor.fb.com/home/default.aspx`

stress expression [193, 219, 23]. Furthermore online abuse can even incite real-world violence [74]. Detection of such content is of growing interest to both online platform maintainers as well as government regulators since it is crucial for operating safe online spaces. Despite this interest, abusive language research faces a multitude of challenges and remains an open problem.

Some of these challenges stem from the fact that abusive language is highly variable. Linguistically, abusive language varies based on the marginalized group it targets. For example "*make me a sandwich*" is a misogynistic phrase used to mock or discredit women [157]. Jewish people are commonly degraded through racial stereotypes such as "*greedy, scheming, and stingy*" [110]. Different populations are therefore abused with different language. Furthermore, even a singular marginalized group faces abuse in a myriad of ways. Overall, abusive language assumes a wide-range of forms. As I go on to explain in Chapter 2, such diversity in content makes the detection of abusive language difficult to operationalize.

This thesis is predicated on two key observations: First, even though researchers agree that abusive language is diverse, abusive language research does not have a cohesive conceptual framework that address this diversity in a structured manner. The vast majority of the literature follows a binary paradigm with related but not equivalent terms (hateful, abusive, toxic, offensive, derogatory, etc). These binary frameworks fail to acknowledge the nuanced diversity in abuse and are often incomparable across studies. Second, reliable detection of abuse requires language resources that reflect diversity in abusive content. However, we lack a principled approach for collecting diverse and representative examples of abuse. Researchers have predominantly relied on keywords to filter data. This can and does introduce inadvertent biases in the resulting corpus [233]. In a sense, these two problems are related to one another because, together, they contribute to a lack of true visibility around the phenomenon of derogatory language.

The contributions in this thesis tackle the problem of diversity within the many aspects of abusive language research. A common trend across these contributions is the exploration of online communities across several avenues aided by a formal taxonomy of abusive expressions.

- In Chapter 3, I establish that self-identifying antagonistic communities are a valid source of abusive language.
- In Chapter 4, I build a precise taxonomy of online abuse and use it to annotate a diverse corpus collected through sampling self-identifying communities.
- In Chapter 5, I apply the taxonomy in the analysis of key elements that define an abuse corpus - how and from where abusive content was collected.
- In Chapter 6, I evaluate language models across taxonomy labels and discover that integrating community context improves abusive language classification.
- In Chapter 7, I study the banning of an antagonistic community as a moderation tool against abusive behaviour.

Together, these contributions highlight how a formal taxonomy and online communities can extensively assist abusive language research. The reminder of this chapter further motivates the role of online communities in abusive language followed by my contributions in further detail.

NOTE: This thesis is dedicated to studying abusive language and therefore contains examples of such content. Even though I have censored particularly explosive and derogatory expressions, reading such material can still be disturbing. I would like to caution the readers to the presence of harmful language in various sections of this manuscript.

1.1 Community-driven abusive language research

Communities both form, and are formed by, coherent linguistic practices [30] and these practices are an integral part of the community identity itself [125, 189, 208]. This holds true for online spaces as well. For example, the language in a automotive forum such as `Car Talk`[2] is primarily automotive. Thus the language of such dedicated communities represents their identity.

We can therefore define abusive language as that practised within abusive communities. Various communities across the Internet are setup for the sole purpose of degrading and derogating marginalized populations. For example `Stormfront` is a popular online forum for white nationalists where majority of the comments are about the "*evils of African Americans, LGBT people, non-white immigrants, and, above all, Jews*, .."[3] . In such communities, which self-identify as antagonistic, abusive language is not only accepted, it is celebrated and promoted. Thus antagonistic communities aggregate a broad variety of abuse which can promote diversity in abusive language resources.

[2] `www.cartalk.com`

[3] `www.splcenter.org/fighting-hate/extremist-files/group/stormfront`

However, the role of communities is not limited to data collection. By associating language with community identity, we can explicitly account for the social environment in which content is generated and shared. Contextualizing language with the source community can help reduce ambiguity in abusive language. Since extraction of intent is challenging, social context can help clarify the nature of text. Take for example these two comments: " *I am genuinly surpised at a suicidal tr*nny*"; "*Just that the tr*nny is dying on me lol.*". From the surface, it is unclear if one or both of them are derogatory. The first comment was made in a toxic community on Reddit - `CringeAnarchy`, while the latter is from `Honda`. This extra information helps us figure out that the second comment is not derogatory and is referring to the *transmission* of a car.

Community-driven research is therefore a promising avenue for studying abusive language.

1.2 Contributions by Chapter

1.2.1 Chapter 3—A Community-Centric View of Abusive Language

To detect abusive language we need abusive language. However, even this initial collection of abusive content is challenging. Therefore, my thesis begins by searching for answers to a simple question - how can we collect abusive language more effectively.

In this chapter, I present a novel approach that aggregates abuse by leveraging the very communities dedicated to it. I introduce "polar communities" - set of self-identifying online communities that antagonize and support the same marginalized groups. In the context of abusive language, self-identification: (1) removes the onus from researchers to judge the abusive nature, (2) provides a strong intuition about the intent of its active users, and (3) indicates a high density of abusive content. The core analysis in this chapter is setup to assess the viability of self-identifying communities as sources of abusive language.

I collect comments from a variety of polar communities across multiple platforms and build language models that train on these comments. I test the performance of these language models through a series of carefully crafted experiments for a broad analysis. I discover that community-based language models are able to successfully identify antagonistic language against random as well as supportive language. These language models perform well even when tested on similar communities but from different platforms but perform poorly when tested on a communities that antagonize a different marginalized group than the one they were trained on.

These results are promising as they affirm that antagonistic communities incorporate a coherent linguistic signature. They also demonstrate that linguistic norms are shared between antagonistic communities with shared target, irrespective of the platform. Furthermore, antagonistic communities with different target groups have different linguistic norms, which speaks to variability in abusive language based on the marginalized population it degrades. Thus, I establish that antagonistic communities can serve as reliable sources of abusive language. However, general detection of abusive content is going to be challenging due the target dependent nature of abusive language.

Authors' Contributions to the Original Manuscript:

- Haji Mohammad Saleem

  - Design of experiments
  - Data collection
  - Execution of experiments
  - Analysis of results
  - Manuscript writing
- Kelly P Dillon
  - Data collection
  - Setting up keyword baselines
- Susan Benesch
  - Guidance in setting up definitions
  - Editorial guidance
- Derek Ruths
  - Guidance in design of experiments and analyses
  - Editorial guidance

1.2.2 Chapter 4—Community Sampling for an Equitable Corpus of Abuse

This chapter is dedicated to meaningful engagement with the diverse nature of abusive language. To this end, I make three contributions:

First, a taxonomy that allows us to speak more precisely about derogatory language. Based on the principle of derogatory variance, it categories pejorative usage in online conversations as - derogatory, non-derogatory non-appropriative, appropriative, and homonym. These categories are further divided into a total of twelve sub-categories which provide a fine-grained understanding of abusive content.

Second, community sampling - a novel methodology for collecting abusive language that promotes diversity of content. Instead of relying solely on keywords, which is the norm for abusive language research, I collect pejorative language from polar communities. This technique allows me to aggregate diverse perspectives around pejorative use.

Third, a large scale human-annotated corpus of abusive language with diverse content. The corpus design process promotes diversity of opinion. Abusive language is subjective as well as interpretive. Different people have different perspectives on abuse, informed by their lived experience. I attempt to integrate a spectrum of perspectives in the final corpus by assembling a diverse cohort of annotators across ethnicity, gender, and sexuality. Combined with community sampling, this inclusive design process helps ensure an equitable corpus for online abuse.

This taxonomy provides the means for fine-grained analysis of abusive content while the corpus presents itself as a strong benchmark for abusive language research.

Authors' Contributions to the Original Manuscript:

- Haji Mohammad Saleem
  - Taxonomy design
  - Ethics Board approval
  - Annotator recruitment
  - Annotation platform setup
  - Annotator training
  - Data collection and sampling
  - Annotator management and reimbursement

  - Annotation validation
  - Perspective bench-marking
  - Manuscript writing

- Jana Kurrek
  - Taxonomy design
  - Annotator selection
  - Annotation guidelines
  - Result visualization
  - Manuscript writing

- Derek Ruths
  - Guidance in design of experiments and analyses
  - Editorial guidance

1.2.3 Chapter 5—Abusive Language across Slurs and Platforms

In the previous chapter, I construct an abusive language corpus with a novel and inclusive design process. However, the design process encompasses two basic elements: 1) select pejorative keywords for data collection and 2) choose an online platform to collect data from. This basic process of filtering data is the norm in abusive language research and is observed across the vast majority of corpora. However, keyword filtering is popular since it is often the only solution available to researchers for parsing large volumes of social media content.

Despite popularity, we as a research community have limited understanding of how these key corpus design elements affect the end result. Are different abusive language corpora equivalent? What can we expect if we only filter data from toxic platforms? How do keywords affect corpus outcomes? To answer these questions, I perform a broad analysis by first collecting data from three different platforms using nine different pejoratives and then annotating the collected data using the fine grained taxonomy presented in the previous chapter. By aggregating and comparing the fine-grained labels across the different platforms and slurs, I am able to demonstrate the lack of homogeneity across the board.

Not only does the amount of abusive content vary across the design elements, the type of language captured varies as well. I discover that toxic platforms provide a higher density of derogatory comments. Different pejoratives collect different kinds of derogatory as well as non-derogatory language. For example, some are more likely to filter sexualized language while others may filter homonym content. Overall these findings further illustrate the diversity in abusive language and provide broad insights into the critical task of building an abusive language corpus.

Authors' Contributions to the Original Manuscript:

- Haji Mohammad Saleem
  - Design of experiments
  - Data collection
  - Data annotation
  - Analysis of results
  - Visualization of results
  - Manuscript writing
- Derek Ruths
  - Guidance in design of experiments and analyses

- Manuscript writing

1.2.4 Chapter 6—Community Context in Abuse Detection

In this chapter, I tackle the difficult task of abusive language detection and investigate the potential role online communities can play in assisting detection frameworks. As I mention, online communities provide important social context and clarification to conversations that occur within them. For example - "*That park looks so fun. I wish we had tr*nny like that locally.*" could be derogatory but its source `skateboarding` community helps clarify that it is referring to transition skating. I seek to integrate source communities within models for abusive language detection to access the latent contextual information.

By creating rich embeddings of Reddit communities, I discover that polar communities cluster by the nature of their support towards marginalized populations. Antagonistic communities cluster with similar antagonistic communities and supportive subreddits display similar trends. Being in close proximity of established antagonistic or supportive communities can help identify other antagonistic and supportive communities.

To study the direct effect of community context, I build and compare different language models with and without the additional information of the source community. By evaluating their performance across the taxonomy labels, I find that the community context improves overall performance by reducing the false-positive errors. Community context especially assists in the identification of reclamatory language by the members of marginalized populations. Through error analysis I discover that a large portion false-negative comments belonged to supportive communities.

But re-annotating such comments, this time with the community context - I identify them as human errors. These comments were actually not derogatory and were labelled as such by the community-driven model. Thus, community context reduces false-positive errors for humans as well as language models and provides a promising direction for context-aware models in abusive language research.

Authors' Contributions to the Original Manuscript:

- Haji Mohammad Saleem
  - Data collection
  - Design of experiments
  - Creating subreddit embeddings
  - Validating subreddit embeddings
  - Perspective bench-marking
  - Execution of `BERT`-based experiments
  - Analysis of results
  - Annotation for error analysis
  - Manuscript writing
- Jana Kurrek
  - Clustering subreddit embeddings
  - Visualizing subreddit clusters
  - Execution of Logistic Regression based experiments
  - Setting up `BERT`-based experiments
  - Error analysis
  - Manuscript writing

- Derek Ruths
    - Guidance in design of experiments and analyses
    - Editorial guidance

1.2.5 Chapter 7—Banning Antagonistic Communities for Abuse Intervention

The prior chapters focus on detecting abusive content by leveraging antagonistic communities. Having identified such communities, a natural progression of thought is why don't we just ban such spaces. While a ban might seem like a logical next step, it hardly guarantees a favourable outcome. At the outset we do not know how such an action would be received nor what might transpire in its aftermath.

In this Chapter, I present a case study on the fallout from a from large scale platform moderation exercise - 'deplatforming' of an entire antagonistic community. Specifically I study the ban of `r/fatpeoplehate` (`FPH`) and analyse user behavior around it. I compare the engagement of the most active `FPH` users before and after the ban against a cohort of random Reddit users. I document efforts by `FPH` users to circumvent the initial ban. I also identity other similar communities, both supportive and toxic, and examine user activity through the volume of comments posted, downvoted, and removed.

I find that overall the ban led to reduced engagement by `FPH` users and many stopped participating all together. Right after the `FPH` users flooded adjacent communities which required control measures by users and moderators of these communities in the form of downvotes and comment removal respectively. `FPH` users also tried to circumvent their ban by creating almost 100 alternate spaces which required further action by Reddit administrators. Thus the banning of `r/fatpeoplehate` was not a singular action but required a collective effort on the part of Reddit users, moderators and administrators for it to be successful. While this work does not directly contribute to the detection of abusive content, it is still critical to overall abusive language research.

Authors' Contributions to the Original Manuscript:

- Haji Mohammad Saleem
  - Design of experiments
  - Data Collection
  - Execution of experiments
  - Planning of analysis
  - Execution of analysis
  - Result Visualization
  - Manuscript writing
- Derek Ruths
  - Guidance in design of experiments and analyses
  - Editorial guidance

CHAPTER 2
Challenges in Abusive Language Research

Despite accelerated interest, abusive language research still faces considerable challenges ranging from how we conceptualize it, how we build language resources to study it, and how our detection frameworks behave. These challenges severely constrain the development of real-world tools for abusive language detection. In this chapter, we go over some of the challenges that make abusive language detection difficult to operationalize.

2.1 Conceptual Challenges

2.1.1 Lack of an objective definition

Abusive language is loosely defined as expression, promotion, or incitement of hatred against a person or group due to their shared characteristics or group membership [150]. However, there is limited consensus on what is abusive, largely due to the fact that abuse is not an entirely objective phenomenon. The derogatory nature of language depends not only on the content but also upon the prevailing social norms as well as the context in which it is used. Furthermore, it is subject to individual and collective interpretation. Such subjectivity leads to a diffused understanding of abuse in which an individual's intuition overlaps but does not necessarily coincide with someone else's understanding. Therefore, it is difficult to reliably identify abusive language consistently, even for humans.

2.1.2 Variation in content

Abusive language lacks a singular form. It varies drastically based on social group that is being targeted. For example, an expression of misogyny can use very different language than that of Islamophobia. Even within a singular target group, abuse can take multiple forms. For example, the same community can be derogated by a variety of pejorative expressions. Furthermore, targeted abuse is localised and varies across language or geographical borders. Islamophobia in India is dissimilar from Islamophobia in United States. Such variation in content makes it extremely difficult to detect abusive language in a generic manner and research typically has to focus on certain forms. For example, though Davidson et al. [61] and Waseem and Hovy [229] study hate speech, their corpus only contains examples of racism and sexism.

2.1.3 Multiple terminologies

The larger research community uses multiple terms and acts to describe different versions of abusive language. These terms, while referring to similar content, do not completely coincide. For example: Waseem and Hovy [229] label racism and sexism and define hate speech as language that attacks, silences, criticizes, misrepresents, stereotypes, or promotes violence against minorities. Davidson et al. [61] distinguish between hate and offensive speech (the distinction is not provided in the paper). Wulczyn, Thain, and Dixon [237] study personal attacks while Aroyo et al. [7] detect toxicity - defined as “comments that are rude, disrespectful or otherwise likely to make someone leave a discussion”. These wide array of terms are used within the same research umbrella. However they are not well defined and their distinguishing features are unclear.

Not only does abuse research employ a multitude of terms, it also focuses on a wide variety of similar yet distinct harmful behaviour. For example, Banko, MacKeen, and Ray [12] identify doxing, identity attack, identity misrepresentation, insult, sexual aggression, and threat of violence as forms of hate and harassment. However, doxing is classified under cyberbullying in other research [43]. Similarly they characterize “dissemination of negative stereotypes” as identity misrepresentation. However, even the “positive” stereotypes cause harm [202, 44, 119]. Abusive language and harmful behaviour is therefore a complex phenomenon which makes comprehensive research challenging.

2.2 Data Challenges

2.2.1 Datasets lack comprehension

Machine learning solutions are data-driven. They require large representative datasets to produce robust classifiers. As I mention in prior sections, abusive language is highly varied. However, these variations not discrete, i.e., there is no exhaustive set of properties that captures each and every form of abusive language. It is therefore not feasible to create a singular resource that can guarantee exhaustive representation. In most cases, researchers limit themselves to a set amount of targets and language around them. For example, Kurrek, Saleem, and Ruths [130] focus on racism against black people, homophobia, and transphobia; Waseem and Hovy [229] study Islamophobia and sexism.

Even within a type of abuse, a corpus needs to be strategically curated to ensure sufficient variety in both both abusive and non-abusive language. Failing to do so can inadvertently introduce sampling bias in the resulting corpus. Systems trained on biased data can develop confounding associations with certain lexical cues and lead to overfitting. For example, in the Wikipedia Talk corpus [67], 58% of comments that contain the term "gay" are labelled as toxic [245]. This imbalance can lead to trained systems associating the term "gay" with toxicity.

2.2.2 Annotation agreement

Not only is abusive language hard to detect, it is particularly difficult to annotate. First, due to the interpretive nature of the task, not *all* instances of language will have a correct answer. The labelling of some comments can fluctuate based on an annotators personal experience and understanding of abuse. This is reflected in low inter-annotator agreement that frequently reported in abuse research [131, 195]. Second, some comments require clarifying information in order to be correctly assessed and can be mislabelled in absence of appropriate context. Context can include who the speaker is, what the conversation is about, where the conversation is happening, etc. However, not all of this information is readily accessible. For example, user-demographics can only be; some comments can be TV and movie quotes. Overall the annotation of corpus is a challenging and error-prone. Noticeably, analysis suggests that up to 10% of Davidson et al. [61] corpus might be mislabelled [3]. Despite the inconsistencies, abusive language research lacks a unanimous set of best practices around data annotation. In most studies, details of annotation process or annotation guidelines are not even provided.

2.3 Detection challenges

2.3.1 Context integration

Even though abusive language derives meaning from both content and context, the majority of research relies completely on linguistic cues for abuse detection. Algorithms are trained solely on the text of online comments devoid of any and all context. The resulting systems lack situational awareness and are unable to make nuanced decisions [102]. Some recent work has started to explore different facets of contextual information with promising results [220, 174]. However, contextual modelling of abusive language is challenging. There is a lack of training datasets that also contain contextual information. Most datasets only provide online comments with human labels. Furthermore, the research community lacks broad empirical analysis on the different aspects of context and their individual as well as collective impact.

2.3.2 Linguistic complexities

User-generated text is generally unstructured. It can contain grammatical, typographical, and spelling errors since such content is not necessarily proofread. However there are other linguistic characteristics which can make the abuse detection even more challenging. Detection of sarcasm in text is a hard task. When language is cloaked in sarcasm or humor, it makes discerning the true intent difficult. Same is true for abusive language, where sarcasm and humor can lead to classification errors [167, 3]. Furthermore, authors can deliberately introduce spelling variations or codes to obfuscate certain words. For example slurs were replaced with words like Skype, Google, and Bing [139, 140]. Such obfuscation attempts introduce 'out of vocabulary' terms which increase classification errors. Homonyms can also create challenges. Some slurs and pejoratives have more than one meaning. For example *tr*nny* is often used in the automotive communities as a short form for transmission. Algorithms need to be provided with enough instances of all possible forms to be able to successfully distinguish abusive uses.

CHAPTER 3
A Community-Centric View of Abusive Language

A variety of online platforms promote discourse. However, anonymity and a low barrier to entry have provided optimal conditions for abusive content to propagate and flourish through such platforms. Abusive content has extreme effects both online and offline. It is therefore critical to curate social media platforms and detect abusive content. However, due to the high volume of content being generated, curation and moderation of abusive language requires algorithmic assistance.

One of the primary challenges to building models for abusive language detection is access to large scale training data. However, due to the variable nature of abusive language, the process of creating such training resources is non-trivial. In this chapter I explore how do we gather abusive language that will allow us to train robust models to detect abusive language?

As I mention in the Introduction, abusive language is produced in a spectrum of online communities. While some of these communities are intolerant of abuse and seek to moderate it, others support and even encourage it. Such communities serve as a congregation of abusive users who seek to degrade and antagonize others. Abusive language can therefore be described as native to such antagonistic communities.

In the following manuscript, I first establish “polar communities” - online communities that antagonize and support the same marginalized population, and compare their linguistic overlap. I leverage these communities to collect data and build language models. I test these models across polarity (supportive content vs abusive content), platform (abusive content from multiple social platforms) and target(out of domain abusive language).

3.1 Manuscript 1: A Web of Hate: Tackling Hateful Speech in Online Social Spaces

Authors: Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths

Published in the Proceedings of the first workshop on Text Analytics for Cybersecurity and Online Safety TA-COS @ LREC 2016.

3.1.1 Abstract

Online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, intimidate, or silence platform users, and some of it can inspire other users to commit violence. Despite widespread recognition of the problems posed by such content, reliable solutions even for detecting hateful speech are lacking. In the present work, we establish why keyword-based methods are insufficient for detection. We then propose an approach to detecting hateful speech that uses content produced by self-identifying hateful communities as training data. Our approach bypasses the expensive annotation process often required to train keyword systems and performs well across several established platforms, making substantial improvements over current state-of-the-art approaches.

3.1.2 Introduction

Online spaces are often exploited and misused to spread content that can be degrading, abusive, or otherwise harmful to people. An important and elusive form of such language is *hateful speech*: content that expresses hatred of a group in society.

Hateful speech has become a major problem for every kind of online platform where user-generated content appears: from the comment sections of news websites to real-time chat sessions in immersive games. Such content can alienate users and can also support radicalization and incite violence [4]. Platform operators recognize that hateful content poses both practical and ethical issues and many, including Twitter, Facebook, Reddit, and gaming companies such as Riot Games, have tried to discourage it, by altered their platforms or policies.

Yet reliable solutions for online hateful speech are lacking. Currently, platforms predominantly rely on users to report objectionable content. This requires labor-intensive review by platform staff and can also entirely miss hateful or harmful speech that is not reported. With the high volume of content being generated on major platforms, an accurate automated method might be a useful step towards diminishing the effects of hateful speech.

Without exception, state-of-the-art computational approaches rely upon either human annotation or manually curated lists of offensive terms to train classifiers [131, 213]. Recent work has shown that human annotators tasked with labeling hate speech have significant difficulty achieving reasonable inter-coder reliability [131]. Within industry, it is generally acknowledged that keyword lists are also insufficient for accurate detection of hateful speech. However, little work has been done to understand the nature of their limitations and to design able alternative approaches. This is the topic of the present work.

This paper makes three key contributions. First, we establish why the problem of hateful speech detection is difficult, identifying factors that lead to the poor performance of keyword-based approaches. Second, we propose a new approach to hateful speech detection, leveraging online communities as a source of language models. Third, we show that such a model can perform well both within a platform and *across* platforms — a feature we believe we are the first to achieve.

We are also aware that automated detection of online speech could be misused to suppress constructive and/or dissenting voices by directing the system at individuals or groups that are not dedicated to expressing hatred. Such a use would be antithetical to our intent, which is to explore and illustrate ways in which computational techniques can provide opportunities to observe and contain harmful content online, without impinging on the freedom to speak openly, and even to express unpalatable or unpopular views. We hope that our work can help diminish hatred and harm online. Furthermore, since our method can be trained on and applied to a wide array of online platforms, this work may help to inform the direction of future research in this area.

3.1.3 Background

Hate and hateful speech

Legal and academic literature generally defines hate speech as speech (or any form of expression) that expresses (or seeks to promote, or has the capacity to increase) hatred against a person or group of people because of a characteristic they share, or a group to which they belong [150]. There is no consensus definition, however. Definitions of this sort are problematic for a number of reasons [13], including that hate speech is defined by prevailing social norms, context, and individual and collective interpretation. This makes it difficult to identify hate speech consistently and yields the paradox (also observed with pornography) that each person seems to have an intuition for what hate speech is, but rarely are two people's understandings the same. This claim is affirmed by a recent study that demonstrated a mere 33% agreement between coders from different races, when tasked to identify racist tweets [131].

A particular ambiguity in the term 'hate speech' is in "hate" itself. That word might refer to the speaker/author's hatred, or his/her desire to make the targets of the speech feel hated, or desire to make others hate the target(s), or the apparent capacity of the speech to increase hatred. Needless to say, we require a rigorous — and formal — definition of a type of speech if we are to automate its detection.

Our initial motivation was to find, and work with, a notion of hate speech that can be operationalised. The work of online platform operators (e.g., Twitter, Facebook, and Reddit) helped to focus this aim. Their concern over the capacity of language to do harm — whether emotional, mental, or physical — logically focuses more on what is *expressed* rather than how it is *intended*. Whereas "hate speech" can imply an inquiry or judgment about intent (e.g. what was this person feeling or wishing?), we propose the term "hateful speech" to focus on the expression of hate — a nuanced, but useful distinction since expression is easier to detect than intent, and more likely to be linked to language's capacity to cause harm.

This leads to our term *hateful speech*: speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity.

Hateful speech is not to be mistaken for "cyber-bullying", another form of troubling online content that has been widely discussed and studied in recent literature.

Cyber-bullying is repetitive, intentional, aggressive behavior against an individual, and it either creates or maintains a power imbalance between aggressor and target [214]. It is often hateful but it does not necessarily denigrate a person based on his or her membership in a particular group, as hateful speech (the subject of the present work) does.

Community-defined speech

As we will discuss in detail later, we use the language that emerges from self-organized communities (in Reddit and elsewhere) as the basis for our models of hateful speech. Our decision is based on a deep sociological literature that acknowledges that communities both form, and are formed by, coherent linguistic practices [30]. Most groups are defined in part by the "relationships between language choice and rules of social appropriateness" forming speech communities [94]. In this way of thinking, the group is defined by the speech and the speech comes to define the group [125, 189, 208, 207].

In the context of this study, this means that hate groups and the hateful speech they deploy towards their target community cannot exist without one another, especially online. Therefore, taking the linguistic attributes particular to a community committed to degrading a specific group is a legitimate and principled way of defining a particular form of hateful speech. To our knowledge, this work represents the first effort to explicitly leverage a community-based classification of hateful language.

Existing approaches to detecting hateful speech

Despite widespread concern about hateful speech online, to our knowledge there have been only three distinct lines of work on the problem of automated detection of hateful speech. One study concerned the detection of racism using a Naive Bayes classifier [131]. This work established the definitional challenge of hate speech by showing annotators could agree only 33% of the time on texts purported to contain hate speech. Another considered the problem of detecting anti-Semitic comments in Yahoo news groups using support vector machines [227]. Notably, the training data for this classifier was hand-coded. As we will discuss in this paper, manually annotated training data admits the potential for hard-to-trace bias in the speech ultimately detected. A third study used a linguistic rule-based approach on tweets that had been collected using offensive keywords [239]. Like manually annotated data, keyword-based data has significant biasing effects as well.

In this work we aim to build on these studies in two ways. First, we will consider a definition of hateful speech that could be practically useful to platform operators. Second, we will develop a general method for the detection of hateful speech that does not depend on manually annotated or keyword-collected data.

Reddit and other online sources of hateful speech

Reddit is currently one of the most actively used social content aggregation platforms. It is used for entertainment, news and social discussions. Registered users can post and comment on content in relevant community discussion spaces called *subreddits.* While the vast majority of content that passes through Reddit is civil, multiple subreddits have emerged with the explicit purpose of posting and sharing hateful content, for example, `r/CoonTown`, `r/FatPeopleHate`, `r/beatingwomen`; all which have been recently banned under Reddit's user-harassment policy [159]. There are also subreddits dedicated to supporting communities that are the targets of hate speech.

Reddit is an attractive test-bed for work on hateful speech both because the community spaces are well-defined (i.e., they have names, complete histories of threaded discussions) and because, until recently, Reddit has been a major online home for both hateful speech communities and supporters for their target groups. For these reasons, throughout this paper, our analyses heavily leverage data from Reddit groups.

Of course, Reddit is not the sole platform for hateful speech. Voat, a recently created competitor to Reddit, along with a vibrant ecosystem of other social content aggregation platforms, provide online spaces for topical discussion communities, hate groups among them. Furthermore, dedicated websites and social networking sites such as Twitter and Facebook are also reservoirs of easily accessible hateful speech.

Important research has investigated the effects of racist speech [161] and sexual harassment [80] in online games. Notably, in this study we have not worked with data from online gaming platforms, primarily because the platforms are generally closed to conventional data collection methods.

3.1.4 The limits of keyword-based approaches

In the same way that hateful groups have defining speech patterns, communities that consist of the targets of hateful speech also have characteristic language conventions. We will loosely call these *support groups.* Notably, support groups and the groups that espouse hateful speech about them often engage in discourse on similar topics, albeit with very different intent. Fat-shaming groups and plus-size communities both discuss issues associated with high BMI, and women and misogynists both discuss gender equity. This topical overlap can create opportunities for shared vocabulary that may confuse classifiers.

In addition, many keyword-based approaches select established and widely known slurs and offensive terms that are used to target specific groups. While such keywords will certainly catch some hateful speech, it is common to express hate in less explicit terms, without resorting to standard slurs and other offensive terms.

For example, hateful speakers refer to migrants and refugees as "parasites" and call African-Americans "animals." While neither of these terms are inherently hateful, in context they strongly denigrate the group to which each term is applied.

We can expect that classifiers trained on overtly hateful keywords will miss such posts that use more nuanced or context-dependent ways of achieving hateful speech.

Furthermore, keywords can be also be obscured through misspellings, character substitutions (by using symbols as letters), using homophones etc. These practices are commonly employed to circumvent keyword-based filters on online platforms [227].

In this section, we study the potential impact of topic overlap on data returned by keyword-based queries (we will consider under-sampling issues in the next section). Here our focus will be on the sample that keyword-based filters return and in later sections we will consider the performance of classifiers built from such samples.

Target Group	**Hate Subreddit**	**Comments**	**Support Subreddit**	**Comments**
Black	CoonTown	350851	racism	9778
Plus	fatpeoplehate	1577681	loseit	658515
Female	TheRedPill	51504	TwoXChromosomes	66390

Table 3–1: Public comments collected from hate and support subreddits on Reddit, for three target groups.

Data

Recently, Reddit user, Stuck_In_the_Matrix[1] , made available large data dumps that contain a majority of the content (posts and comments) generated on Reddit[2] . The data dumps, collected using the Reddit API, are organized by month and year. The data date back to 2006 and are regularly updated with new content. We use all comments from January 2006 through January 31, 2016 and expanded the dataset with each update. Each file corresponds to a month of Reddit data, and every line is a json object of a Reddit comment or post.

For our analysis, we identify three commonly targeted groups on Reddit — African-American (black), plus-sized (plus) and women. For each of the target groups, we select the most active support and hate subreddits. To create our datasets, we extract all user comments in the selected subreddits from the data dumps described above, in October 2015. The details on the selected subreddits and the number of the extracted comments are provided in Table 3–1.

[1] www.reddit.com/user/Stuck_In_the_Matrix/

[2] couch.whatbox.ca:36975/reddit/

Methods

For each of the selected subreddits, we use labeled Latent Dirichlet Allocation (LLDA) to learn the topics that characterize them, against a baseline Reddit language. This baseline is intended to push the LLDA to remove non-topical vocabulary from the two subreddit topics; it consists of a sample of 460,000 comments taken at random from the Reddit data scrape (none of the posts belonged to any of the subreddits of interest). Prior to topic modeling, stop words, punctuation, URLs, and digits were stripped from the comments and for the purpose of balanced analysis, an equal number of comments was selected from the subreddit and the random sample. We use JGibbLDA for the topic inference [183].

Results

In Table 3–2, we present the 15 most topical words from each subreddit. The top terms in the topics are consistent with the target/support communities. For example, the term "women" was ranked highly in subreddits that concern women (whether positively or negatively referenced) and "weight" is the highest ranked topic for subreddits discussing plus-sized individuals and lifestyle.

We observe a substantial overlap in vocabulary of hate and support subreddits, across all three target communities (see bold words in Table 3–2). While in the case of a black target group, we observe a Jaccard Index (*JI*) of 0.28, the overlap is higher in the case of female targets with *JI* at 0.50 and much higher for plus-size targets, with a *JI* of 0.76.

The implication of this shared vocabulary is that while keywords can be used to detect text relevant to the target, they are not optimal for detecting targeted hateful

Black		**Plus-size**		**Female**	
CoonTown	racism	FPH	loseit	TRP	TwoXCr
nigger	**white**	**weight**	**weight**	**women**	**time**
white	racism	**calorie**	**calorie**	**girl**	**women**
black	**black**	**time**	**time**	**time**	**feel**
shit	**racist**	**work**	**food**	woman	**work**
time	**race**	**food**	**eating**	**shit**	**year**
fucking	**time**	**feel**	**week**	**work**	**fuck**
fuck	person	**eating**	**work**	**year**	**shit**
race	point	**week**	**feel**	**life**	weight
year	feel	**lose**	**lose**	**fuck**	**fucking**
hate	comment	**year**	**diet**	guy	person
racist	american	women	**body**	point	**life**
live	post	**diet**	exercise	friend	**girl**
work	issue	**body**	**goal**	post	love
jew	asian	start	loss	**feel**	pretty
crime	color	**goal**	**year**	**fucking**	food
Jaccard Index: 0.28		JI: 0.76		JI: 0.50	

Table 3–2: Top discovered topics from support and hate subreddits for the three targets. The bold terms signify those that are present in both the hate and support vocabulary.(FPH: fatpeoplehate, TRP: TheRedPill, TwoXCr: TwoXChromosomes)

speech. Shared vocabulary increases the likelihood of tagging content that is related to the target but not necessarily hateful, as hateful and increases false positives. We therefore require more robust training data.

3.1.5 A community-driven model of hateful speech

A key objective of our research is to avoid the issues associated with using manual annotation and keyword searches to produce training data for a classifier. As noted previously, sociological literature acknowledges that communities are formed by coherent linguistic practices and are defined, in part, by their linguistic identity

[94]. Thus, the opportunity considered here is to leverage the linguistic practices of specific online communities to empirically define a particular kind of hateful speech.

Since linguistic practices coincide with the identity of a community using them, we can define hateful speech as discourse practiced by communities who self-identify as hateful towards a target group. The members of the community contribute to the denigration of the target and, therefore, share a common linguistic identity. This allows us to develop a language model of hateful speech directly from the linguistic conventions of that community without requiring manual annotation of specific passages or keyword-based searches. This approach has a number of advantages over these practices.

First, a community-based definition removes the interpretive challenge involved in manual annotation. Membership in a self-organized community that is committed to denigration of a target group through the hatred of others is an observable attribute we can use to surface hateful speech events.

Second, unlike prior work, our method does not require a keyword list. We identify communities that conform to the linguistic identity of a self-organized hateful groups and use such communities to collect data. This data is used to learn the language model around the linguistic identity for detection. This removes any biases implicit in the construction of a keyword list (i.e., in the words included in or excluded from the list).

Third, a community-based definition provides a large volume of high quality, current, labeled data for training and then subsequent testing of classifiers. Such large datasets have traditionally been difficult to collect due to dependence on either manual annotation (annotation is slow and costly) or keyword searches (stringent keywords may turn up relatively few hits).

This approach generalizes to other online environments (such as Voat and other hateful speech-focused web forums) in which communities declare their identities, intentions, and organize their discussions. Any online (or, even, offline) communication forum in which all participants gather for the understood purpose of degrading a target group constitutes a valid source of training data.

In the following subsections, this approach is validated through three analyses. First, we demonstrate that the hate speech communities identified actually employ distinct linguistic practices: we show that our method can reliably distinguish content of a hateful speech community from the rest of Reddit. We also show that our approach substantially outperforms systems built on data collected through keywords.

Second, we show that our approach is sensitive to the linguistic differences between the language of hateful and support communities. This task is notably difficult given the results we reported above, showing that such communities share many high-frequency words.

Finally, we use our Reddit-trained classifier to detect hateful speech on other (non-Reddit) platforms: on Voat and hateful speech web forums (websites devoted to discussion threads attacking or denigrating a target community). For both, we find that our method performs better than a keyword-based baseline.

Data collection

Reddit: We use Reddit as the primary source for the hateful communities and leverage the linguistic practices of these communities to empirically define and develop language models for target-specific hateful speech. In all three of our studies, we focus on the aforementioned three target groups: black people, plus-sized individuals, and women. For each, we select the most active hateful and support subreddits and collect all the publicly available comments present in the data dumps provided by `Stuck_In_the_Matrix`. The details on the dataset are provided in Table 3–1. We also collect a random sample of 460,000 Reddit comments to serve as negative examples.
Voat: Voat, a content aggregator similar to Reddit, also hosts active discussion communities, called *subverses*, few of which identify as hateful. We select Voat because of its similarity to our original source[3] . Since the two websites cater to a similar userbase, the generated linguistic identities should be similar in sub-communities with similar themes. Therefore, the language model of hateful communities on Reddit should match, to an extent, with the language model of similar hateful communities on Voat.

For the target groups, we identify hateful subverses - `v/CoonTown`, `v/TheRedPill` and `v/fatpeoplehate` - sub-communities that share their name with their counterparts on Reddit and target blacks, plus-size individuals, and women, respectively. In the absence of an API, we use web-scraping libraries to retrieve all publicly available

[3] `thenextweb.com/insider/2015/07/09/what-is-voat-the-site-reddit-users-are-flocking-to/`

Target	Subverse	Comments	Web forum	Comments
Black	`CoonTown`	3358	`shitskin`	3160
Plus	`fatpeoplehate`	31717	-	
Female	`TheRedPill`	478	`mgtowhq`	20688

Table 3–3: Target-relevant hateful comments collected from Voat subverses and web forums.

comments posted to the selected subverses between July 2015 and January 2016. We also collect a set of 50,000 comments (from the same time period) from a random sample of subverses to serve as negative examples (Table 3–3).

Web forums: We also use stand-alone web forums that are dedicated to expressing hate or contempt for the target communities. These web forums are social platforms that provide their users with discussion boards, where users can create threads under predefined topics and other users can then add comments in these threads. We, therefore, select web forums for their discussion-based communities and user-generated content. Again, due to the lack of APIs, we use, as data, comments that were collected by web-scraping libraries from numerous threads of their discussion boards during October 2015.

For the black target group, we use `Shitskin.com`: our dataset consists of 3,160 comments posted to 558 threads from three of website's boards: "Primal Instinct", "Crackin the whip!" and "Underground Railroad." For the female target group, we use `mgtowhq.com`: this dataset consists of 20688 comments posted to 4,597 threads from the "MGTOW General Discussion" board. Finally, as a source of negative examples, we use the "random" discussion board on `topix.com`: this dataset consists of nearly 21,000 comments from 2458 threads. To our knowledge, no large fat-shaming forum exists, thus we do not include this target group in this phase of the study (Table 3–3). All comments have posting times between July 2015 and January 2016.

Methods

Before the classification process, we preprocess all the data by eliminating URLs, stopwords, numerals and punctuation. We further lowercase the text and remove platform-relevant noise (e.g., comments from house keeping bots on Reddit like AutoModerator). The text is finally tokenized and used as input for the classification pipeline.

We use multiple machine learning algorithms to generate the language models of hateful communities. From the analysis of the prior work, we identify the commonly-used algorithms and employ them in our analysis. Specifically, we use naive Bayes (NB), support vector machines (SVM) and logistic regression (LR). We do this in order to assess the merits of our insight into using community-defined data collection.

The algorithms take as input, tokenized and preprocessed arrays of user comments along with the label of the community they belong to. We use a sparse representation of unigrams with *tfidf* weights as our feature set. In future investigation, we would like to add part of speech tags and sentiment score as features.

For performance evaluation, we use the standard measures: accuracy, precision, recall and F1-Score. We also use Cohen's κ as a measure of agreement between the observed and expected labels. κ helps in evaluating the prediction performance of classifiers by taking in account any chance agreement between the labels.

Baseline comparison Our aim is to assess the impact of using community-based text compared with keyword-based text as training data. Due to space limitations, here we report only a logistic regression classifier trained on keyword-collected data (SVM and NB showed comparable performance).

The specific keywords used are generated from the comments collected from hateful Reddit communities. For a given target group, we generate three sets of keywords for each: (1) keywords generated between hate subreddits and a random sample of Reddit comments using LLDA, as in Section 3, (2) keywords generated between hate subreddits and a random sample of Reddit comments using χ^2 weights (χ^2I), and (3) keywords generated between hate and support subreddits using χ^2 weights (χ^2II). To generate the training datasets, we use the top 30 keywords and from a separate random sample of Reddit comments, collect samples that contain at least one of the keywords as positive samples and samples that contain no keywords as negative samples. For each keyword type and each target, we aggregate 50,000 positive and 50,000 negative samples for training.

Results and Discussion

Community language vs. hateful speech. It may seem that, by comparing classifiers on the task of detecting hateful community posts, we are equating language produced by a hateful community with hateful language. Certainly, they are not always the same. Some content is likely non-hateful chatter. One alternative for excluding such noise is manual coding of testing data. Given the existing issues with such labeled data, we avoid such manual labeling. Furthermore, a comparison of the two approaches is not fair due to the associated trade-offs. The community definition, as mentioned, relies on the assumption that all the content in a hateful community is hateful, which might not always be true. However, such an assumption allows us to generate large training datasets with relative ease. We therefore allow the presence of some noise in the training data for ease of training data generation and favouring recall. On the other hand, manual annotation promises less noisy datasets at the expense of time and resources, which limits the size of training datasets. It would be very laborious to produce datasets as large as those generated with our community approach. Also, since manual annotation relies heavily on personal perception, it can also introduce noise in the datasets. In other words, manual annotation does not allow us to generate large training sets, and also cannot provide completely noise-free data.

Another option, however, is to focus on the precision ($\frac{TP}{TP+FP}$) of the classifier. Precision indicates the classifier's ability to identify only content from the hateful community. The construction of the test datasets is such that hateful speech should only exist in the hateful community posts. Thus, a method that detects hateful content should strongly favor including only content from hateful communities — yielding high precision. Crucially, in the discussions that follow, we find that a community-based classifier demonstrates much higher precision than keyword-based methods. Thus, by either measure (F1 or precision), our community-based classifier outperforms the baselines.

Hateful groups have distinct linguistic signatures. In Table 3–4(a), we see the performance of the three classifiers when classifying a balanced corpus of hateful posts and randomly selected (non-hateful speech) Reddit posts with 10-fold cross validation. The dataset consists of all the comments collected from the relevant hate subreddit (Table 3–1) as positive samples and an equal number of random Reddit

(a) Assessing the distinct nature of language emerging from hate groups.

	Accuracy			**Precision**			**Recall**		
	NB	SVM	LR	NB	SVM	LR	NB	SVM	LR
Black	0.79	0.81	0.81	0.78	0.84	0.87	0.82	0.74	0.73
Plus	0.78	0.78	0.79	0.78	0.81	0.82	0.79	0.75	0.73
Female	0.77	0.80	0.81	0.71	0.81	0.84	0.90	0.77	0.75

	F1-Score			**Cohen's κ**		
	NB	SVM	LR	NB	SVM	LR
Black	0.80	0.79	0.79	0.58	0.61	0.61
Plus	0.78	0.78	0.77	0.56	0.57	0.57
Female	0.79	0.79	0.79	0.55	0.60	0.61

(b) Assessing sensitivity between the language of hate and support groups.

	Accuracy			**Precision**			**Recall**		
	NB	SVM	LR	NB	SVM	LR	NB	SVM	LR
Black	0.80	0.79	0.79	0.80	0.80	0.78	0.85	0.82	0.86
Plus	0.83	0.85	0.85	0.85	0.84	0.84	0.79	0.86	0.86
Female	0.79	0.78	0.78	0.78	0.79	0.80	0.79	0.77	0.77

	F1-Score			**Cohen's κ**		
	NB	SVM	LR	NB	SVM	LR
Black	0.82	0.81	0.82	0.57	0.56	0.55
Plus	0.82	0.85	0.85	0.66	0.69	0.70
Female	0.79	0.78	0.78	0.57	0.56	0.57

Table 3–4: The performance of the three classification algorithms across the three target groups, with a 10 fold cross-validation. (a) Hateful comments are classified against random comments. (b) Hateful comments are classified against comments from support communities. In both cases, the classifier is able to distinguish hate speech from negative cases. (NB: Naive Bayes, SVM: Support Vector Machines, LR: Logistic Regression)

comments as negative samples. We observe the three classifiers perform almost identically. Naive Bayes slightly outperforms others on Recall and F1-score, while Logistic Regression is a slightly better performer on the other metrics. Also, the performance of the classifiers is consistent across the three target groups. Analysis of κ suggests that observed labels after the classification process are in moderate to substantial agreement with the expected labels.

Community-based approach is sensitive to the linguistic differences of hate and support communities. In Section 3.1.4, we showed that hateful and support communities for a target group have a shared vocabulary: the two communities often engage in discourse on similar topics, albeit with quite different intent. Since the shared keywords are not effective in the discrimination process, recognizing the distinction between hate and support communities can be challenging. We set up a classification task for identifying comments from support and hate communities, carried out with a 10-fold cross-validation. The performance of the task is presented in Table 3–4(b). We observe that this performance is close to the performance of our system against a random collection of Reddit comments (Table 3–4(a)). Therefore, even with shared vocabulary, our system is sensitive to the distinction in linguistic characteristics of hateful and support communities for the same target.

Comparison to baseline. In all cases considered, a classifier trained on community-based data outperforms a keyword-based classifier (Table 3–5). Notably, the keyword-based classifier for the women-target group performed best, suggesting that hateful community language associated with the keywords used for collection are more representative of hateful speech (compared to other communities).

From a precision perspective, we find that the community-based classifier outperforms the baselines by between 10% and 20%, indicating that the community-based classifier is including far fewer incorrect cases of hateful speech (false positives). When we look at the true positive posts that have been detected exclusively by the community-based classifier (i.e., that the keyword-based approach missed), we find many that are clearly hateful, but in ways that do not use specialized slurs. Several examples from `r/CoonTown`:

1. "*I don't see the problem here. Animals attack other animals all the time.*"
2. "*Oy vey my grandparents vuz gassed ven dey vaz six years old!*"
3. "*DNA is rayciss, or didn't you know?*"
4. "*Are they going to burn their own town again? Yawn.*"

These examples characterize different (and important) ways in which speech can be hateful without using words that typically operate, largely independent of context, as slurs. In Example 1, African-Americans are described as animals, employing a word that is not usually a slur, to denigrate them. In Example 2, historical

(a) Baseline performance over Reddit data.

Target	**Accuracy**			**Precision**			**Recall**			**F1-Score**			**Cohen's κ**		
	LDA	χ^2I	χ^2II	LDA	χ^2I	χ^2II	LDA	χ^2I	χ^2II	LDA	χ^2I	χ^2II	LDA	χ^2I	χ^2II
Black	0.59	0.63	0.57	0.61	0.71	0.62	0.52	0.44	0.40	0.56	0.54	0.48	0.18	0.26	0.15
Plus	0.53	0.57	0.53	0.54	0.60	0.55	0.35	0.40	0.34	0.42	0.48	0.42	0.06	0.14	0.06
Female	0.68	0.70	0.70	0.65	0.69	0.74	0.71	0.71	0.60	0.68	0.70	0.66	0.35	0.40	0.40

(b) Baseline performance over Voat data.

Black	0.62	0.63	0.62	0.65	0.73	0.68	0.48	0.40	0.40	0.55	0.51	0.51	0.24	0.26	0.23
Plus	0.56	0.60	0.57	0.58	0.65	0.61	0.35	0.40	0.36	0.43	0.50	0.45	0.11	0.20	0.14
Female	0.67	0.69	0.67	0.68	0.71	0.74	0.63	0.63	0.50	0.65	0.67	0.60	0.35	0.38	0.34

(c) Baseline performance over web forum data.

Black	0.66	0.62	0.57	0.72	0.77	0.67	0.53	0.35	0.31	0.61	0.48	0.42	0.32	0.24	0.15
Female	0.78	0.79	0.77	0.81	0.83	0.87	0.75	0.74	0.64	0.78	0.78	0.74	0.56	0.58	0.54

Table 3–5: Baseline performance on multiple platforms with three keyword-generating methods: LDA, χ^2I and χ^2II. Classification was done using Logistic Regression.

Target	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Cohen's κ**
Voat					
Black	0.82	0.87	0.74	0.80	0.64
Plus	0.81	0.85	0.74	0.79	0.62
Female	0.74	0.76	0.71	0.73	0.49
Web forum					
Black	0.82	0.87	0.77	0.82	0.65
Female	0.77	0.83	0.69	0.75	0.54

Table 3–6: Performance of language models trained on Reddit communities and tested on data from Voat and web forums.

context (the gas chambers in Nazi concentration camps), culturally stereotyped language ("*Oy vey*"), and spelling to imitate an accent ("*ven dey vaz*") are successfully used to express contempt and hatred, without any slur or even any word that, like "*animals*" in the first example, is sometimes pressed into service as a slur. The third example, like the second, parodies an accent, and here it is notable that while "racist" might be a keyword use for collection, it's unlikely that "*rayciss*" would be used. Finally Example 4 achieves its effect by attacking a group through an implication of stereotyped action without even actually naming them at all (as opposed to Example 1, in which the targets were called "*animals*").

Community-trained systems can be deployed on other platforms. Often training data for hateful language classification can be hard to obtain on specific platforms. For this reason, methods that work across platforms (trained on one platform, applied on another platform) present significant advantages.

For the analysis, we continue with the same three target groups and train our language model, using logistic regression, with comments from relevant Reddit communities and then test it on data we collected from other platforms. The performance

Training	Testing	Acc	Precision	Recall	F1	κ
CoonTown	fatpeoplehate	0.58	0.72	0.26	0.38	0.15
CoonTown	TheRedPill	0.55	0.6	0.22	0.32	0.08
fatpeoplehate	TheRedPill	0.58	0.65	0.3	0.41	0.15
fatpeoplehate	CoonTown	0.54	0.61	0.23	0.34	0.08
TheRedPill	CoonTown	0.51	0.53	0.28	0.36	0.03
TheRedPill	fatpeoplehate	0.60	0.65	0.41	0.51	0.19

Table 3–7: Cross target performance of classification systems trained on data that belongs to a target community different than the one tested on.

of the system, (Table 3–6), is very similar to the results we obtain when testing on Reddit (Table 3–4(a)). This said, we must be careful not to overstate our method's generalizability. While, certainly, the degree of generalizability observed is noteworthy (particularly given past work), these platforms all feature similar posting conventions: posts are not length restricted, are made within well defined discussion threads, and have a clear textual context. Our method will likely perform well on any such forum-based system. Platforms, which involve quite different conventions, particularly those that are predominantly populated by short-text posts (e.g., Twitter and Facebook), will likely involve additional work. Nonetheless, we do believe that the community-based approach presents opportunities for these other platforms as well.

Hateful classifiers are not target-independent. Hateful conversations are thematic and major topics discovered from conversations are target related (Table 3–2). Not surprisingly, our system performs poorly when tested across targets. We train the classifier on one target and test it on another. The results (see Table 3–7) provide a strong indication that hateful speech classification systems require target-relevant training.

Imbalanced Datasets. We use balanced datasets for our analysis. Since this assumption may or may not hold for different data sources, we perform some initial analysis on imbalanced datasets. As the actual composition of data sources can be variable, we generate testing sets with the ratio of hateful content to non-hateful content at 1:10, 1:100, 1:1000. Our preliminary results are similar to the performance on a balanced test set. These results are encouraging but require further analysis. We hope to overcome the challenges of dataset-shift due to mismatch in the composition of testing and training datasets in future work.

Detailed Error Analysis

In order to better understand the performance of our system, we manually inspect a set of erroneously classified posts from the `r/CoonTown` training/testing dataset. We characterize the kinds of issues we observe and discuss them here.

Type I errors. These posts arise when non-hate group posts are labeled as hate-group posts. Notably, we observe that some of these errors are actually racist comments that originated from other communities in Reddit.

1. "*well jeez if u pit a n*gger against a cunt what do u expect*"
2. "*Triskaid is a fucking n*gger.*"

In both of the cases the comments were in fact racist and were therefore correctly labeled. This, of course, points out a potential (though, we would argue minor) weakness of our approach, which is that hate groups are not the *only* source of hateful language — simply the most high-density source.

More frequently, Type I errors featured non-racist comments which had been mislabeled. This is likely due to the fact that not all content in a hateful community is hateful: some is simply off-topic banter among community members. This adds noise during the training phase which manifests as classification errors. While certainly an issue, given the dramatic improvement in overall classification performance, we consider this an acceptable trade off at this stage in the research. Future work should consider ways of focusing training data further on the distinctly hateful content produced by these communities.

Type II errors. In most cases where hateful-speech community posts were incorrectly labeled as non-hateful, we primarily find that these were, in fact, non-racist posts that were made to the hateful subreddit. Here are a few examples:

1. "*and you're a pale virgin with a vitamin d deficiency.*"
2. "*Whats the deal with you 2? And besides, we're all on the same side here..*"
3. "*IP bans do literally nothing, it only takes a moment to change it.*"
4. "*I can't believe Digg is still up. I can't believe Reddit is still up.*"

Posts like these constitute noise, in terms of our community-based definition of hateful speech, discussed above. Nonetheless, our system was able to correctly identify them as non-hateful. Taken together with the Type I errors, it appears that the noise implicit in our community-definition of hateful speech yields a modest increase in Type I error, but can somewhat be removed by the classifier in the form of Type II errors (which are not, in fact, errors).

A very small number of other Type II errors are examples of hateful speech, but that target a community other than blacks (in the cases we saw, primarily Jewish people):

1. "*Peace and harmony? Yeah that's why they stole that land (now k*keriel) and killed the civilians that lived there before. Did I mention they STILL kill the Palestinians to this day and cover it up? Fuck them.*"
2. "*quit kissing k*keass*'
3. "*You sound like a jew. In a system ruled by money, money can buy anything. Everything is capitalisms fault. But I get why you'd support capitalism since your "people" invented the whole shebang*"
4. "*Losing weight isn't even hard, stop eating like a fucking landwhale, drink lots of water and move your fatass*"

Although these comments are hateful, since they are not directed at black people, the system is technically performing according to specification.

Our system missed some cases of obvious racism, such as the following examples. However, such cases constitute only a small fraction of the comments in Type II error.

1. "*Ok Korea - you know your duty in the impending 'blackification' of the globe? I know where I stand*"
2. "*Black people are terrible.*"
3. "*Pretty soon we will need a dedicated sub for black-on-senior sexual assaults.*"
4. "*Who is the target audience? I would think black literacy levels would prevent "n*g lit" from ever being a viable book market.*"

Overall, our analysis of Type II errors indicated that the vast majority of mislabeled comments are not racist and are, therefore, correctly labeled. This suggests that the actual performance of our method is likely higher than what we report.

3.1.6 Conclusion

The presence of hateful speech on online platforms is a growing problem with a need for robust and scalable solutions. In this work, we investigated the limitations of keyword-based methods and introduced a community-based training method as an alternative. Our work makes two key contributions.

First, we highlight two major mechanisms that hurt the performance of keyword-based methods. The shared vocabulary between hateful and support communities causes training positive examples to contain non-hateful content. Also, because keyword lists focus on more widely known slurs, these lists miss many instances of hateful speech that use less common or more nuanced constructions to express hatred all too clearly.

Our second contribution is the idea of using self-identified hateful communities as training data for hateful speech classifiers. This approach both involves far less effort in collecting training data and also produces superior classifiers.

The promising results obtained in this study suggest several opportunities for future work. Foremost is the extension of this approach to other non-forum-based platforms. Twitter and Facebook, for example, are heavily used platforms which mainly feature short-text messages. Such content presents unique challenges that will require new or modified approaches. Another direction involves looking at other high-signal features (syntax, n-grams, and sentiment scores).

In these and other initiatives, we believe that community-based data may play an essential role in producing both better detectors of hateful speech, and a richer understanding of the underlying phenomenon.

Bibliography

[4] James Allan. “The Harm in Hate Speech”. In: *Constitutional Commentary* 29.1 (2013), pp. 59–80.

[13] Jamie Bartlett et al. “Anti-social media”. In: *Demos* (2014), pp. 1–51.

[30] Mary Bucholtz and Kira Hall. “Identity and interaction: A sociocultural linguistic approach”. In: *Discourse studies* 7.4-5 (2005), pp. 585–614.

[80] Jesse Fox and Wai Yen Tang. “Sexism in online video games: The role of conformity to masculine norms and social dominance orientation”. In: *Computers in Human Behavior* 33 (2014), pp. 314–320.

[94] John J Gumperz. “The speech community”. In: *Linguistic anthropology: A reader* 1 (2009), p. 66.

[125] Olivier Klein, Russell Spears, and Stephen Reicher. “Social identity performance: Extending the strategic side of SIDE”. In: *Personality and Social Psychology Review* 11.1 (2007), pp. 28–45.

[131] Irene Kwok and Yuzhou Wang. “Locate the Hate: Detecting Tweets against Blacks.” In: *AAAI.* 2013.

[150] Toby Mendel, M Herz, and P Molnar. “Does International Law Provide for Consistent Rules on Hate Speech?” In: *The content and context of hate speech: Rethinking regulation and responses* (2012), pp. 417–429.

[159] Jessica Moreno, Ellen Pao, and Alexis Ohanian. *Removing harassing subreddits.* 2015. URL: `www.reddit.com/r/announcements/comments/39bpam/removing_harassing_subreddits/`.

[161] Lisa Nakamura. "Don't hate the player, hate the game: The racialization of labor in World of Warcraft". In: *Critical Studies in Media Communication* 26.2 (2009), pp. 128–144.

[183] Xuan-Hieu Phan and Cam-Tu Nguyen. "Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference". In: (2006).

[189] Stephen D Reicher, Russell Spears, and Tom Postmes. "A social identity model of deindividuation phenomena". In: *European review of social psychology* 6.1 (1995), pp. 161–198.

[207] Russell Spears and Martin Lea. "Panacea or panopticon? The hidden power in computer-mediated communication". In: *Communication Research* 21.4 (1994), pp. 427–459.

[208] Russell Spears and Martin Lea. *Social influence and the influence of the'social'in computer-mediated communication.* Harvester Wheatsheaf, 1992.

[213] I-Hsien Ting et al. "An approach for hate groups detection in facebook". In: *The 3rd International Workshop on Intelligent Data Analysis and Management.* Springer. 2013, pp. 101–106.

[214] Robert S Tokunaga. "Following you home from school: A critical review and synthesis of research on cyberbullying victimization". In: *Computers in human behavior* 26.3 (2010), pp. 277–287.

[227] William Warner and Julia Hirschberg. "Detecting hate speech on the world wide web". In: *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics. 2012, pp. 19–26.

[239] Guang Xiang et al. "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus". In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM. 2012, pp. 1980–1984.

[END OF MANUSCRIPT]

In the previous manuscript, we see that, online communities that target a marginalized group share vocabulary with communities that support the same group. This indicates that supportive and antagonistic communities discuss similar topics, with opposite polarities.Language models trained on such data are able to reliably distinguish antagonistic comments from random as well as supportive comments. Furthermore, community-driven models are even able to identify comments from similar antagonistic communities across platforms. However, when identifying abuse against a community they were not trained on, these models perform poorly. This helps illustrate the problem of diversity in abusive language - different marginalized populations are targeted with different language. We can assert that a framework for generalized detection of abuse would require expansive training resources.

Overall, antagonistic and supportive communities are viable resources of diverse abusive language. In the next chapter I build a taxonomy that allows for precisely addressing this diversity and then leverage online communities to construct an abusive language corpus.

CHAPTER 4
Community Sampling for an Equitable Corpus of Abuse

Abusive language frameworks, like other machine learning solutions, are data-driven. They rely on large amounts of high quality data to successfully train classification algorithms. However, the sheer diversity within abusive language makes it challenging to reliably detect. Abusive language varies highly across marginalized populations. In the last chapter, we observe that a model, when trained on one kind of abusive language, performed poorly out of domain. Even for a singular marginalized group, abusive language can take varied forms, making detection non-trivial. For meaningful analysis and detection of abusive language we require:

1. a formal system to refer to the various forms of abusive language that accurately captures its diversity, and
2. a large-scale corpus that incorporates this diversity.

To tackle the first issue, I focus on slurs - pejoratives that degrade their intended targets. In essence, they are the unit form of abusive language, since pejorative expressions encode derogatory force within themselves. Focusing on pejorative language allows us to: 1) filter for abuse from the vast volumes of user-generated content, and 2) study a tractable form of abusive language. After a comprehensive exercise in open coding, I introduce a formal taxonomy which takes into account the various contexts in which slurs can be used. The slur taxonomy is a fine grained vocabulary of both abusive and non abusive content and helps promote nuance in abusive language research.

For the second issue, I rely on online communities. The previous chapter introduces polar communities and in this chapter I leverage them. I present community sampling - a novel approach that collects data from a collection of polar communities. Such strategic sampling helps ensure diversity of content within the aggregated corpus. I further ensure diversity of opinion during annotation by assembling a diverse cohort of annotators. The construction of this corpus encodes variation in content as well as assessment. Another defining feature of this the corpus is the presence of pejorative expressions in each comment, which serves as a strong benchmark for methods that over-fit on such language.

4.1 Manuscript 2: Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage

Authors: Haji Mohammad Saleem[†], Jana Kurrek[†], and Derek Ruths

Published in the proceedings of the Fourth Workshop on Online Abuse and Harms @ EMNLP 2020

4.1.1 Abstract

Abusive language classifiers have been shown to exhibit bias against women and racial minorities. Since these models are trained on data that is collected using keywords, they tend to exhibit a high sensitivity towards pejoratives. As a result, comments written by victims of abuse are frequently labelled as hateful, even if they discuss or reclaim slurs. Any attempt to address bias in keyword-based corpora requires a better understanding of pejorative language, as well as an equitable representation of targeted users in data collection. We make two main contributions to this end. First, we provide an annotation guide that outlines 4 main categories of online slur usage, which we further divide into a total of 12 sub-categories. Second, we present a publicly available corpus based on our taxonomy, with 39.8k human annotated comments extracted from Reddit. This corpus was annotated by a diverse cohort of coders, with Shannon equitability indices of 0.90, 0.92, and 0.87 across sexuality, ethnicity, and gender. Taken together, our taxonomy and corpus allow researchers to evaluate classifiers on a wider range of speech containing slurs.

[†] These authors made equal contributions.

4.1.2 Introduction

Detecting abusive language is important for two substantive reasons. First is the mitigation of harm to individuals. Exposure to hate speech can result in a wide range of psychological effects, including degradation of mental health, depression, reduced self-esteem, and greater stress expression [193, 219, 23]. Second is the broader impact of unregulated speech on the participation gap in social media [107, 169]. Overexposure to hateful language results in user desensitization [206] and radicalization [168], both of which have been shown to worsen racial relations [199]. Moreover, hateful echo-chambers promote a “spiral of silence” that discourages counter-speech in conversations online [71].

Access to large-scale training data is the first step towards robust automated systems for abusive language detection. While industry researchers can access moderator logs and user reports, proprietary data is not the standard for academics. Instead, pejorative keywords are commonly used as filters in the data collection process. These include, but are not limited to, slurs and other curated lists of profane language [229, 228, 123, 190], terms borrowed from Hatebase, a multilingual repository for hate speech [200, 61, 78, 72], offensive hashtags [41, 89], and manually selected threads or subreddits [84, 97, 186]. Although the drawbacks of keyword-based approaches are known to researchers, there are currently no clear alternatives to this technique [229, 61, 72].

There has been a recent focus on how technical choices involving data curation can introduce systemic bias in the resultant corpus. For instance, Wiegand, Ruppenhofer, and Kleinbauer [233] discover that terms like *football*, *announcer*, and *sport* have the strongest correlation to abusive posts in Waseem and Hovy [229]. Furthermore, Davidson, Bhattacharya, and Weber [60], Xia, Field, and Tsvetkov [238] and Sap et al. [196] reveal how classifiers trained on data with systemic racial bias have a higher tendency to label text written in African-American English as abusive. Cited examples include: “Wussup, nigga!”, and “I saw his ass yesterday”. Left unaddressed, bias has a real impact on users. Automated recruiting tools used by Amazon.com were shown to discriminate against women [52]. Similarly, Microsoft released a public chatbot that learned to share racist content on Twitter [224]. A common solution is to debias language representations [26]. However, these methods conceal but do not remove systemic bias in the overall data [91].

A way of beginning to address the issue of racial and gender bias is therefore to understand the implications of forced sampling. Our paper focuses specifically on data that is collected using derogatory keywords and we make two main contributions to this end. First, we provide an annotation guide that outlines 4 main categories of online slur usage, which we further divide into a total of 12 sub-categories. Second, we present a publicly available corpus based on our taxonomy, with 39.8k human annotated comments extracted from Reddit. We also propose an approach to data collection and annotation that prioritizes inclusivity both by design and application:

Inclusivity by Design. Data selection and annotation achieves weighted group representation. We sample from a variety of subreddits in order to capture non-derogatory slur usage. We then hire a diverse set of coders under strict ethical standards as a means of engaging the perspectives of various target communities.

We encourage opinion diversity by pairing annotators into teams based on maximum demographic differences.

Inclusivity by Application. Our coding guidelines are extensible to language that targets multiple protected groups. We collect data using the slurs: *f*ggot*, a pejorative term used primarily to refer to gay men, *n*gger*, an ethnic slur typically directed at black people, especially African Americans, and *tr*nny*, a derogatory slur for a transgender person. This is only time we mention the actual slurs. From hereon, We refer to each term as the f-slur, n-slur, and t-slur, respectively. We specifically choose these slurs because they enable us to study discrimination across sexuality, ethnicity, and gender.

Our work does not directly eliminate bias in existing datasets. Rather, it aids in truly understanding the different ways in which slurs can be used online so that models can be trained and assessed more effectively.

4.1.3 Related Work

Existing Hate Speech Corpora

The earliest and most notable corpus for hate speech research is Waseem and Hovy [229]. It contains 16k comments from Twitter, annotated according to the offense criteria of McIntosh [148]. Waseem [228] is an extension of this corpus by 6,909 comments and it considers amateur as well as expert annotations. The authors make use of offensive hashtags for data collection, but it was not until Nobata et al. [167] that slurring language was formally introduced as a sub-problem of hate speech. This paper uses a variety of linguistic features, such as modal words, insulting and hate blacklist words, and politeness words, in order to separate the three

Authors	**Size**	**Platform**	**Annotation**	**Agreement**
Keyword based data collection				
Qian [186]	34k 22k	Gab Reddit	Hate Speech	Unknown
Waseem [229]	16k	Twitter	Racism, Sexism	$\kappa = 0.84$
Waseem [228]	7k	Twitter	Racism, Sexism	$\kappa = 0.34$
Golbdeck [89]	35k	Twitter	Hate Speech, Threats, Harassment, Offense	$\kappa = 0.84$
Chatzakou [41]	9k	Twitter	Aggressors, Bullies, Spammers	Agreement = 0.54
Davidson [60]	25k	Twitter	Hate Speech, Offense	Agreement = 0.92
Rezvan [190]	25k	Twitter	Harassment	$\kappa \sim 0.81$
Founta [78]	80k	Twitter	Hate Speech, Spam, Abuse	Unknown
ElSherief [72]	2k	Twitter	Hate Speech	$\alpha = 0.622$
Jha [111]	1k	Twitter	Sexism	$F\kappa = 0.74$
Silva [200]	539.5m	Twitter Whiser	Hate Speech	Not applicable
Fersini [76]	3k	Twitter	Sexism	Unknown
Basile [14]	19.6k	Twitter	Hate Speech, Target, Aggressiveness	F8 = 0.83 0.70, 0.73
Zampieri [242]	14.1k	Twitter	Offense, Target	$F\kappa = 0.83$*
Manual selection				
Gao [84]	1.5k	Fox News	Hate Speech	$\kappa = 0.98$
Hammer [97]	30k	Youtube	Threats	Unknown
Proprietary data				
Sprugnoli [209]	15k	WhatsApp	Cyberbullying	SDC = 0.80 - 0.88
Nobata [167]	1.2m	Yahoo	Hate Speech	$F\kappa = 0.21$ (AMT) $F\kappa = 0.46$ (Trained)
Random data selection				
deGibert [87]	10k	Stormfront	Hate Speech	$\kappa \sim 0.62$
Napoles [163]	10k	Yahoo	Positive Conversations	$\alpha = 0.79$ (Group) $\alpha = 0.71$ (Trained)
Other methods				
Wulczyn [237]	100k	Wikipedia	Harassment, Attacks	$\alpha = 0.45$
Kennedy [121]	20k	Twitter, Reddit The Guardian	Harassment	Agreement = 0.88

Table 4–1: An overview of the main corpora on abusive language and similar behaviours. $F\kappa$ is Fleiss' Kappa, κ is Cohen's Kappa, SDC is the Sørensen–Dice coefficient, and areement refers to raw disagreement. (* on 21 tweets)

notions of hate, derogation, and profanity based on their relative degrees of harm to the target. These guidelines inspired the Fox News user comments corpus of Gao and Huang [84]. Both works emphasize the capacity for hateful language to exist in implicit and explicit forms and collect the explicit form using derogatory keywords. Silva et al. [200] is a target-based analysis of the explicit form. They leverage the syntactic structure “`I <intensity><user intent><hate target>`”, where each hate target is one of 1,078 terms selected from Hatebase, in order to identify ten top targets of hate within Twitter and Whisper content. Next, Davidson et al. [61] investigate intentional group-based humiliation and derogation. They reinforce the role of slurs as archetypal representations of hate by acknowledging that “tweets with the highest predicted probabilities of being hate speech tend to contain multiple racial or homophobic slurs.” More recently, Gibert et al. [87] sample from a white supremacist sub-forum and, in doing so, encourage community-based filtering. The emerging theme from these research efforts is the consensus that we require an alternative to random sampling for reliably capturing hateful content. What that alternative is remains unclear but keywords are currently the dominant choice.

Other researchers have expanded on this definition and shown that it is applicable to more nuanced categories of online misbehaviour, such as abuse, threats, personal attacks, and cyberbullying. For instance, Khodak, Saunshi, and Vodrahalli [123] is a self-annotated corpus for sarcasm on Reddit. Sprugnoli et al. [209] focuses on cyberbullying within WhatsApp conversations. Rezvan et al. [190] points out sexual, appearance-related, intellectual, and political harassment on Twitter. Hammer et al. [97] is a corpus for detection of violent threats on YouTube. Holgate et al.

[99], Cachola et al. [34], and Pamungkas, Basile, and Patti [173] examine vulgarity and swearing. A number of corpora on mixed behaviours have also been produced. Golbeck et al. [89] is a study on harassment and offense on Twitter. Chatzakou et al. [41] labels Twitter users, not comments, as aggressors, bullies, or spammers. Founta et al. [78] considers spam in conjunction with abuse, bullying, and aggression on Twitter. Napoles, Pappu, and Tetreault [163] works on the converse of the problem. This paper uses Yahoo! News data to advance a corpus on constructive conversations.

We have collected a list of the major English-language corpora and summarized their sizes, platforms of focus, annotation schemes, and agreement scores in Table 4–1. With that said, the study of online misbehavior has been extend beyond the traditional focus on English. It now includes resources in Italian, Indonesian, Hindi-English, Tunisian, etc. [195, 103, 128, 24, 95, 160, 49].

Slurs

To model the contents of slur-based data, it is crucial that we first examine the properties of slurs themselves. Slurs are pejoratives that derogate based on in-group membership, that is, they categorize targets based on institutionally defined archetypes [56]. Studies on slurs are built on the recognition by Kaplan [115] that meaning in natural language comes from convention and from context: a sentence is *expressively correct* if it is true by interpretation; a sentence is *descriptively correct* if it is literally true.

Hom [101] advocates in favor of the expressive view of slurs. He identifies nine adequacy conditions that characterize and explain racial epithets: A slur exhibits

(1) derogatory force. The force of any slur is (2) variable across epithets and (3) fundamentally offensive, independently of the intents and beliefs of the speaker. While slurs are capable of being (7) reclaimed or (8) used towards a non-derogatory, non-appropriative end, they are generally (4) taboo unless (6) their force changes over time. This is because slurs are (5) meaningful insofar as they contribute to the truth-conditions of the sentence in which they arise. Hom's account of slurs is (9) generalizable across pejoratives.

Hom implies that there are three main categories of slur usage, which are derogatory, non-derogatory non-appropriative, and appropriative. His adequacy conditions are central to our research. The three categories are the basis of our annotation scheme and they enable us to make assessments of abuse with ambiguous user intent.

4.1.4 Inclusive Design Process

Random sampling of slur-based data allows for proportional representation because the share of each usage in the corpus is reflective of its probability of occurrence online. However, this approach is not equitable. Less common usages, such as reclamation, discussion, and counter-speech, are not captured. Consequently, language models can overfit on pejoratives and further codify institutional biases [36, 85]. A top-down approach to debiasing is simply insufficient. We advocate in favor of affirmative action during data collection and make an effort to represent a wider range of slur usages through community targeting. We also tailor our study to include individuals that belong to targeted communities, both as authors and annotators.

f-slur	**n-slur**	**t-slur**
Supportive communities		
askgaybros	BlackPeopleTwitter	transgendercirclejerk
gaybros	Blackfellas	traaaaaaannnnnnnnnns
lgbt	blackladies	asktransgender
ainbow	beholdthemasterrace	ainbow
LGBTeens	AgainstHateSubreddits	transgender
Antagonistic Communities		
4chan	CoonTown	TumblrInAction
ImGoingToHellForThis	uncensorednews	MGTOW
The_Donald	WhiteRights	Braincels
CringeAnarchy	GreatApes	metacanada
TheRedPill	european	GenderCritical
General Discussion Communities		
funny	todayilearned	rupaulsdragrace
pics	videos	cars
politics	changemyview	Drama
AskReddit	worldnews	AdviceAnimals
atheism	movies	unpopularopinion

Table 4–2: This table presents the major supportive, antagonistic, and general discussion subreddits that were used in data collection. Their range of views towards the targets of each slur facilitates equitable representation.

Data Collection

We use the Pushshift Reddit corpus [15] and filter for the three slurs (f-slur, n-slur, t-slur) and their plurals. The data ranged from October 2007 to September 2019 at the time of filtering. We extracted a total of 2.6 million comments. We applied the following filtering process:

Author Level. We remove comments written by users with no history in order to leave open the possibility of a future analysis with user meta-data. We remove comments written by users that were identified as bots. We limit the number of comments written by the same author.

Comment Level. Reddit comments vary in length, with an upper limit of 40,000 characters. For ease of annotation, we remove comments from the top and bottom quartiles by length. We limit our corpus to English-language comments and use the Compact Language Detector v3[1] to detect them.

Community Level. Communities that antagonise or support a group talk about similar topics but with opposing valence [194]. To capture such polarity, we compile a list of subreddits based on their disdain for, neutrality towards, or support of the f-slur, n-slur, and t-slur (see Table 4–2). We do this by building on an existing list of toxic Reddit communities [35]. We consider the name, rules, extent of moderation, description text, and polarity of comments containing slurs (overall score) of each subreddit in our assessment of whether or not to include them. We then extract the top comments in terms of polarity.

Our post-filter corpus has 40,000 comments, sourced from 2704 individual subreddits and 37,133 unique authors. The median and maximum number of comments per author is 1 and 5.

Taxonomy Design

Our coding guide is based on the three major categories of slur usage identified in Hom [101]. By open coding data collected using slurs, we identify a fourth major category as well as twelve subcategories. The complete taxonomy, along with examples for each subcategory, is provided in Table 4–3. In general, comments containing

[1] `github.com/google/cld3`

more than one slur were labelled according to the most derogatory usage. The four main categories are explained below:

Derogatory Usage (DER). Any usage that is understood to convey contempt towards a targeted individual or group.

Appropriative Usage (APR). Meaningful usage by the targeted group for an alternate, non-derogatory purpose. Text belonging to this label loses its derogatory force.

Non-Derogatory, Non-Appropriative Usage (NDG). Meaningful usage by targeted or non-targeted groups for an alternate non-derogatory, non-appropriative purpose. Text belonging to this label retains its derogatory force.

Homonyms (HOM). A slur with one or more non-derogatory alternative meanings.

Annotator Selection

Following approval by the university Research Ethics Board (REB), we shared messages on social media and university mailing lists as well as physical posters across faculties in order to look for participants. The application consisted of eight short answer questions, in which candidates were asked to disclose their name, email, field and year of study, age, sexuality, ethnicity, and gender. We specifically collected the demographic information in free-form text. The free-form allows participants to choose best demographic identifiers for themselves. The demographic information is confidential and used solely for selecting annotators and creating their teams.

All demographics were collapsed into categories (see Figure 4–1) primarily based on the classification structure approved as a departmental standard by Statistics

Slur Usages	Example Text
DEROGATORY	
Attribution	he’s an ugly [f-slur] with greasy hair.
Community Focus	lol don’t be a [f-slur]
Stand Alone	[t-slur]
Sexualization	I love the taste of a nice hot [t-slur] load
Self-Deprecation	as mizkif i can agree i look like a [f-slur]
APPROPRIATIVE	
Reclamation	get in [t-slur] Formation everyone, it’s time to march against the tyranny of heteronormatives trying to appropriate OUR WORDS
NON-DEROGATORY, NON-APPROPRIATIVE	
Counter Speech	[t-slur] is a slur please don’t use it.
Direct Quotations	actual quote: de [n-slur] woman is de mule uh de world so fur as ah can see.
Discussion	You could call someone a [f-slur] in the 70s and 80s with absolutely no recourse.
Recollection	I never got so much shit until I graduated high school. :— I get called a [f-slur] by some random clitdick almost every day I have class.
Sarcasm	Yeah because apparently [f-slur] all of a sudden isn’t a slur used against homosexuals.
HOMONYMS	
	transmissions are beautiful pieces of engineering, why not have a [t-slur] tattoo?
	[f-slur] Hill, 969th tallest peak in Massachusetts... why even count at that point?
	Damn talk about being able to skate anything. Rips [t-slur] then throws in kickflip back lips on rails.

Table 4–3: Our taxonomy of slur usage, with 4 main categories broken down into 12 subcategories. Examples are provided for each subcategory and further details can be found in the Appendix (Section 4.1.7).

Canada (2017). Of the four hundred and twelve applications received, 20 participants, ranging between 19 and 65 years of age (M = 26.7, SD = 10.8), were chosen using iterative proportional fitting. Overall, our annotator cohort has a Shannon equitability index of 0.90, 0.92, and 0.87 across sexuality, ethnicity, and gender. We

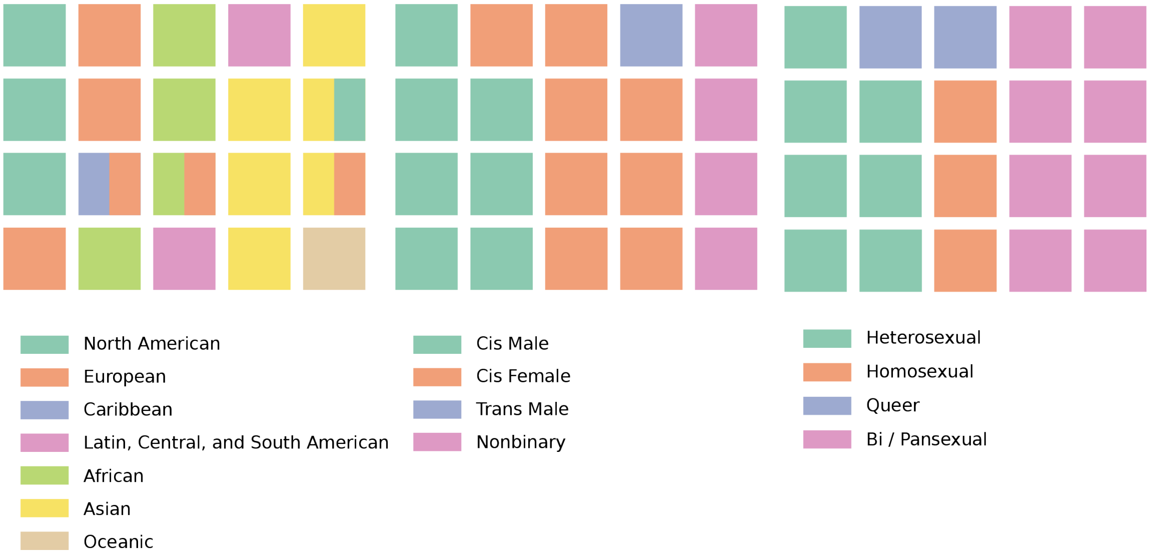


Figure 4–1: The diverse demographic details of our annotator cohort, aggregated on ethnicity, gender and sexuality.

did not have the REB clearance to perform any further analysis on the relationship between annotator demographics and annotations. We leave this as an area for future work.

Training and Annotation

A 4-session on-campus training program was developed for annotators to attend over 2 days. On Day 1, we presented the annotation scheme obtained through open coding. Annotators were then guided through two group annotation exercises of 20 and 40 comments respectively. On Day 2, annotators were randomly divided into 4 teams. Each team completed 2 rounds of 200 training annotations. After each round, they discussed their annotations and the reasons behind their labels. The discussion was aimed at fostering a common understanding of the annotation process.

The final annotations were divided into 4 tasks of 10,000 comments each. The 20 annotators were grouped into 10 teams of 2. The team creation process maximized the demographic distance between members across sexuality, ethnicity, and gender. It was treated as an assignment problem and solved using the Kuhn-Munkres algorithm. Each team annotated 1000 comments per task and annotators were grouped into new pairs for each subsequent task. Comments with no disagreement were added to the final corpus. Comments with disagreement were resolved by the authors. The final annotations were performed remotely on the open source text annotation tool Doccano [162].

4.1.5 Labeled Corpus

40,000 Reddit comments were annotated, of which 189 were removed as noise. The remaining 39,811 were closely split across slurs: 13,290, 13,267 and 13,267 for f-slur, n-slur and t-slur respectively. In total, 20,531 comments were labelled derogatory, 16,729 non-derogatory, 1,998 homonym, and 553 appropriative. We anticipated a large portion of derogatory comments in our corpus because our data is slur-based. However, only 52% of comments were labelled as such. We attribute this to our community-targeted data collection process and efforts to sample from supportive subreddits.

Label Distribution Across Slurs

In Figure 4–2, we present the label distribution across slurs. We observe that roughly 59% of comments collected using the f-slur and t-slur were labelled as derogatory. In comparison, about 37.9% of comments containing the n-slur were similarly labelled. The majority of found homonyms include the t-slur, which accounts for

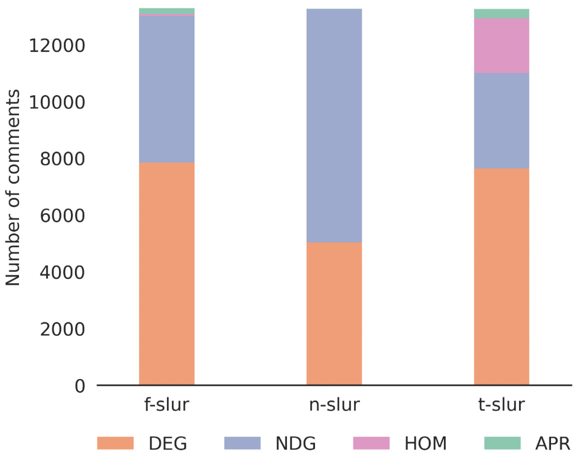


Figure 4–2: Label distribution across slurs in the annotated corpus.

95.9% of the label. This is largely because the term is used in automotive communities to mean vehicle transmission (see Figure 4–3) and in skateboarding communities to describe skating transition. The remaining homonyms include the f-slur, with the meaning “bundle” or in reference to a form of British meatball. The n-slur has the smallest share of homonyms (0.02%) and appropriative (0.16%) comments.

Label Distribution Across Subreddits

In Figure 4–3, we present the label distribution across the 50 most common subreddits in our corpus. The graph is sorted by the proportion of derogatory comments in each subreddit. Consequently, it can be seen as a scale of derogatory behavior. On the far right are communities that we had previously identified as antagonistic. Many of their comments were labelled as derogatory and examples include `MGTOW`, `CoonTown`, `4chan` and, `The_Donald`. In the middle we find general discussion subreddits such as `videos`, `todayilearned`, and `politics`. They generally have an even split of derogatory and non-derogatory labels. On the far left we observe mostly

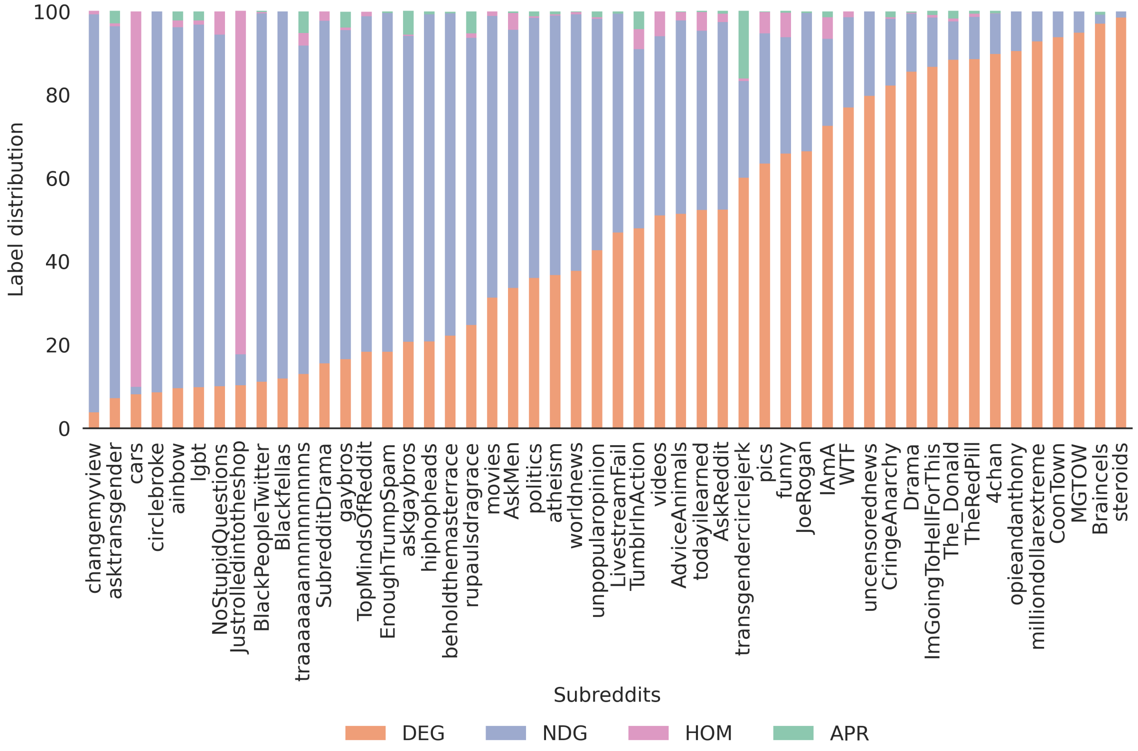


Figure 4–3: Normalized label distribution across the 50 most common subreddits in our corpus, sorted by their portion of derogatory comments.

supportive subreddits, with small portions of derogatory comments. Automotive subreddits like `cars` have a large number of homonyms. Meanwhile, subreddits such as `traaaaaaannnnnnnnnns`, `askgaybros`, and `rupaulsdragrace` contain significant portion of appropriative speech. These findings align with our initial hypothesis about supportive, antagonistic, and general discussion communities.

Agreement Analysis

Both annotators agreed on the same label for 31,034 of the comments in our corpus. The remaining 8,777 comments were resolved by the authors. Overall we

	Agreement (%)	Cohen's κ
overall	78.6	0.60
f-slur	79.7	0.58
n-slur	75.4	0.51
t-slur	80.5	0.65

Table 4–4: Raw and inter-rater agreement. We achieve moderate to substantial agreement with Cohen's κ.

achieve a raw agreement score of 78.6%, corresponding to a Cohen's κ of 0.60. Our scores indicate substantial agreement and are in line with what has been observed in the literature (see Table 4–1). We obtain similar agreement across the three slurs, which are presented in Table 4–4.

APR had the highest amount of disagreement, with 67.99% comments requiring resolution, followed by NDG (35.36%), and HOM (31.58%). DEG was the lowest at 9.034%. During the resolution process, we identified three probable causes for disagreement:

Label Overlap. Discussions of derogation or reclamation created ambiguity and were falsely labelled as DER or APR, rather than NDG. A similar issue arose in comments acknowledging slurs as homonyms. For instance: "*When i was telling my skate friends about me being trans i asked them if they knew why it was so ironic that i love skating [t-slur] so much.*".

Satire. In `transgendercirclejerk`, which is a subreddit that self-identifies as a "parody for trans people, mocking all transgender-related topics", our annotators found many derogatory comments (see Figure 4–3). However, the sarcastic or satirical nature of these comments was not always evident: "*We don't need gun control we need [T-SLUR] CONTROL! [t-slurs] are not in the Constitution or Bible,*

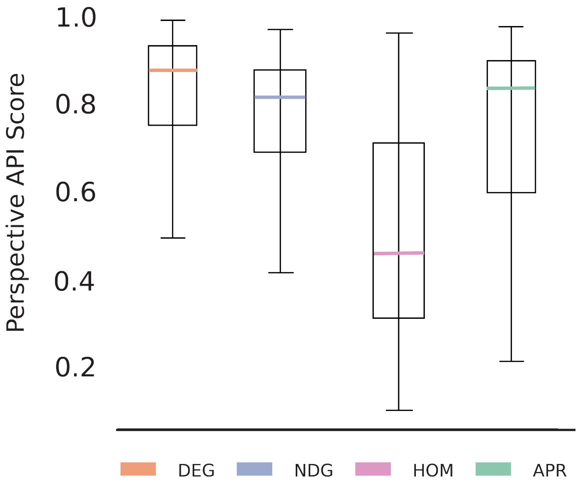


Figure 4–4: Benchmarking the Perspective API. Scores indicate a comment's degree of toxicity.

like guns are! If we don't outlaw t-slurs, only [t-slurs] will have outlaws!". We leave this area for future work.

Lack of Context. In an independent assessment of label reliability, we re-annotated 100 DEG comments from `transgendercirclejerk` with complete access to user and thread history. 44 of our labels did not match those submitted by annotators. For instance, the following comment came from a transgender poster: "*LA LA LA CAN'T HEAR YOU I'M STUCK IN [T-SLUR] REALITY*" but was mislabelled. This testifies the difficulty of annotating appropriative language without context. Other instances that requires context are reference to lyrics and dialogues from pop culture.For example "Dead [n-slur] Storage" from the movie Pulp Fiction.

Benchmarking the Perspective API

We use a state-of-the-art model for derogatory content detection to assess whether current classifiers are subject to overfitting on pejoratives. We choose the Perspective

API by Conversation AI, which “identifies whether a comment could be perceived as toxic to a discussion”. We obtain the toxicity scores for 100 random comments for each of the DEG, NDG, HOM, and APR labels. The results are summarized in Figure 4–4. As expected, the overall score distribution is high for DEG. However, it is equally high for NDG and APR comments. This perfectly illustrates the issue of potentially biased models failing to identify non-derogatory content.

Further analysis of toxicity scores across comments underlines the challenges faced by existing models. First, instances of slur reclamation received high toxicity scores. For example: “*Psh my [t-slur] agony sits atop that steed with militant fervour. The world shall hear me roar, I AM A [T-SLUR] FREAK!!!! /uj Not even kidding, I’m 100% out as a [t-slur] freak. World can suck my shenis*” and “*When I’ve got a guy I’m crushing on I will sometimes say ‘He makes me feel like a silly [f-slur] all over again*’” have toxicity scores above 0.93. Reclamation is an attempt at empowerment and community cohesion. The mislabelling of such examples further censors communities already targeted by hate. Second, recollections of past harassment received high toxicity scores. For example: “*A homeless dude called me a spic [f-slur] once while I was with my ex*” is rated as high as 0.889. This belittles victims’ experiences with abuse, rather than protecting them from it. Finally, counter speech received high toxicity scores. For example: “*Ummmm, yeah no, [t-slur] is a slur and youre ignorant as hell*” is rated 0.953. This undermines community-level efforts at removing derogatory language. Overall, these three outcomes are counterproductive to the detection process since empowering and vulnerable conversations of targeted communities may be flagged down.

4.1.6 Conclusion

We present a comprehensive taxonomy and large-scale annotated corpus for online slur usage. Our findings are an attempt at integrating a qualitative understanding of slurs into their usage in natural language. We believe that they provide a significant contribution to the hate speech research community, not only as resources for training machine and deep learning models, but also as a means of achieving a nuanced understanding of the phenomenon of slurs. We encourage researchers to replicate and expand our efforts by studying language that targets other marginalized communities. With that said, our corpus is a challenging benchmark that will help expose over-fitting on pejoratives and our taxonomy introduces a systematic approach for dealing with derogatory keywords and epithets. Our corpus can be accessed by emailing the authors.

4.1.7 Appendix - A Taxonomy of Slur Usage

Derogatory Slurs

Derogatory slur usage is any usage that is understood to convey contempt towards a targeted individual or group [101]. It is divisible into the following five subcategories:

1. Attribution
2. Community Focused Slurs
3. Stand-Alone Slurs
4. Sexualization
5. Self-Deprecation

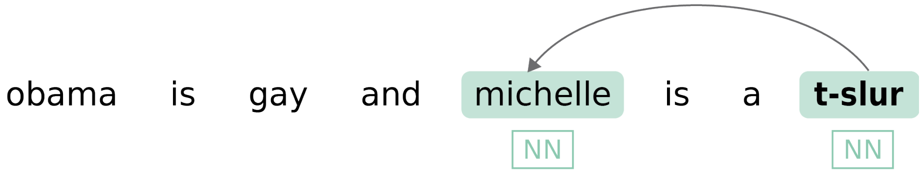


Figure 4–5: Noun-Noun Structure of Attribution

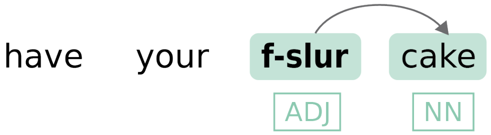


Figure 4–6: Adj-Noun Structure of Attribution

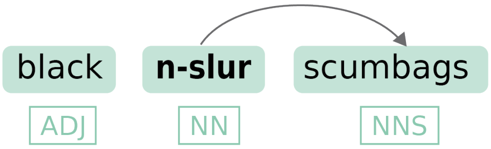


Figure 4–7: Adj-Noun-Noun Structure of Attribution

Attribution. In general, a slur may be used as an attributive noun or as an attributive adjective. An attributive adjective typically follows an Adjective-Noun structure. Its structure can be recursed any number of times within a noun phrase, but attributive order and scope may affect the meaning of the sentence [217]. An attributive noun follows a free-form or bound-form compound structure [92] and slurring adjectives may be used as the base for noun inflections [233]. Examples of noun inflections are: *ni***ring*, *tr***iversary*, and *f**gy*.

Implicit attributive nouns target an entity that is unnamed but present in the sentence. This entity has attributes that are conventionally implicated by the slur. Such is the case with the *n-slur* in "*n-slur*, I'm tired of this". Two statements are made by the speaker: the first, (you are the *n-slur*), is implicit, and the second, (I'm tired of this), is explicit. The sentence is therefore semantically equivalent to "(you are the *n-slur*) & (I'm tired of this)". Similarly, in "not get downvoted by every *f-slur* on reddit for everything I say", *f-slur* acts as an implicit attributive noun because it suggests that (everyone who downvotes me is the *f-slur*). Other examples of implicit attributive nouns are nicknames and informal vocatives [92].

Community Focused Slurs. Every slur has a neutral counterpart that exists at a lower descriptive level in the sentence. It is called the Non-Pejorative Correlative (NPC) and substituting one term for the other is a truth-conditionally, but not expressively, equivalent replacement. Hom [101] formally defines the NPC for a slur as the expression that picks out the supposed extension of the epithet without derogating members of that extension:

1. that means literally nothing to a *n-slur*
2. that means literally nothing to an *African-American*

Community focused slur usage is separated from other sub-categories of slur usage because it is generalizable to all community members, rather than a finite set that is referenced by the speaker. Attribution is, however, taken to be community-focused when it is accompanied by negation. The following example illustrates why this is so:

1. lol don't be a *f-slur*
2. lol (do not you) be a *f-slur*
3. lol (do not you) be *homosexual*

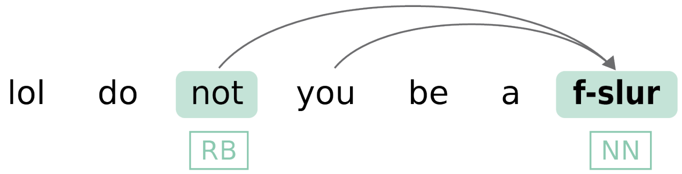


Figure 4–8: Attribution is taken to be community-focused when it is accompanied by negation. The inclusion of *do not* redirects the derogatory force associated with *f-slur* onto *homosexual*, rather than *you*.

Stand Alone Slurs. A slur is stand-alone when it is the only, single or repeated, word in a phrase:

- *f-slur*
- *n-slur*! *n-slur n-slur n-slur*
- *t-slur*;))!!!!

Sexualization. A slur is sexualized when it is used as imagery for the purpose of fetishization, pornography, prostitution, or objectification. We isolate sexualized pejorative language from language in a violent or humorous context because the inclusion of the slur introduces a transactional element to derogation. Not only is the target community degraded, but, especially in the case of pornography, they are portrayed as a sexual object for consumption.

- A very hot black *t-slur* is mastu `URL`.
- I love the taste of a nice hot *t-slur* load
- it wasn't just hardcore porn, it was hardcore *t-slur* porn! fact check'd!

Self-Deprecation. A slur is used for self-deprecation when it is negatively attributed to the speaker. Its appropriative counterpart is slur reclamation, which is characterized by a positive attribution to the speaker.

Appropriative Slurs

Appropriative slur usage is defined as meaningful usage by the targeted group for an alternate, non-derogatory purpose. It contains one subcategory:

1. Reclamation

Reclamation. A slur is reclaimed when it loses its derogatory force, variation, and autonomy [101]. Applications of slur reclamation include, but are not limited to, self-, individual-, and group-targeted empowerment. These may take the form of positive attributions or connotations:

- i just don't see the point in pointing out how different we are when we're trying to be treated as equal. not all of it is about that but some of it is and it's that part that gets recognized by the public. i also feel no association to any of it. i'm a *f-slur* because i like dick...not rainbows. i hate rainbows.
- well now i kinda feel like the odd *t-slur* out lol(sorry bout the t word guys, all in good spirits.) i think i tried it because ive read about alot of other transmen who enjoyed it after they began transitioning
- get in *t-slur* Formation everyone, it's time to march against the tyranny of heteronormatives trying to appropriate OUR WORDS

Non-Derogatory Non-Appropriative (NDNA) Slurs

NDNA slur usage is meaningful usage by targeted or non-targeted groups for an alternate non-derogatory, non-appropriative purpose [101]. A comment falling under this use case retains the derogatory force of the slur. Its subcategories are:

1. Referential Language
2. Counter Speech

3. Sarcasm

Referential Language.

Discussion of the slur, its origin, or acceptable use cases;

- ask him if he would say *n-slur* if he had a black roommate
- I didn't really know *t-slur* was that inappropriate.
- I always get downvoted for saying *t-slur* or *sh*male*
- Just 5-6 years ago you could say *f-slur* and noone would care

Direct Quotes, excluding paraphrasing and quotes involving non-embedded standalone slurs;

- what a about when harry truman said i think one man is just as good as another so long as he's not a *n-slur* or a chinaman?
- there is no dead *n-slur* storage. you know why? cause storing dead n-slur ain't my fuckin' business that's why!
- the day dave found out the girl he likes is a *t-slur* screenshot vote history on srscmharts ¡ this comment posted by a bot — report an error
- actual quote: "de *n-slur* woman is de mule uh de world so fur as ah can see."

Recollection of a time when the slur was used;

- i used to troll this sub until i realized that it's literally a bunch of teenaged boys who are using what must be their newfound knowledge of sex and stds to see how many different ways they can call someone a *f-slur* so they can impress the older neighbor kids next door.

- one of my friends who happens to be a lesbian has always called me *t-slur*-butt as a term of endearment. to be honest, being called sir is far worse than being referred to as a *t-slur* for me…
- she says *f-slur* the whole time and you can tell she takes great joy in throwing it around
- i get called a *f-slur* by some random clitdick almost every day i have class.

Counter Speech. Counter speech is a discouraging response to derogatory content. A slur used in counter speech responds directly to an instance of derogation, in defense against a comment made by a single speaker or group.

- lmao cmon fam we JUST talked about not using the word *dyke*.
- anyone can say nigga or *n-slur* as long as they don't use it in a derogatory sense.

Sarcasm. A slur used for non-derogatory sarcastic purposes inverts the intuitive truth-conditions of a sentence [37]. As Wilson [234] argues, such is the case when speakers endorse an embedded, ironic meaning that is contrary to the one that they literally convey. For example in "imagine the difficulty of being white though. you have to cross the street every time you see a black man on the sidewalk, that could get exhausting. plus, you won't be able to say *n-slur* in public anymore, how are you supposed to sing along at the rap concert that you paid for?? think about it." the implicature that being white is difficult is inverted by sarcasm to draw attention to cultural appropriation and the derogatory *n-slur*.

Sarcasm is made explicit by Reddit users via the signals: *s*, *rj*, *uj*, or *#sarcasm*

- 1 *t-slur* year = 5 cis years
- yes. straight people are welcome to apply for their *f-slur* license. we'll start you off with a *f-slur* learner's permit, which permits you to say the f-word between the hours of 8am and 11pm, only in the presence of a fully-licensed homosexual. after a trial period, you may be approved for wider use.

Homonyms

A homonym contains one or more dual, non-derogatory alternative meanings [2] .

1. *ch*nk*/n. (i) a small cleft, slit, or fissure; (ii) a weak spot that may leave one vulnerable; (iii) informal, an ethnic slur usually referring to a person of Chinese ethnicity;
2. *f*g*/v. (i) to tire by strenuous activity; n. (ii) cigarette; (iii) unit of measurement; (iv) used as an insulting and contemptuous term for a male homosexual;
3. *tr*nny*/n. (i) vehicle transmission; (ii) slang for transgender person;
4. *b*tch*/n. (i) the female of the dog or some other mammal; (ii) informal, a malicious, or overbearing woman; used as a generalized term of abuse and disparagement for a woman; (iii) informal, something that is difficult, objectionable, or unpleasant; (iv) informal, complaint;
5. *k*ke*/n. (i) insulting and contemptuous term for a Jewish person; (ii) nickname of Enrique;
6. *d*ke*/n. (i) British spelling of dike; (ii) offensive, lesbian; (iii) reference to Dick van Dyke;

[2] Definitions by Merriam-Webster Dictionary

Examples are shown below:

- *chink* in the armor
- but as a whole we form a mighty *f-slur*
- i wouldn't buy a used honda *t-slur* from that era. that's when they had a spate of crappy transmissions and extended the warranties on them until over 100km

[14] Valerio Basile et al. “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2019, pp. 54–63.

[15] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[23] Robert J Boeckmann and Jeffrey Liew. “Hate speech: Asian American students’ justice judgments and psychological responses”. In: *Journal of Social Issues* 58.2 (2002), pp. 363–381.

[24] Aditya Bohra et al. “A dataset of hindi-english code-mixed social media text for hate speech detection”. In: *Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media*. 2018, pp. 36–41.

[26] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems*. 2016, pp. 4349–4357.

[34] Isabel Cachola et al. “Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media”. In: *Proceedings of the*

27th International Conference on Computational Linguistics. Association for Computational Linguistics, 2018, pp. 2927–2938.

[35] Justin Caffier. "Here Are Reddit's Whiniest, Most Low-Key Toxic Subreddits". In: *Vice.com* (2017). URL: `www.vice.com/en_us/article/dyz377/trump-supporters-at-his-tulsa-rally-wrote-a-terrible-trump-2020-reelection-song`.

[36] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[37] Elisabeth Camp. "Sarcasm, pretense, and the semantics/pragmatics distinction". In: *Noûs* 46.4 (2012), pp. 587–634.

[41] Despoina Chatzakou et al. "Mean birds: Detecting aggression and bullying on twitter". In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 13–22.

[49] Yi-Ling Chung et al. "CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2819–2829.

[52] James Cook. "Amazon scraps 'sexist AI' recruiting tool that showed bias against women". In: *The Telegraph* (2018). URL: `www.telegraph.co.uk/technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/`.

[56] Adam M Croom. “The semantics of slurs: A refutation of coreferentialism”. In: *Ampersand* 2 (2015), pp. 30–38.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019, pp. 25–35.

[61] Thomas Davidson et al. “Automated hate speech detection and the problem of offensive language”. In: *Eleventh international aaai conference on web and social media*. 2017.

[71] Megan Duncan et al. “Staying silent and speaking out in online comment sections: The influence of spiral of silence and corrective action in reaction to news”. In: *Computers in Human Behavior* 102 (2020), pp. 192–205.

[72] Mai ElSherief et al. “Hate lingo: A target-based linguistic analysis of hate speech in social media”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[76] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. “Overview of the Task on Automatic Misogyny Identification at IberEval 2018.” In: *IberEval@ SEPLN*. 2018, pp. 214–228.

[78] Antigoni Founta et al. “Large scale crowdsourcing and characterization of twitter abusive behavior”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 2018.

[84] Lei Gao and Ruihong Huang. “Detecting Online Hate Speech Using Context Aware Models”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., 2017, pp. 260–266.

[85] Nikhil Garg et al. “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[87] Ona de Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018, pp. 11–20.

[89] Jennifer Golbeck et al. “A large labeled corpus for online harassment research”. In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 229–233.

[91] Hila Gonen and Yoav Goldberg. “Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them”. 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)*. 2019, pp. 609–614.

[92] Philip B Gove. ““Noun Often Attributive” and “Adjective””. In: *American Speech* 39.3 (1964), pp. 163–175.

[95] Hatem Haddad, Hala Mulki, and Asma Oueslati. “T-HSAB: A Tunisian Hate Speech and Abusive Dataset”. In: *International Conference on Arabic Language Processing*. Springer. 2019, pp. 251–263.

[97] Hugo L Hammer et al. "THREAT: A Large Annotated Corpus for Detection of Violent Threats". In: *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2019, pp. 1–5.

[99] Eric Holgate et al. "Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 4405–4414.

[101] Christopher Hom. "The semantics of racial epithets". In: *The Journal of Philosophy* 105.8 (2008), pp. 416–440.

[103] Muhammad Okky Ibrohim and Indra Budi. "A dataset and preliminaries study for abusive language detection in Indonesian social media". In: *Procedia Computer Science* 135 (2018), pp. 222–229.

[107] Henry Jenkins. *Confronting the challenges of participatory culture: Media education for the 21st century*. Mit Press, 2009.

[111] Akshita Jha and Radhika Mamidi. "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data". In: *Proceedings of the second workshop on NLP and computational social science*. 2017, pp. 7–16.

[115] David Kaplan. "The meaning of ouch and oops. explorations in the theory of meaning as use". 1999.

[121] George Kennedy et al. "Technology solutions to combat online harassment". In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 73–77.

[123] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. "A Large Self-Annotated Corpus for Sarcasm". In: *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*. 2018.

[128] Ritesh Kumar et al. "Aggression-annotated Corpus of Hindi-English Code-mixed Data". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[148] Peggy McIntosh. *White privilege: Unpacking the invisible knapsack*. 1988.

[160] Hala Mulki et al. "L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language". In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 111–118.

[162] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. 2018. URL: `github.com/doccano/doccano`.

[163] Courtney Napoles, Aasish Pappu, and Joel Tetreault. "Automatically identifying good conversations online (yes, they do exist!)" In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[167] Chikashi Nobata et al. "Abusive language detection in online user content". In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 145–153.

[168] Julie Norman and Drew Mikhael. "Youth radicalization is on the rise. Here's what we know about why." In: *The Washington Post* (2017). URL: `www.washingtonpost.com/news/monkey-cage/wp/2017/08/25/youth-radicalization-is-on-the-rise-heres-what-we-know-about-why/`.

[169] Tanya Notley. “Young people, online networks, and social inclusion”. In: *Journal of Computer-Mediated Communication* 14.4 (2009), pp. 1208–1227.

[173] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. “Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media”. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 6237–6246.

[186] Jing Qian et al. “A Benchmark Dataset for Learning to Intervene in Online Hate Speech”. 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)*. Association for Computational Linguistics, 2019, pp. 4757–4766.

[190] Mohammadreza Rezvan et al. “A quality type-aware annotated corpus and lexicon for harassment research”. In: *Proceedings of the 10th ACM Conference on Web Science*. 2018, pp. 33–36.

[193] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. “Prevalence and psychological effects of hateful speech in online college communities”. In: *Proceedings of the 10th ACM Conference on Web Science*. 2019, pp. 255–264.

[194] Haji Mohammad Saleem et al. “A web of hate: Tackling hateful speech in online social spaces”. In: 2016.

[195] Manuela Sanguinetti et al. “An italian twitter corpus of hate speech against immigrants”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[196] Maarten Sap et al. “The risk of racial bias in hate speech detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1668–1678.

[199] Yaye Nabo Sène. “Hate speech exacerbating societal, racial tensions with ‘deadly consequences around the world’, say UN experts”. In: *UN News* (2019). URL: `news.un.org/en/story/2019/09/1047102`.

[200] Leandro Silva et al. “Analyzing the targets of hate in online social media”. In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[206] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. “Exposure to hate speech increases prejudice through desensitization”. In: *Aggressive behavior* 44.2 (2018), pp. 136–146.

[209] Rachele Sprugnoli et al. “Creating a whatsapp dataset to study pre-teen cyberbullying”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 51–59.

[217] Robert Truswell. “Attributive adjectives and the nominals they modify”. PhD thesis. University of Oxford, 2004.

[219] Brendesha M Tynes et al. “Online racial discrimination and psychological adjustment among adolescents”. In: *Journal of adolescent health* 43.6 (2008), pp. 565–569.

[224] James Vincent. “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”. In: *The Verge* 24 (2016). URL: `www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist`.

[228] Zeerak Waseem. “Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter”. In: *Proceedings of the first workshop on NLP and computational social science*. 2016, pp. 138–142.

[229] Zeerak Waseem and Dirk Hovy. “Hateful symbols or hateful people? predictive features for hate speech detection on twitter”. In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[233] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. “Detection of abusive language: the problem of biased datasets”. 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)*. 2019, pp. 602–608.

[234] Deirdre Wilson. “The pragmatics of verbal irony: Echo or pretence?” In: *Lingua* 116.10 (2006), pp. 1722–1743.

[237] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex machina: Personal attacks seen at scale”. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 1391–1399.

[238] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. “Demoting Racial Bias in Hate Speech Detection”. In: *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 2020, pp. 7–14.

[242] Marcos Zampieri et al. “Predicting the Type and Target of Offensive Posts in Social Media”. 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). Association for Computational Linguistics, 2019, pp. 1415–1420.

[END OF MANUSCRIPT]

In the previous manuscript I leverage online community structures to create a novel corpus of abusive language that ensures diversity of content as well as opinion. I collect data from a variety of sources and assemble a demographically diverse cohort of annotators to label it. This corpus serves as a strong baseline to assess for the validity of abuse detection systems as it can elicit overfitting on pejorative language.

Despite being a valuable resource for abusive language research, the `Slur-Corpus` 1) uses three slurs for data collection, restricting its focus to three marginalized groups, and 2) collects data from Reddit. These design decisions were logistical necessities. The sheer diversity of abusive language makes it impossible to build a single resource that encompasses it all.

However these design designs do limit the overall corpus. Abusive language affects groups outside the three I focus on. Different marginalized groups follow different linguistic styles. Furthermore, abusive language is present and rampant on social media platforms other than Reddit. Online platforms themselves enforce different social and linguistic norms, through user interface and platform policy. Therefore, changes in either of two design decisions can affect the final corpus. However, there is little understanding on the after effects of these design decisions.

The taxonomy provides a formal system for a fine-grained analysis of abusive language. As the next course of action, I perform a broad analysis of how design decisions affect the final distribution of an abusive language corpus.

CHAPTER 5
Abusive Language across Slurs and Platforms

In the previous chapter I construct an abusive language corpus by filtering for pejorative expressions on Reddit. This general procedure - keyword selection → filtering data from online platforms → human annotation - is the norm across abusive language research. The resulting corpora are defined by two key design elements - the keywords and the online platform. Despite the popularity of this approach, we have no insight in how these design elements can shape the outcome. Can we still expect a similar corpus if we choose a different set of keywords? How do platforms themselves influence the collected data? We have no concrete answers to such questions.

Abusive language research is largely focused on building solutions to detect abuse. Surprisingly, as critical as data is to machine learning solutions, there is limited research that actually focuses on the construction of abusive language resources themselves. The following manuscript expands on the contributions of Chapter 4 and is the first attempt at systematic analysis of abusive corpus construction.

I specifically study the effects of two key design decisions - keywords and source platforms - on an abusive language corpus. Using the slur taxonomy, I perform fine-grained analysis of user comments collected from three different online platforms using nine different slurs. This broad analysis provides insights into the heterogeneous nature of abusive language corpora.

5.1 Manuscript 3: Composition of an Abusive Language Corpus : Decisions Matter

Authors: Haji Mohammad Saleem and Derek Ruths

In submission

5.1.1 Abstract

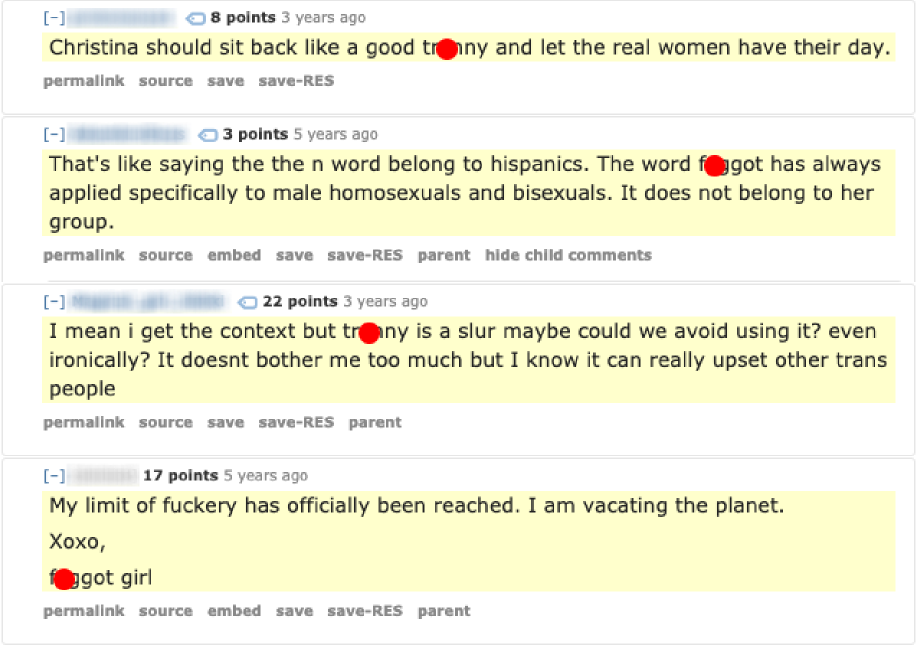
The prevalence of abusive language online has inspired broad research dedicated to combating it. An essential component of this research is access to representative samples of abusive speech. Since proprietary data on media platforms is not made public, researchers rely on filtering content (using keywords often in the form of slurs) followed by human annotation as a norm to collect abusive language. These resulting corpora play a central role in downstream research. While some issues have been identified in existing abusive language datasets, there is no research that systematically analyzes the crucial task of constructing a corpus in the first place. A fine-grained understanding of these resources and their creation simply does not exist. In this paper, we seek to bridge this gap by studying the role of two critical factors in the design of a dataset: (1) the slurs used to filter for abusive language and (2) the source where abusive language is filtered from. We perform a broad analysis by selecting nine slurs that derogate different communities and use them to filter comments from three distinct social media platforms. The collected comments were coded by an expert following established guidelines to determine the nature of slur usage. Overall, we find our corpus to be extremely heterogeneous, with almost a quarter of the comments labelled non-derogatory. We discover that both the slur and the source induce dramatic differences in the distribution of collected language, both in terms of amount and type of derogatory language. While the amount of

derogatory content can be influenced through targeted community sampling, the type of language obtained depends heavily on the slur itself. We also determine that non-trivial amounts of reclamation and counter speech are systematically present, both of which are actually used to diffuse, disarm, or deflect abusive language. On the other hand pornography was identified as one of the major sources of abusive content, with open questions as to how it should be handled by the research community at large. Our findings demonstrate that abusive language corpora are not homogeneous and and require thoughtful design.

5.1.2 Introduction

The proliferation and negative societal impacts of abusive language online have stimulated diverse and urgent research on the topic. Prior and ongoing work include characterization of abusive content online [130, 230], platform design interventions [241, 186, 21] and algorithmic detection methods [167, 175]. Though diverse in their aims and means, all research on abusive language online shares a common need for data - specifically representative samples of the abusive language being considered in the study.

A variety of factors conspire to make the collection of representative abusive language content difficult and prohibitively time intensive. Foremost, abusive content, while entirely too common, constitutes a very small portion of the volume of posts on any given platform. Filters are needed to collect such posts — which yields an enduring and problematic paradox in the abusive language research literature: how does one filter for abusive speech in order to design filters to detect it?



[–] [redacted] 8 points 3 years ago

Christina should sit back like a good tr[redacted]nny and let the real women have their day.

[–] [redacted] 3 points 5 years ago

That's like saying the the n word belong to hispanics. The word f[redacted]ggot has always applied specifically to male homosexuals and bisexuals. It does not belong to her group.

[–] [redacted] 22 points 3 years ago

I mean i get the context but tr[redacted]nny is a slur maybe could we avoid using it? even ironically? It doesnt bother me too much but I know it can really upset other trans people

[–] [redacted] 17 points 5 years ago

My limit of fuckery has officially been reached. I am vacating the planet.

Xoxo,

f[redacted]ggot girl

permalink source save save-RES
permalink source embed save save-RES parent hide child comments
permalink source save save-RES parent
permalink source embed save save-RES parent

Figure 5–1: Example user comments collected from Reddit that illustrate the diversity of slurring language. The first comment is derogatory while the others are different forms of non-derogatory comments, specifically discussion, counter speech and reclamation respectively.

The practical solution adopted, essentially everywhere, across the literature is to use pejoratives, in the form of slur words to identify posts with high likelihood to be derogatory. With rare exception, the process to produce a corpus is: (1) choose a slur word, (2) collect posts containing this word, (3) possibly have human coders perform a second pass filter. Despite systemic issues that emerge from biased filtering [60, 176, 196, 233], it endures as the main corpus creation method for the simple fact that, to date, there are no credible alternative approaches known, without the access to proprietary data such as moderator logs and user reports.

Despite this reality, the question remains: how credible are these slur-based corpora? What specific issues do they systematically have? Even if a slur-based corpus is the best we can do, answers to these questions can help researchers and developers in this space deliver more accurate findings and more responsible tools. Answering these questions requires, foremost, an understanding of what content is actually returned by slur-based filters: by which we mean, what different kinds of speech contains a slur word? For example slurs can be used in blatant hate speech, be reclaimed by the target community, or be part of counter speech calling out objectionable behaviour or just being discussed. We provide examples of user comments collected from Reddit in Figure 5–1 to further illustrate the differences. These examples are semantically distinct and use the same slur for different purposes, including some that are not abusive. We, therefore, need to understand the extent to which inclusion of different speech types in a corpus poses a challenge to the validity of online abusive language research and tool building. While there has been some work showing the limitations of keyword-based corpora[194], to our knowledge there have been no studies that take a fine-grained view of what type of content actually appears in such a corpus and what the broader implications are of ignoring these compositional details.

The purpose of our work is to provide the first broad and detailed assessment of what kind of speech a slur-based corpus contains. At the outset, we fully acknowledge that the space of factors influencing the composition of a corpus is too expansive for us to claim a comprehensive assessment. Our goal here is to provide a foundational study that (1) provides motivation and a methodological framework for assessing the composition of slur-based corpora and (2) assesses two key factors (the slur word itself and the platform on which it is used) that influence the composition of speech types in slur-based corpora. To do this, we collect data using several slurs: nine different pejorative words across three different platforms. We employ a functional taxonomy for slur-containing language to characterize what types of speech are in the collected corpus [130].

Among our findings, we discover that the slur word and platform can induce very different distributions of speech type. We observe that different slurs induce different kinds of derogatory content - which points to one reason classifiers built on slur datasets may struggle. We also find that the kind of non-derogatory content collected varies dramatically across slurs. The functional taxonomy we employ allows us to connect these findings to specific societal harms - which underscores the urgent need to accommodate for these distributional differences in research and tool development.

Overall our findings underscore how complex and varied slur-based corpora are - and how important it is to take into account corpus composition in order to ensure research validity and minimize unintentional societal harms of the research.

5.1.3 Abusive language is heterogeneous

Contemporary abusive language corpora are often treated as homogeneous within the dichotomy of derogatory and non-derogatory content. Most studies attempt to detect abusive content in a generic manner. However, treating such language as uniform in nature is a major oversight that can have unintended consequences. In reality, abusive content is varied. Researchers who claim to be addressing abusive language in general are actually addressing specific forms of abuse, based on their

	DEG	NDNA	APR	HOM
Derogatory force	Yes	Yes	Diminished	None
Derogatory use	Yes	No	No	No

Table 5–1: The four major categories of slur usage differ from each other based on the derogatory force of word and its usage for derogation.

specific corpus. The corpora, defined by the parameters of biased sampling focus on specific types or targets of abuse. For example, while advertised as abusive language, some datasets may only include examples of racism and sexism [61, 229]. Abusive content does not generalize across a variety of targets [194]. Furthermore, failure to acknowledge corpus characteristics can introduce racial and gender bias in detection algorithms [60, 196, 176] or harmless confounders highly correlated to abuse [233]. Therefore, it is crucial to understand the nuances of abuse language corpora in order to properly identify any unintended biases or limitations that might be introduced.

Our research aims to add such nuance by rooting itself in sociological literature around slurs to illustrate the variability in speech acts that contains slurs. We employ the taxonomy introduced by Kurrek, Saleem, and Ruths [130] which provides a fine grained distinction between the derogatory and non-derogatory usage of slurs. Our work is a practical application of this taxonomy and helps shed light on the diversity of language within an abusive language corpus and the implications of such diversity for research and its application in the real world. Slurs are the vehicles of derogatory force due to their capacity to impact their intended targets in deep and explosive ways [101]. However, not every usage of the slur brings harm to the target community. The taxonomy breaks down slur usage in four major classes based on derogatory variance (Table 5–1).

Derogatory (DEG) The derogatory force of a slur derogates the targeted individuals or groups. This derogation is independent of the intention or the attitude of the speaker. The usage of the slur serves to uphold the power structure, normalizing hateful attitudes and harmful discriminatory practices towards its target.

Non-Derogatory Non-Appropriative (NDNA) While the slur retains its autonomous derogatory force, it is used in a meaningful non-derogatory manner. Slurs, while still derogatory, do not directly derogate their intended targets. Such uses are often pedagogical / counter-speech where the derogatory nature of the word is recognized but not weaponized.

Appropriative (APR) Groups and individuals that serve as canonical targets of a slur may re-purpose it to take back control of the word and alter its power structure. Appropriation happens within the target group and serves to foster camaraderie and group membership. Appropriation works towards transforming the meaning of the slur to lessen or to eliminate its derogatory force. However, not all in-group uses are appropriative and usage in anger, contempt, or self-loathing can still be derogatory.

Homonyms (HOM) Some slurs have alternate meanings in the form of homonyms. In such usages, the word does not retain any derogatory force. Some slurs are part of proper nouns and are used without the canonical derogation. Finally some slurs might appear in a different language as a false friend, where they share the form but not the meaning. For example, the phrase "*ch*nk* in the armour", American actor "Dick Van *D*ke*", Swedish word "*sl*t*" which translates to end.

DEROGATORY	
Attribution	attributing the negative connotation of a slur to a target. *god, youre such a f*ggot anon*
Community-Focus	using a slur to directly refer to the target community. *Alison Brie is ALMOST white, but still making me lean towards liking k*ke women better*
Self-Deprecation	self directed derogation by the speaker. *I feel like a dirty karma wh*re*
Sexualization	derogation and sexualization of target through a slur. *Wh*re Whipped Into Submission* <URL>
Stand-Alone	slurs by themselves have derogatory autonomy and perpetuate derogation. *b*tch*
NON-DEROGATORY NON-APPROPRIATIVE	
Counter-Speech	speech acts condemning abusive behaviour and derogatory usage of slurs. *Yeah, fuck you for sl*t shaming. It's just sex; people use apps and take nude selfies these days, including the heteros. A simple copy and paste of the article title would have sufficed.*
Direct-Quotations	reiteration of quotes that contain a slur. *"I think one man is just as good as another so long as he's not a n*gger or a Chinaman" -Harry Truman*
Discussion	pedagogical discussions around slurs without derogating intended targets. *I'm trans, and using the term tr*nny to abbreviate transmission doesn't bother me in the slightest.It's old as hell. Now when people get clever with homonyms, sure, get mad.*
Recollection	recalling a past incident of someone else using a slur. *Tonight I've been called a f*ggot & a ni**er by a Trumpster.*
Sarcasm	using a slur in a sarcastic context without derogating intended targets. *Yeah because apparently f*ggot all of a sudden isn't a slur used against homosexuals*
APPROPRIATIVE	
Reclamation	*Ch*nk here... Highly recommend soy sauce and tea eggs!*
HOMONYM	
Homonyms	*I listened to Van D*ke Parks's Song Cycle yesterday and I was both amazed and confused.*

Table 5–2: Derogatory and NDNA usages of slurs are further divided into five categories each. We provide a brief description of each sub-category as well as examples to illustrate their semantic differences. These are actual examples from our corpus.

The four major categories are semantically distinct in the nature of the derogatory force of the slur word and whether that force is channelled towards the derogation of a target. The differences are illustrated in Table 5–1. We provide details on the further sub-division of the derogatory and NDNA categories in Table 5–2 along with examples of each. The nuanced differences between these categories are not only semantic, but can lead to syntactic differences as well. We further elucidate these differences using additional comments from our corpus. For example, Community-focused usage of slurs substitutes the target community with the slur. In the following comment: "Get wanked off by a *ch*nk* in Chinatown all the massage parlours there are fronts." the slur directly refers to a Chinese person. On the other hand, in "Don't be such a *ch*nk* AutoModerator.", the speaker is attributing the negative properties of the slur to AutoModerator. Similarly, sexualized usage of slurs consists of its own distinct language style. Overall, the taxonomy provides a total of 12 subcategories based on how the slurs are being deployed by users on online social media. Each subcategory adds nuance to how the derogatory nature of slurs is being operationalized to abuse the targeted communities or discarded to limit such abuse. Furthermore, all these categories are not equal and require thoughtful interaction catering to individual characteristics. Mishandling of different categories can have different implications. Erroneous mislabeling of appropriative language has far greater consequences than that of homonyms, since appropriative language is a tool for communities facing abuse to alter the power structure. Similarly, it is also important to preserve usage of slurs in counter-speech due to its role in moderating online abuse. Researchers need to be aware of these intrinsic distributions within their corpora to ensure the validity of their methods and analyses.

Misogynistic	**Racial & Ethnic**	**LGBT-related**
*b*tch*	*ch*nk*	*d*ke*
*sl*t*	*k*ke*	*f*ggot*
*wh*re*	*n*gger*	*tr*nny*

Table 5–3: The list of slurs that were used for data collection. Note: given the nature of the these words, we censor all occurrences of them in this paper.

5.1.4 Abusive language collection and annotation

The construction of an abusive language corpus is a labour intensive task and requires researchers to make numerous design decisions throughout the process. Given the variable and nuanced nature of derogatory as well as non-derogatory language, these key decisions can dramatically influence the resulting corpus. Currently, as a research community, we know little of just how these influences might work. Corpora play a crucial role in the sphere of abusive language research and we seek to further our understanding of their creation process. We study how different key decisions yield different kinds of speech in a corpus. Specifically we investigate the role of two key factors: (1) the kind of abusive language that should be included in the corpus and (2) the source from which such abusive content should be collected . The former determines the social groups that are targeted by abusive language and means to capture the language that targets them. The latter includes identifying online communities and platforms where such language might be prevalent and available. Together, these two particular factors play a large part in defining the resulting corpus, for example a Twitter corpus for sexism and racism [229].

In this study, we replicate the standard practise in the abusive language literature for corpus creation. Abusive language varies across targets. Therefore, we assemble a diverse corpus of derogatory language that targets multiple social groups, using a variety of slurs. We choose to focus on three major types of online abuse: misogyny, racial and ethnic derogation, and LGBT-related abuse. We select slurs which target social groups from each class. Our choice of slurs is based on those commonly studied in academic literature [101, 55, 108, 124, 98, 9]. In total we select nine separate slurs which target seven distinct social groups (women, Asian people, Jewish people, Black people, lesbians, gays, and transgender people). The slurs are overt with well established meanings in mainstream North American culture. The full list is presented in Table 5–3.

Typically a project researching abusive language online will focus on a singular online platform [167, 61, 229] in its analysis. However, language across platforms, abusive or otherwise, is not identical. Social platforms affect the language produced through their features, norms and guidelines. For example, Twitter only allows short sentences through their character limit[104]. Similarly, online communities exhibit linguistic norms as a reflection of socialization[132, 166], making language a function of the community it originates from. Platform guidelines control what kind of content is allowed and therefore produced on a platform [112]. In order to determine the extent to which platforms might affect abuse corpora, we choose three different social media platforms. They platforms are popular both amongst users as well as researchers. We chose these platforms because they are commonly used in the literature, it was possible to access data through APIs / scrapers, and due to their functional differences as described below.

4chan. A popular message board with a dedicated community of users. Compared to other platforms that push for persistent user identity, 4chan maintains a culture of anonymity [19]. The platform is split into 70 topic-based sub-communities called boards. In this paper, we focus on two of the popular boards: (i) *Random or /b* - 4chan's first and most popular board and (ii) *Politically Incorrect or /pol* - the board for discussion of news, world events, political issues, and other related topics. We scraped these two boards for comments that contain the relevant slurs. We specifically chose the two boards as they are known to be toxic. The 4chan portion of our corpus thus represents data extraction from extremely antagonistic communities.

Reddit. An aggregator of user-submitted text, links, photos, and videos. Reddit has thousands of sub-communities called *subreddits* on a wide and growing range of topics: from world news headlines, to animal GIFs, to fan forums. The popularity and prominence of material on the site is determined by voting from the Reddit community. Users need pseudo-anonymous usernames to create content. We use the Pushshift Reddit dataset [15] and create a random sample of comments that contain the relevant slurs. The Reddit portion of our corpus represents data extraction across diverse communities.

Twitter. A popular microblogging service. Twitter users follow other users to access content created by these users. The relationship of following and being followed requires no reciprocation. Unlike the past two platforms, Twitter does not have sub-communities. The platform does not require users to provide their real names and can be accessed through a unique pseudo-anonymous username. However, it is common practice on the platform to create accounts with identifiable information [179]. We used the Twitter API to collect tweets that contain the relevant slurs. Twitter represents data collection in absence of self-identifying communities.

From the filtered data, we randomly sample 100 comments for each slur, for each of the three platforms, resulting in a total of 2700 comments. This allows us to assemble a corpus which is diverse in targets as well as the sources it is derived from. We go on to annotate our corpus with fine grained labels, as described in the previous section. Note that we understand the limitations of a single annotator. However, annotations with such fine-grained understanding of slur usage required expert knowledge and access to multiple experts was not feasible. Our analysis allow us to understand the compositional distributions that emerge from varying choices of slur and source.

When we aggregate these annotations by platform or slur, we can observe a distribution of counts of posts over the taxonomic categories. This distribution reveals what kind of language is present in the collection - and is the central focus of this study. Hereafter, we call this the *taxonomic distribution* of the platform, slur, or whatever grouping is being applied.

5.1.5 Results

Following the manually intensive annotation task we tabulate the expert labels. These labels are much more fine-grained than contemporary research. We seek to study the distribution of language that emerges through systematic biased sampling. We present our observations here.

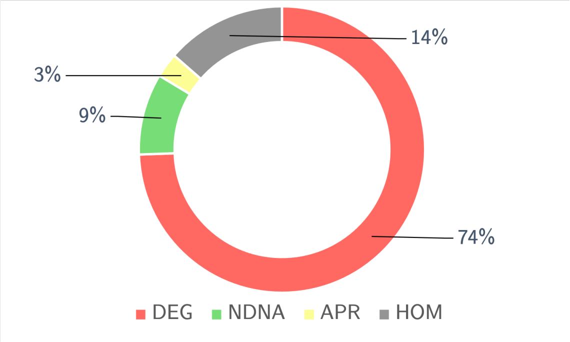


Figure 5–2: Breakdown of the macro categories of ways slurs are used in our corpus. The vast majority of the slur usages were derogatory. We find significant Homonym usage as well.

A vast majority of comments are derogatory

In Figure 5–2, we present the overall breakdown of the 4 major categories of slur usage in our corpus. Almost three quarters of the comments use slurs in a derogatory manner. This observation is not entirely surprising and certainly the reason keyword based sampling is popular. Pure random sampling would only manage to capture tiny proportions of abusive language [78]. On the flip side, a quarter of the comments were labelled non-derogatory, in various forms. The portion of non-derogatory comments is as high as 40% on platforms where we sampled comments randomly (Figure 5–3). Even though slurs are primarily viewed as vehicles of derogation, they are frequently used in non-derogatory contexts. An imbalanced corpus can cause algorithms to over-fit on such lexical cues [233]. Researchers should make sure that the volume and variety of non-derogatory content is great enough for detection systems to avoid overfitting.

		Non Derogatory	
Derogatory		Non Appropriative	
Attribution	1184	Counter-Speech	56
Community-Focus	486	Direct-Quotations	30
Self-Deprecation	28	Discussion	86
Sexualization	249	Recollection	67
Stand-Alone	63	Sarcasm	9
Total	2010	*Total*	248
Appropriative		Homonym	
Reclamation	75	Homonyms	367

Table 5–4: Breakdown of the micro categories of ways slurs are used in our corpus. Apart from macro observations, we find that slurs are quite often used to sexualize the target. While using slurs to refer to the communities they target was common, attributing a slur to someone as an insult was the most prevalent form of slur usage.

Attribution is the most common form of slur usage.

Slurs can be used to derogate their targets in a variety of ways. We present our observations in Table 5–4. About a quarter of the derogatory comments were community-focused. In these cases, slurs directly refer to their canonical target communities. Canonical target of a slur is the member of a community that a slur traditionally derogates. For example, *f*ggot* derogates gay men while *n*gger* derogates black people. In the comment provided in Table 5–2, the slur directly refers to Jewish women. In contrast, more than half of the derogatory comments used slurs in an attributive manner. In attribution, slurs are used as insults and the inherent derogatory nature of the slur is projected onto the targeted person. The targeted person may or may not be the canonical target of the slur. Using the comment in Table 5–2, the target of the slur in the comment is not necessarily a gay man.

While these two subcategories cover most of the derogatory comments, they are semantically distinct. Community focus will always refer to the targeted community at large but through attribution anyone can be derogated using a slur which can create complex constructs. For example, if *f*ggot* is used against a hetero-normative person, it insults the target while also derogating gay men.

Sexualization is a significant portion of slur usage.

There was an abundance of sexualized content in our corpus. Almost 10% of the comments used slurs for sexualization of the target, especially in pornography. Slurs, namely *tr*nny*, *wh*re*, and *sl*t*, are often used in descriptive titles for porn. When such content ends up on mainstream social media, it is likely to be picked up in keyword-based data collection. We further observe that porn titles often share phrases with instances of derogatory language. In Table 5–5 we provide examples of similar usage of slurs being in pornographic and non-pornographic contexts. Even though pornographic content is similar to other instances of abusive language, the

Porn	*tr*nny* sex women being striped asleep
Non-Porn	oh sure you'll go skin deep but won't play *tr*nny* sex #autocorrect
Porn	Dirty Black *Sl*t* Gets A Deep Schlong Ramming ¡URL¿
Non-Porn	You're a dirty fucking *sl*t* who loves taking the cock you want every cock you see
Porn	Asian *Wh*re* Fucked And Sucks Cock For Mouthful Of Cum ¡URL¿
Non-Porn	I fucked a 90lb. asian *wh*re* today, literal massage parlor *wh*re,* super petite.

Table 5–5: Examples of slurs being used in pornographic and non-pornographic context. We find the usage slurs in pornographic context can mimic usage of slurs in non-pornographic context. The first set of examples both talk about "*tr*nny* sex". Similarly the second set of examples refer to dirty "*sl*t*" and the final set refers to "asian *wh*res*".

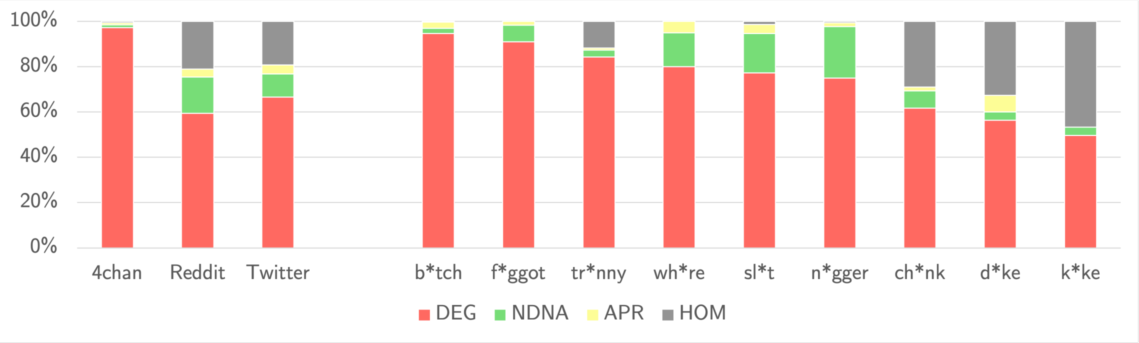


Figure 5–3: Breakdown of the macro categories of slur usage across platform and slur word. Both the choices affect the final composition of the corpus. Sourcing data from an extremely antagonistic community led to a very high amount of derogatory comments. While some slurs are used entirely in a derogatory manner, others have high homonym usages.

research community has been silent on whether it should be included in abusive language corpus. Given the plethora of porn on mainstream social media, this is an important design consideration with important social implications.

Derogatory composition changes with community sampling.

The notion of community sampling has been established in previous research [194]. We observe further evidence of this as presented in Figure 5–3. For Reddit and Twitter, we randomly sampled from the entire platform. However, for 4chan, we performed targeted sampling entirely from antagonistic communities (`/pol, /b`). While the taxonomic distribution (the proportion of different classes of slur usage) was similar on Reddit and Twitter, comments from 4chan were overwhelmingly derogatory. Specifically, almost 98% off all the comments collected from 4chan were found to be derogatory in nature, significantly higher compared to 60% and 67% on Reddit and

	4chan	Reddit	Twitter	b*tch	f*ggot	tr*nny	wh*re	sl*t	n*gger	ch*nk	d*ke	k*ke
DEROGATORY												
Attribution	501	365	318	249	251	56	181	110	131	51	86	69
Community	312	109	65	22	5	106	22	19	54	115	68	75
Self-Deprecation	3	18	7	6	7	3	6	1	3	1	0	1
Sexualization	47	24	178	4	1	87	31	100	14	0	12	0
Stand-Alone	13	19	31	3	9	1	0	2	23	18	3	4
NON DEROGATORY NON APPROPRIATIVE												
Counter-Speech	3	20	33	2	1	3	22	19	2	4	1	2
Direct-Quotations	1	21	8	3	2	2	3	7	7	3	2	1
Discussion	3	61	22	1	12	4	7	10	34	8	4	6
Recollection	2	37	28	1	6	0	12	13	23	7	3	2
Sarcasm	2	5	2	0	1	0	1	3	2	1	1	0
APPROPRIATIVE												
Reclamation	8	32	35	8	5	3	15	12	5	5	22	0
HOMONYM												
Homonyms	5	189	173	1	0	35	0	4	2	87	98	140

Table 5–6: Breakdown of the macro categories of slur usage across platform and slur word. We observe very little presence of non-derogatory usage of slurs on 4chan. Some slurs are used entirely as attributive insults and others often to sexualize.

Twitter respectively. It stands to reason that similarly the portion of non derogatory content can be directly affected by sampling from supportive communities.

Individual slurs dictate the final composition.

The intrinsic composition of a corpus is heavily dependent on the characteristics of the slurs for its construction. The slurs _b*tch_ and _f*ggot_ are used entirely in an attributive manner (84% of all comments - Table 5–6). Other slurs have more prominent use in referring to the targeted community (_tr*nny_ and _ch*nk_). Similarly, some slurs are more likely to be used in sexualized contexts, as we mention in a previous subsection. Thus different slurs are likely to filter different kinds of derogatory content, affecting the composition of the resulting corpus.

Homonym and appropriation vary drastically across slurs.

A community can reclaim a slur for non-derogatory purposes, to show a sense of intimacy and solidarity. Reclamation can also be a deliberate tool to subvert institutionalized norms [20]. However, reclamation relies on efforts made by targeted community. For example, *queer* is widely considered to be reclaimed within the LGBT+ community [109] while *f*ggot* still remains largely homophobic [39]. While *n*gger* maintains its pejorative meaning, it is used widely by the black community within their conversation or even against oppressors [122, 210]. Thus, the presence of reclamation is entirely dependent on the choice of the slur and how reclaimed it is.

Similarly, homonyms are dependents on slurs, where some slurs might have homonyms while others may not. Homonyms have 4 major sources: (1) alternate meanings: *tr*nny* → transmission, *f*ggot* → meatballs, *ch*nk* → a crack; (2) proper names: *d*ke* → Dick Van Dyke - American actor, *k*ke* → Kike Hernandez - professional baseball player; (3) false friends: *sl*t* → Swedish word for end; (4) misspellings: *k*ke* → like. Homonym usages are therefore incidental on the slur in question.

5.1.6 Discussion

This study surfaced three high-level findings that should inform any effort to build or use an online abusive language corpus.

Different slurs/platforms, different datasets. Foremost, we found that the platform and choice of slur word resulted in different taxonomic distributions. For example, sourcing from just antagonistic communities of 4chan resulted in 98% derogatory content. On the other hand, 48% of the comments filtered using *k*ke* were actually homonyms. Taxonomic difference were even more noticeable for sub categories of both derogatory and non-derogatory content (Table 5–4). This establishes the most fundamental point of our study: that simply using the same slur-based filtering process does not mean that we obtain similar or comparable corpora.

On the surface, this may not be surprising - one of our sources was a 4chan board known for derogatory and inflamatory languages [19]. Worth noting, prior work has also established the taxonomic distributions can even differ between subreddits [130]. Further, that different slurs yield different kinds of content can seem rather intuitive. While this all may seem obvious, many research studies seem to operate on the opposite assumption. These employ multiple slur words (e.g., [229, 8]) and treat the resulting content as one homogeneous dataset. In practical terms, this signals an implicit belief that either taxonomic distributions do not differ or that taxonomic distributions do not matter to the modeling or analytical work. Our first finding establishes that taxonomic distributions *do* differ. And, as we discuss next, our study provides strong indications that the taxonomic distributions matter.

Different slurs, different kinds of derogatory content. We observed that different slurs yielded marked differences in the kind of offensive content collected. In particular, slurs differed in which derogatory category tended to occur the most frequently. For example, “b*tch”, when derogatory, was almost always used via attribution (which amounts to a personal attack). Attribution accounted for over 90% of the derogatory comments. In contrast, usages of “slut” - another misogynistic slur - were evenly split between attribution and sexualization (objectifying women in general). Specifically, attribution and sexualization account for 47% and 43% of the derogatory comments respectively. Consider that the syntax, supporting vocabulary, and underlying intent of sexualization and attribution can be quite different. Both these slurs are part of *hatebase*, a popular lexicon used build abusive corpora [61]. The implication of this is that analyses or modeling exercises are being run on quite different kinds of offensive content. At the very least, the mixing of derogatory content types from different slur words can skew the priors learned by models. At the worst, this unintentional mixing of derogatory content types makes for a much harder classification problem. A promising direction for future work is to consider the performance gains to be had by modeling these different derogatory content types separately.

Different slurs, different kinds of non-derogatory content. We see similar distributional differences in the non-derogatory content categories. The gender and sexuality slurs provide a compelling example here. Non-derogatory uses of the slur “tr*nny” are overwhelmingly due to homonyms (mostly people talking about car parts). In fact, 75% of the non derogatory comments were homonyms. In contrast, non-derogatory uses of the term “n*gger” are largely due to discussion of the slur itself or recollection of past abuse by the victims. The two categories account for 45% and 30% of the non-derogatory usage of the slur respectively. Failing to acknowledge that these non-derogatory types are different has both practical and ethical implications.

The practical implications follow along similar lines as was discussed earlier: different content type distributions can skew a model's underlying priors and, worse, complicate the modeling task itself (the model now needs to learn the markers for two very different kinds of non-derogatory uses). Even more than attribution and sexualization, homonym and discussion instances of non-derogatory slur use look *very* different.

The ethical implications of failing to differentiate non-derogatory content relate to the unintended consequences of false positives. Most modeling exercises report F1 scores. However, an F1 score doesn't capture the very different societal impacts of deploying a classifier that accidentally flags reclamation compared to one that accidentally flags discussion compared to one that accidentally flags homonyms. Reclamation is an act of self-empowerment by the target community - false positive effects on reclamation deprive target communities of this important tool. Discussion of slurs can serve informal educational or deliberative purposes. As a result, false positive effects on discussion can shut down organic efforts among users to improve their understanding of the impact of and responses to slurs. As researchers, we must not make the mistake of thinking that all false positives are equal in their impact. Our finding of the distributional differences between slurs underscores the need for greater awareness of these in the preparation and use of slur-based datasets.

Overall, we consider the most important insight of this study to be that using the same slur-based filtering process on different slurs (or platforms) doesn't produce similar or comparable datasets. As we have discussed, those differences will have an impact on both the quality of methods AND the unintended and negative societal impacts methods could have. When we mix slurs without an understanding of their taxonomic distributions, we open our studies and our methods to serious risks to their validity. While robust alternatives to slur-based collection methods may not exist, mapping the content type distributions of datasets is a promising tool for ensuring that we better understand the task complexity and false positive implications of our work.

5.1.7 Related Works

Our study is rooted in two distinct areas of research: the first, sociolinguistics, depicts slurs as institutional phenomena; and the second, computational literature, isolates the role of slurs in the creation of abusive language corpora.

Variation in slurs

A slur is defined in sociology and linguistics as a word or expression that can be used to demean targets based on their membership to a racial, ethnic, religious, gender, or sexual-orientation group [192]. While the subjectivity of derogation undermines agreement on the nature of comments that contain slurs, it is generally agreed upon that slurs possess a common capacity to dehumanize their targets [108]. However, this derogation is not a static phenomenon and rather varies in assorted ways. All slurs are not equally offensive and some are more offensive than others [25, 101, 6, 108, 110, 232]. Furthermore not all usages of the same slur convey contempt with equal force and any given usage impacts certain audience members more forcefully than others[54, 55, 100, 184]. Overall slurs demonstrate three types of variations in the offense that they generate as described by Popa-Wyatt and Wyatt [185]:

Word-variation: Slurs vary in the degree of offense they cause. This offense differs across slur words within community groups (intra-group variation) and across community groups exposed to the same slur words (inter-group variation).

Use-variation: Offense does not only vary across slurring words themselves. It also varies across uses of the same word based on the context and the speaker. In some cases, members of the target community even succeed in reclaiming the terms, capturing the power it previously encapsulated [101] [6].

Audience-variation: The same utterance of a slur offends different members of the audience differently. Therefore along with the speaker and the content, the audience also decides how offensive a slur is.

These three variations demonstrate that slurs generate offense through a complex process that depends on a multiple factors. Given the pivotal role that slurs play in abusive language research, a better understanding of their offense patterns is essential for the creation and future application of slur-based corpora.

Slurs and to abusive language corpora

Access to online discourse is central to studying abusive language online. Much of the discourse happens on proprietary platforms and collecting such data and isolating abusive examples is non-trivial. One way to gather such data is to utilize user reports. For example, in their analysis, Nobata et al. [167] use user-reported comments from Yahoo! Finance and News. However, this approach requires privileged access to reporting data, which is not publicly available. Another solution is to sample data from antagonistic communities [194, 84, 87, 209]. However, not all content generated in such communities is abusive.

In the majority of contemporary literature, researchers employ a combination of keywords and human annotation to create corpora for their studies. Waseem and Hovy [229] released a value-laden collection of 16K tweets annotated for hate speech. They bootstrapped the corpus collection, by searching for common slurs and terms pertaining to religious, sexual, gender, and ethnic minorities through the Twitter API and then manually annotate it. The authors mention their corpus also contains tweets that have clearly offensive words but remain non-offensive in their usage of these words. Davidson et al. [61] also released their dataset, for which they use *Hatebase.org* as their source for slur words and phrases. The authors then manually code 25K of the collected tweets, 5% of which are coded as hate speech. The authors go on to comment on the imprecision of the hatebase lexicon. Another valuable dataset has been released by Golbeck et al. [89] which contains 35K manually annotated tweets. The tweets are collected using 10 terms that contain derogatory terms for races and religions, offensive hashtags and offensive phrasing, followed by intensive coding. Multiple other studies have used similar methods [228, 72, 103]. Furthermore, 30% of the studies in the proceedings of 2018 Abusive Language Workshop use one or more of the aforementioned corpora [77, 153, 127, 3, 220, 226, 116].

Prevalence of slurs in other research areas

While we focus specifically on abusive language, our work has broader implications for research that studies online bad behaviour. The past decade has seen a significant increase in research on a wide-variety of online bad behaviours, including personal attacks [177, 237, 205], cyberbullying [158, 66, 209], harassment [240, 70], toxicity, and antisocial behaviour. While these forms of online bad behaviour have their own defining characteristics, we lack a clear dividing line between them. Their detection therefore brings forth similar challenges to those faced by abusive language researchers.

Manual coding for online incivility detection relies on a list of “offensive/mean words” [141]. The same list was used in the analysis of controversial comments [1], abusive user posts [42], offensive language [81, 106, 90, 170], and profanity in video game content [182]. Similarly, works that detect cyberbullying often use profanities to filter comments out [8] or in for analysis [188, 204]. Given the widespread usage of slurs in existing literature, across a wide variety of contexts, a better understanding of their capacity for offense is necessary not only for the detection and moderation of abusive language, but also for an overarching analysis of online bad behaviour.

Bibliography

[1] Aseel Addawood et al. “Telling apart tweets associated with controversial versus non-controversial topics”. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. 2017, pp. 32–41.

[3] Betty van Aken et al. “Challenges for Toxic Comment Classification: An In-Depth Error Analysis”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 33–42.

[6] Luvell Anderson and Ernie Lepore. “Slurring words”. In: *Noûs* 47.1 (2013), pp. 25–48.

[8] Zahra Ashktorab and Jessica Vitak. “Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 3895–3905.

[9] Lauren Ashwell. “Gendered slurs”. In: *Social Theory and Practice* 42.2 (2016), pp. 228–239.

[15] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[19] Michael S Bernstein et al. “4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.” In: *ICWSM*. 2011, pp. 50–57.

[20] Claudia Bianchi. “Slurs and appropriation: An echoic account”. In: *Journal of Pragmatics* 66 (2014), pp. 35–44.

[21] Michał Bilewicz et al. “Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment”. In: *Aggressive behavior* 47.3 (2021), pp. 260–266.

[25] Renée Jorgensen Bolinger. “The pragmatics of slurs”. In: *Noûs* 51.3 (2017), pp. 439–462.

[39] Allison T Casar. “Queer Stance: Metalinguistic attitudes towards slur reclamation among LGBTQ young adults”. In: *Lavender Languages AND Linguistic Conference* (2021).

[42] Hao Chen, Susan Mckeever, and Sarah Jane Delany. “Presenting a labelled dataset for real-time detection of abusive user posts”. In: *Proceedings of the International Conference on Web Intelligence*. ACM. 2017, pp. 884–890.

[54] Adam M Croom. “How to do things with slurs: Studies in the way of derogatory words”. In: *Language & Communication* 33.3 (2013), pp. 177–204.

[55] Adam M Croom. “Slurs”. In: *Language Sciences* 33.3 (2011), pp. 343–358.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019, pp. 25–35.

[61] Thomas Davidson et al. “Automated hate speech detection and the problem of offensive language”. In: *Eleventh international aaai conference on web and social media*. 2017.

[66] Karthik Dinakar, Roi Reichart, and Henry Lieberman. “Modeling the detection of textual cyberbullying”. In: *fifth international AAAI conference on weblogs and social media*. 2011.

[70] Maeve Duggan. *Online harassment*. Pew Research Center, 2014.

[72] Mai ElSherief et al. “Hate lingo: A target-based linguistic analysis of hate speech in social media”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[77] *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018.

[78] Antigoni Founta et al. “Large scale crowdsourcing and characterization of twitter abusive behavior”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 2018.

[81] Rolf Fredheim, Alfred Moore, and John Naughton. “Anonymity and online commenting: The broken windows effect and the end of drive-by commenting”. In: *Proceedings of the ACM web science conference*. 2015, pp. 1–8.

[84] Lei Gao and Ruihong Huang. “Detecting Online Hate Speech Using Context Aware Models”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., 2017, pp. 260–266.

[87] Ona de Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018, pp. 11–20.

[89] Jennifer Golbeck et al. “A large labeled corpus for online harassment research”. In: *Proceedings of the 2017 ACM on web science conference.* 2017, pp. 229–233.

[90] Viktor Golem, Mladen Karan, and Jan Šnajder. “Combining Shallow and Deep Learning for Aggressive Text Detection”. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018).* 2018, pp. 188–198.

[98] Amanda Haynes and Jennifer Schweppe. “STAD: Stop Transphobia and Discrimination Report”. In: *Transgender Equality Network Ireland* (2017).

[100] Christopher Hom. “Pejoratives”. In: *Philosophy compass* 5.2 (2010), pp. 164–185.

[101] Christopher Hom. “The semantics of racial epithets”. In: *The Journal of Philosophy* 105.8 (2008), pp. 416–440.

[103] Muhammad Okky Ibrohim and Indra Budi. “A dataset and preliminaries study for abusive language detection in Indonesian social media”. In: *Procedia Computer Science* 135 (2018), pp. 222–229.

[104] Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. “Brevity is the soul of Twitter: The constraint affordance and political discussion”. In: *Journal of Communication* 69.4 (2019), pp. 345–372.

[106] Myungha Jang and James Allan. “Explaining Controversy on Social Media via Stance Summarization”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval.* ACM. 2018, pp. 1221–1224.

[108] Robin Jeshion. “Expressivism and the Offensiveness of Slurs”. In: *Philosophical Perspectives* 27.1 (2013), pp. 231–259.

[109] Robin Jeshion. “Pride and Prejudiced: on the Reclamation of Slurs”. In: *Grazer Philosophische Studien* 97.1 (2020), pp. 106–137.

[110] Robin Jeshion. “Slurs and stereotypes”. In: *Analytic Philosophy* 54.3 (2013), pp. 314–329.

[112] Jialun’Aaron’ Jiang et al. “Characterizing Community Guidelines on Social Media Platforms”. In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing.* 2020, pp. 287–291.

[116] Mladen Karan and Jan Šnajder. “Cross-Domain Detection of Abusive Language Online”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2).* 2018, pp. 132–137.

[122] Randall Kennedy. *Nigger: The strange career of a troublesome word.* Vintage, 2008.

[124] Hannah Kia, Kinnon Ross MacKinnon, and Melissa Marie Legge. “In pursuit of change: Conceptualizing the social work response to LGBTQ microaggressions in health settings”. In: *Social work in health care* 55.10 (2016), pp. 806–825.

[127] Rohan Kshirsagar et al. “Predictive Embeddings for Hate Speech Detection on Twitter”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2).* 2018, pp. 26–32.

[130] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. “Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 138–149.

[132] Wan Shun Eva Lam. “Language socialization in online communities”. In: *Encyclopedia of language and education* 8.301 (2008), p. 11.

[141] Suman Kalyan Maity et al. “Opinion conflicts: An effective route to detect incivility in Twitter”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–27.

[153] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. “Neural Character-based Composition Models for Abuse Detection”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 1–10.

[158] Michael J Moore et al. “Anonymity and roles associated with aggressive posts in an online forum”. In: *Computers in Human Behavior* 28.3 (2012), pp. 861–867.

[166] Dong Nguyen and Carolyn Rose. “Language use as a reflection of socialization in online communities”. In: *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. 2011, pp. 76–85.

[167] Chikashi Nobata et al. “Abusive language detection in online user content”. In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 145–153.

[170] Mariam Nouh, Jason RC Nurse, and Michael Goldsmith. “Understanding the radical mind: Identifying signals to detect extremist content on twitter”. In:

2019 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE. 2019, pp. 98–103.

[175] Ji Ho Park and Pascale Fung. "One-step and Two-step Classification for Abusive Language Detection on Twitter". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 41–45.

[176] Ji Ho Park, Jamin Shin, and Pascale Fung. "Reducing Gender Bias in Abusive Language Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 2799–2804.

[177] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. "Deep Learning for User Comment Moderation". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 25–35.

[179] Sai Teja Peddinti, Keith W Ross, and Justin Cappos. "User Anonymity on Twitter". In: *IEEE Security & Privacy* 15.3 (2017), pp. 84–87.

[182] Quang Anh Phan and Vanessa Tan. "Play with bad words: A content analysis of profanity in video games". In: *Acta Ludica-International Journal of Game Studies* 1.1 (2017), pp. 7–30.

[184] Mihaela Popa-Wyatt. "Not All Slurs are Equal". In: *Phenomenology and Mind* 11 (2016), pp. 150–157.

[185] Mihaela Popa-Wyatt and Jeremy L Wyatt. "Slurs, roles and power". In: *Philosophical Studies* 175.11 (2018), pp. 2879–2906.

[186] Jing Qian et al. "A Benchmark Dataset for Learning to Intervene in Online Hate Speech". 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). Association for Computational Linguistics, 2019, pp. 4757–4766.

[188] Rahat Ibn Rafiq et al. “Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network”. In: *Social Network Analysis and Mining* 6.1 (2016), p. 88.

[192] Katherine Ritchie. “Social Identity, Indexicality, and the Appropriation of Slurs”. In: *Croatian Journal of Philosophy* 17.2 (50) (2017), pp. 155–180.

[194] Haji Mohammad Saleem et al. “A web of hate: Tackling hateful speech in online social spaces”. In: 2016.

[196] Maarten Sap et al. “The risk of racial bias in hate speech detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1668–1678.

[204] Devin Soni and Vivek K Singh. “See no evil, hear no evil: Audio-visual-textual cyberbullying detection”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–26.

[205] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. “Automatic identification of personal insults on social news sites”. In: *Journal of the American Society for Information Science and Technology* 63.2 (2012), pp. 270–285.

[209] Rachele Sprugnoli et al. “Creating a whatsapp dataset to study pre-teen cyberbullying”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 51–59.

[210] Kameron Johnston St Clare et al. “Linguistic Disarmament: On How Hate Speech Functions, the Way Hate Words Can Be Reclaimed, and Why We Must

Pursue Their Reclamation". In: *Linguistic and Philosophical Investigations* 17 (2018), pp. 79–109.

[220] Elise Fehn Unsvåg and Björn Gambäck. "The Effects of User Features on Twitter Hate Speech Detection". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 75–85.

[226] Cindy Wang. "Interpreting Neural Network Hate Speech Classifiers". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 86–92.

[228] Zeerak Waseem. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter". In: *Proceedings of the first workshop on NLP and computational social science*. 2016, pp. 138–142.

[229] Zeerak Waseem and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[230] Zeerak Waseem et al. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 78–84.

[232] Daniel Whiting. "It's not what you said, it's the way you said it: slurs and conventional implicatures". In: *Analytic Philosophy* 54.3 (2013), pp. 364–377.

[233] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. "Detection of abusive language: the problem of biased datasets". 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). 2019, pp. 602–608.

[237] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex machina: Personal attacks seen at scale”. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 1391–1399.

[240] Dawei Yin et al. “Detection of harassment on web 2.0”. In: *Proceedings of the Content Analysis in the WEB* 2 (2009), pp. 1–7.

[241] Li-Yin Young. “The effect of moderator bots on abusive language use”. In: *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*. 2018, pp. 133–137.

[END OF MANUSCRIPT]

This work squarely focuses on the developing a broader understanding around the critical but unexamined task of constructing abusive language resources. The vast majority of these language resources follow similar design paradigms that include two defining choices - keyword filters and source platforms. Through this analysis, I uncover that language resources that follow this paradigm are not equivalent, despite the shared construction process. Both the choice of how to filter for abusive language and where to filter it from, can and do affect the outcome. Different choices for either affects the amount as well as the kind of abusive language we capture - another reelection of the diversity problem in abusive language. Thus the resulting corpora are heterogeneous and systems trained on one resource can behave differently when tested on another, raising concerns about the generalizability of abusive language detection frameworks.

CHAPTER 6
Community Context in Abuse Detection

The past chapters focus on problem of diversity in abusive language - that is - abusive content takes a variety of forms. To this end, I build a taxonomy that provides precise terminology to address content diversity and construct a corpus with diverse perspectives around pejorative expressions. Combined, they allow for a more honest engagement with abusive language.

Equipped with these tools, I now tackle detection of abusive language. Contemporary research highlights the issue of bias in detection frameworks through systematic misclassificaiton of conversations from marginalized populations as abusive. Such biases are likely due to faulty correlations learnt by detection frameworks, leading to false-positive errors. We can overcome some of these biases by providing detection frameworks: (1) access to marginalized perspectives to help rectify faulty correlations and (2) situational information that can assist in additional clarification.

The corpus I introduce in Chapter 4 sources content from supportive communities and thus already contains broad perspectives of marginalized populations. In this chapter I focus on further reducing false-positive errors in abuse detection through contextualization. Additional information such as user -history and -network has been shown to improve performance of detection frameworks. However, such information fails to account for how users behave as a whole. Since these conversations are happening within online communities, integrating this social context can help reduce ambiguity in abusive language detection.

My corpus was collected from wide variety of communities, which puts me in a unique position to conduct a series of experiments that assess the significance of community content in abuse detection.

6.1 Manuscript 4: Enriching Abusive Language Detection with Community Context

Authors: Haji Mohammad Saleem, Jana Kurrek, and Derek Ruths

In submission

6.1.1 Abstract

Uses of pejorative expressions can be benign or actively empowering. When models for abuse detection misclassify such expressions as derogatory, they inadvertently censor productive conversations held by marginalized groups. One way to engage with marginalized perspectives is to add context around conversations. Previous research has leveraged user- and thread-level data, but it often neglects where these conversations take place. This paper highlights how community context can improve classification outcomes in abusive language detection. We make three main contributions to this end. First, we demonstrate that online communities cluster by the nature of their support towards marginalized groups. Second, we establish that community context improves accuracy and reduces false positive rates of state-of-the-art abusive language classifiers. Third, we reveal that community context can make human annotation procedures more robust. Our findings suggest a promising direction for context-aware models in abusive language research.

6.1.2 Introduction

Existing models for abuse detection struggle to grasp subtle knowledge about the social environments that they operate within. They do not perform natural language understanding and consequently cannot generalize when tested out-of-distribution [133, 17]. This problem is often the result of imbalances in training data, which encourage language models to overestimate the significance of certain lexical cues. For instance, Wiegand, Ruppenhofer, and Kleinbauer [233] observe that "commentator", "football", and "announcer" end up strongly correlated with hateful tweets in the Waseem and Hovy [229] corpus. These correlations are caused by focused sampling and are not observed in abusive expressions at large.

When models rely on pejorative or demographic words, they can encode systemic bias through *false positives* [10, 120]. For example, research has established that detection algorithms are more likely to classify comments written in African-American Vernacular English (AAVE) as offensive [60, 238]. Benign tweets like "*Wussup, n*gga*!" and "*I saw his ass yesterday*" both score above 90% for toxicity [196]. Similarly, Zhang et al. [245] analyze the Wikipedia Talk page corpus [67] and find that 58% of comments that contain the term "gay" are labelled as toxic, even though toxic comments only constitute 10% of the corpus. This again enables the misclassification of positive phrases like "*she makes me happy to be gay*". Even Twitter accounts belonging to drag queens have been rated higher in terms of average toxicity than the accounts associated with white nationalists [171]. These findings underline how language models with faulty correlations can facilitate the censorship of productive conversations held by marginalized communities.

Productive conversations containing slurs are common, and they take many forms [101]. For example, research inspired by the #MeToo movement has focused on the detection of sexual harassment disclosures by victims [137, 48, 136, 62, 47]. However, research on minority narratives has not been sufficiently integrated into the literature on abusive language detection. The distinction between actual sexist messages and messages calling out sexism is rarely addressed in the field [45]. A similar trend is seen with sarcasm. Humor and self-irony can be employed as coping mechanisms by victims of abuse [86], yet they constitute frequent sources of error for state-of-the-art classifiers [222]. For example, the median toxicity score for language on `r/transgendercirclejerk`, a "parody [subreddit] for trans people", is as high as 90% [130]. More broadly, transgender users are "excluded, harmed, and misrepresented in existing platforms, algorithms, and research methods" relating to social network analysis [211].

Meaningful improvements in abuse detection require a thoughtful engagement with the perspectives of marginalized communities and their allies. This is a necessary step towards making language models socially aware, and it may even lead to novel methods for platform moderation [223]. For example, counter speech generation has been shown to effectively combat abuse [212, 18, 197, 236], but this solution has limited applications at-scale since we cannot reliably detect counter speech in the first place.

One way to ensure that machine learning frameworks are socially conscientious is to add context around conversations. Past research has explored the contextual information within conversation threads [178, 248], user demographics [220, 79], user history [220, 187, 58, 248], user profiles [220, 79], and user networks [220, 248, 154, 174] with varying degrees of success in improving performance. However, most modelling efforts for abuse detection neglect a major aspect of online conversations: the community environment within which they take place.

Online communities adhere to specific socio-linguistic conventions that reinforce group identity. This phenomenon is easily observed on Reddit, where explicit community structure is part of platform design. The majority of comments on the pro-Trump subreddit `r/The_Donald` delegitimize liberal ideas [149, 203]. Similarly, the collection of "manosphere" and "red pill" subreddits espouse and encourage misogynistic ideologies [211, 88, 144]. Thus, community identity dictates content produced within the digital space. We note that community structures are not limited to Reddit. They are also components of the platform designs for 4chan, Facebook, Gab, Voat, etc. Community structures are also present across other social media platforms (e.g., Twitter), though in a less explicit manner [201].

In this paper, we investigate the importance of community context for abusive language detection. Our contributions are as follows:

- We demonstrate that online communities cluster both by topic and by the nature of their support towards marginalized groups.
- We establish that community context improves accuracy and reduces false positive rates of state-of-the-art abusive language classifiers.
- We discover that community contextualized algorithms can detect false positive errors in human annotations, and context can generally be used to make human annotation procedures more robust.

6.1.3 Related Work

Methods in Abusive Language Detection

Abusive language detection is a relatively new field of research, with "very limited" work from as recently as 2016 [229]. Early methods featured Naive Bayes [2, 131, 135], SVMs [33, 61, 64, 93, 218, 227], Random Forests [33, 61, 64, 93, 218, 227], Decision Trees [33, 61, 64, 93, 218, 227], and Logistic Regression [33, 61, 64, 93, 218, 227].

However, recent developments in NLP have directed the field towards neural and Transformer-based approaches. CNNs [83, 145, 151, 155], LSTMs (+ Attention) [151, 155, 118, 215, 40], and GRUs [247, 155] have been widely used in the literature. As of 2019, researchers have begun adopting pre-trained language models. Contemporary work leverages BERT, DistilBERT, ALBERT, RoBERTA, and mBERT [5, 22, 59, 172, 216]. In fact, Bodapati et al. [22] note that seven of the top ten performing models for offensive language identification at SEMEVAL-2019 were BERT-based. A similar trend was seen at SEMEVAL-2020, where "most teams used some kind of pre-trained Transformers" [243]. Regardless of architecture, methods in abusive language detection can generally be divided into content- and context- based approaches.

Content-based approaches rely on comment text for feature engineering. Researchers have used TF-IDF weighted n-gram counts as well as distributional embeddings for text representation [61, 167, 221], POS tags or dependency relations for encoding syntactic information [61, 167, 164], and the frequencies of hashtags, URLs, user mentions, emojis, etc. for detecting platform-specific tokens. Lexicons are also popular for capturing sentiment [38], politeness [167], emotion [142, 84], hate [126, 41] and clout [246]. The central assumption behind content-based abusive language detection is that comments can be exclusively assessed using textual features. However, this assumption neither holds in theory nor in practice as linguistic structures are discourse-determined, and that discourse is shaped by social, historical, and political context [29, 198]. Semantics cannot be completely interpreted using content cues alone. Even human annotators struggle to classify comments that involve satire or homonymy in the absence of broader information [130]. In light of these concerns, researchers are increasingly identifying the importance of user or conversational features to their detection frameworks. Five current trends within context-based approaches are outlined below:

Conversational Context. Attempts have been made to situate abusive comments within conversation threads. Threads have been studied using preceding comments [178, 117], discussion titles [84, 178], and counts for aggressive comments specifically [248, 114]. The position of a comment in a thread, including whether it is at the start or end, has also been considered [114]. Finally, researchers have analyzed conversation graphs for topological indicators of abuse [174].

User Demographics. Researchers have attempted to incorporate user-level context through demographic signals for age, location, and gender. Age is extracted from user disclosures, but these disclosures can be unreliable if users have an incentive to view adult-rated content [58]. Previous work has inferred gender from user names [229, 220], expressions in user biographies [229, 220], and in-game avatar choices [11]. These methods are problematic because names are not necessarily a reliable indicator of gender. Location information obtained through geo-coding has also been used to analyze hateful tweets [75]. However, accurate demographic information is rarely available and therefore its usage is not popular in the literature.

User History. Patterns in user behaviour are indicators of user history. For instance, metrics that count daily logins [11], favourites [220], and posting history [248, 220] have been used to track past activity. Other works focus directly on previous comments. For example, Dadvar et al. [58] look for the prevalence of profanity in a user's past comments. Conversely, Qian et al. [187] encode all historical posts by a user. Similarly, Ziems, Vigfusson, and Morstatter [248] create TF-IDF vectors derived from a user's timeline.

User Profile. Various profile metadata have been studied as a proxy for digital identity. Usernames have been included in detection efforts [84], while user anonymity has been correlated with hateful messaging [244, 237]. Efforts by users to curate their identities have also been incorporated through the presence of updated profile pictures [220] or biographies [152]. User popularity measured through verified account status [248], counts for followers [79] or in-game friends [11] is also common. Some other profile features include profile language [82] and account age [79].

User Network. Homophily in social networks induces the clustering of users based on shared identities. These clusters have been shown to represent collective ideologies [51, 96] and moralities [63], motivating researchers to examine local user networks for markers of abusive behaviour. Interaction and connection-based social graphs have also been analyzed using traditional metrics, such as Jaccard's similarity [248] and eigenvalue or closeness centrality [41, 79, 220, 174] and are also the basis for obtaining user embeddings. [191, 154, 46].

Methods in Community Profiling

User networks only capture localized information around a single user, overlooking how a group of users behave as a whole. There are connection- and content-based solutions for explicit community profiling that exist outside of the current literature on abusive language detection. Connection-based solutions evolve out of the idea that similar communities house similar users. In contrast, content-based solutions claim that similar communities house similar content.

Connection Based Representations. ector representations of online communities have been shown to encode semantic relationships [143]. Popular techniques for obtaining these representations require the construction of a community graph. Kumar et al. [129] construct a bipartite multigraph between Reddit users and subreddits. An edge $u_i \rightarrow s_j$ is added for each post by a user u_i in a subreddit s_j. The resulting graph is then used to learn subreddit embeddings by a "node2vec-style" approach.

Martin [143] create a symmetric matrix $\mathbf{X}$ of subreddit-subreddit user co-occurrences, where $\mathbf{X}_{ij}$ is the number of unique users who have commented at least ten times in the subreddits s_i and s_j. Skip-grams with negative sampling [134] or GloVe [181] can then be used to obtain subreddit embeddings. Here, subreddits inherit the role of words, and user co-occurrences inherit the role of word co-occurrences. Waller and Anderson [225] also treat communities as "words" and users who comment in them as "contexts" and adapt word2vec for community representations.

Finally, the subreddit graph proposed in Janchevski and Gievska [105] contains edges $s_i - s_j$ weighted by the number of shared users between the subreddits s_i and

Label	**Count**		**Stats**	**Count**
DEG	20531	51%	Users	36962
NDNA	16729	49%	Posts	34610
HOM	1998		Subreddits	2691
APR	553			
Total	39811			

Table 6–1: Slur-Corpus details. Derogatory and non-derogatory comments have an almost equal share. The corpus contains comments from a vast variety of users, posts and subreddits.

s_j. A key difference is that the authors only consider users who participate in at least ten subreddits, after which they use node2vec to generate node embeddings.

Content Based Representations. Content-based solutions for community profiling rely on methods for document similarity. Janchevski and Gievska [105] average the word2vec representations for the top 30 words in each subreddit, ranked by TF-IDF score. To the best of our knowledge, this research is limited and the encoding of semantic relationships between communities has not yet been established.

6.1.4 Methodology

Data Overview

For our analysis we choose the Slur-Corpus by Kurrek, Saleem, and Ruths [130], which consists of 40k human-annotated Reddit comments. Every comment contains a slur and is labelled as either derogatory (DEG), appropriative (APR), non-derogatory non-appropriative (NDNA), or homonym (HOM). The corpus is nearly evenly split between derogatory and non-derogatory (APR, NDNA, HOM) slur usages, with 51% of comments labelled DEG (see Table 6–1). Crucially, this corpus explicitly provides and labels comments that contain a slur but are not abusive.

The Slur-Corpus is one of few community-aware resources for abusive language detection. The data is sampled over the course of a decade (October 2007 to September 2019), reflecting a variety of users and language conventions. Every comment is published with the subreddit from which it is sourced, and the authors curate content across a number of antagonistic, supportive, and general discussion subreddits. As opposed to random sampling, this method guarantees the representation of targeted and minority voices. We see this as crucial for investigating the role of social context within abusive language conventions.

subsubsectionDefinitions Subreddits are niche communities dedicated to the discussion of a particular topic, with users participating in subreddits that engage their personal interests. As a result, subreddits often exhibit language specificity, and that specificity is crucial for making inferences about slur usages.

Consider the slur *tr*nny*. The comment "I am genuinly surpised at a suicidal *tr*nny*" from r/CringeAnarchy is derogatory. In contrast, "So do I. Just that the *tr*nny* is dying on me lol." from r/Honda is non-derogatory because *tr*nny* is used as a homonym. Both of these subreddits adhere to different linguistic norms and appeal to different user bases. Quantifying such differences is important. Niche or small automotive subreddits are likely to be related to r/Honda, and their users may also use *tr*nny* to mean *transmission*.

Constructing Subreddit Embeddings

We construct subreddit embeddings based on user comment co-occurrence. This method is in line with prior work on the subject [143, 129, 225], but it extends existing studies by considering data collected at a much larger scale. We use all publicly available Reddit comments prior to September 2019 in order to generate lists of users that comment in each subreddit [15] and then store frequency counts for each list. In total, we identify 998K unique subreddits and 42.7M unique authors over the course of 12 years. We see a long tail because many subreddits receive little participation.

Next, we identify active users, defined as being any user with at least ten comments in any given subreddit. We exclude bot accounts and focus on top subreddits as classified by activity. This leaves 10.4K subreddits and 12.2M unique users. With this data, we build a subreddit adjacency matrix $\mathbf{A}$, where $\mathbf{A}_{ij}$ is the number of co-occurring users in subreddits i and j. We use `GloVe` [181] to generate dense subreddit embeddings from $\mathbf{A}$ and run it over 100 epochs with a learning rate of 0.05 and a representation size of 150.

Evaluating Subreddit Embeddings

There are two conditions that we would like to capture in our tests for subreddit similarity. The first condition is that subreddit representations exhibit compositionality, i.e. similar subreddits have similar constituent subreddits. The second condition is that subreddit representations permit analogical reasoning, i.e. subreddit similarity is preserved under analogical argument. We rely on vector algebra to model each of these two conditions.

Similarity. The similarity between two subreddits S_i and S_j is simply the cosine similarity of their representations:

$$sim(S_i, S_j) = \frac{\vec{S_i} \cdot \vec{S_j}}{|\vec{S_i}||\vec{S_j}|}$$

Composition Test

city	+	*sport*	=	*team*
montreal	+	hockey	=	habs
toronto	+	baseball	=	Torontobluejays
toronto	+	hockey	=	leafs
toronto	+	nba	=	torontoraptors
chicago	+	baseball	=	CHICubs
chicago	+	hockey	=	hawks
chicago	+	nba	=	chicagobulls
chicago	+	nfl	=	CHIBears
boston	+	baseball	=	redsox
boston	+	hockey	=	BostonBruins
boston	+	nba	=	bostonceltics
boston	+	nfl	=	Patriots

Analogy Test

city	:	*team*	::	*city*	:	*team*
boston	:	BostonBruins	::	toronto	:	leafs
boston	:	redsox	::	toronto	:	Torontobluejays
boston	:	bostonceltics	::	toronto	:	torontoraptors
boston	:	Patriots	::	chicago	:	CHIBears
team	:	*sport*	::	*team*	:	*sport*
redsox	:	baseball	::	BostonBruins	:	hockey
redsox	:	baseball	::	bostonceltics	:	nba
redsox	:	baseball	::	Patriots	:	nfl
university	:	*city*	::	*university*	:	*city*
mcgill	:	montreal	::	UBC	:	vancouver
mcgill	:	montreal	::	UofT	:	toronto
mcgill	:	montreal	::	uAlberta	:	Edmonton

Table 6–2: Examples of subreddit embedding evaluation, based on our composition and analogy tests.

Composition Tests. We find a subreddit S_k that represents the sum of a pair of subreddits S_i and S_j. We create $\vec{V} = \vec{S_i} + \vec{S_j}$ and then compute $S_k = max(\{sim(\vec{V}, \vec{S_x}) \; \forall x : 1, ..., n\})$. We run the composition test to identify local sports team subreddits from combinations of sport and city subreddits ($\overrightarrow{sport} + \overrightarrow{city} = \overrightarrow{team}$). We base these tests on the evaluations of Martin [143].

Analogy Tests. We find a subreddit S_n such that $\vec{S_i} : \vec{S_j} :: \vec{S_m} : \vec{S_n}$ for a triad of subreddits S_i, S_j and S_m. We create $\vec{V} = \vec{S_i} - \vec{S_j} + \vec{S_m}$ and then compute $S_n = max(\{sim(\vec{V}, \vec{S_x})\forall x : 1, ..., n\})$. We run the analogy test to identify:

1. A local team given a city and sport:
   $\overrightarrow{city} : \overrightarrow{team} :: \overrightarrow{city'} : \overrightarrow{team'}$
2. A sport given a team and its city:
   $\overrightarrow{team} : \overrightarrow{sport} :: \overrightarrow{team'} : \overrightarrow{sport'}$
3. A city given a university
   $\overrightarrow{university} : \overrightarrow{city} :: \overrightarrow{university'} : \overrightarrow{city'}$

We base these tests on the evaluations of Waller and Anderson [225].

In total, we ran 157 composition tests and 6349 analogy tests. In 47% of cases, the correct answer to a composition test was the top subreddit using cosine similarity. In 81% of cases, the correct answer was in the top five most similar subreddits. We observe 66% and 84% accuracy for the analogy tests, evaluated on the top subreddit and top five most similar subreddits. Examples are highlighted in Table 6–2.

Context Insensitive Classifiers

To assess the importance of social context, we run a series of experiments first without and then with access to community information. We describe our context free models here.

LOG-REG. Our first classifier is a Logistic Regression with L2 regularization. We preprocess the corpus by lowercasing and stemming the text, removing stop words, and masking user mentions and URLs prior to tokenization. Each token is then weighed using TF-IDF to create unigram, bigram, and trigram features. We use scikit-learn [180] to create the classification pipeline.

BERT. Our second classifier is BERT [65]. We use BERT-BASE pre-trained on uncased data with AdamW optimizer [138], which has a final linear layer. We use BERTForSequenceClassification from huggingface [235] for implementation. It takes BERT's top-level embedding of the [CLS] token as input. Fine-tuning was done over four epochs with a batch size of 32. We choose a learning rate of 2e-05 and epsilon 1e-8. All BERT experiments were performed on Google Colab with Tesla V100-SXM2-16GB GPU.

[CLS] *comment* [SEP]

PERSPECTIVE. Our third classifier is a commercial tool for toxicity detection that is publicly available through the PERSPECTIVE API [1] . It is a CNN-based model that is trained on a high volume of user-generated comments across social media platforms. While the tool is updated by PERSPECTIVE, the API cannot be retrained, fine-tuned, or modified. We use 0.8 as the threshold for our derogatory label.

All experiments were run using 5-fold cross validation in order to label the entire corpus. Moreover, we use stratified sampling to ensure a uniform distribution of slurs, subreddits, and labels across all folds.

[1] www.perspectiveapi.com

Context Sensitive Classifiers

LOG-REG-COMM. We use the same setup as in LOG-REG, however, we include an additional feature for the name of the subreddit that each comment is sourced from. This is done to incorporate a social prior with which the algorithm can contextualize the comment text.

BERT-COMM. We extend BERT for social context sensitivity. In this setup, we concatenate the name of the source subreddit at the beginning of every comment before passing it to BERT.

$$\texttt{[CLS]}\ subreddit + comment\ \texttt{[SEP]}$$

BERT-COMM-SEP. In our second variant for context sensitivity, we use the sentence entailment format for BERT. This model concatenates the comment with the source subreddit, separated by BERT's [SEP] token. The model is then fine tuned in the same way as our other BERT models.

$$\texttt{[CLS]}\ comment\ \texttt{[SEP]}\ subreddit\ \texttt{[SEP]}$$

BERT-COMM-NGH. We use the same sentence entailment format in our last variant of context sensitive BERT. However, we provide additional context by incorporating the direct neighbourhood of the source subreddit. To get this neighbourhood, we use our trained GloVe embeddings from Section 6.1.4. We extract the five most similar subreddits to each source subreddit. These subreddits are concatenated with the source subreddit, and the model is trained in the sentence entailment format.

$$\texttt{[CLS]}\ comment\ \texttt{[SEP]}\ subreddit\ sim - subreddit - [1:5]\ \texttt{[SEP]}$$

where $sim - subreddit - [1:5]$ are the top-5 most similar subreddits to the source subreddit.

6.1.5 Results

Subreddits Cluster Around Polarity Towards Social Groups

We want to examine the structure of subreddits based on their polarity towards specific communities and social groups. With reference to the classification guidelines proposed in Kurrek, Saleem, and Ruths [130], we identify fifteen supportive, antagonistic, and general discussion subreddits. We use our `GloVe` embeddings to extract the five most similar subreddits to each of them (see: Table 6–3). We make three main observations.

Supportive subreddits			
gaybros	Blackfellas	trans	TwoXChromosomes
askgaybros	blackladies	transpositive	TrollXChromosomes
gay	BlackHair	ask_transgender	relationships
gaymers	racism	MtF	AskWomen
lolgrindr	AsABlackMan	transadorable	UpliftingNews
gaybrosgonemild	BlackPeopleTwitter	transpassing	books
Antogonisitc subreddits			
4chan	CoonTown	GenderCritical	MGTOW
ImGoingToHellForThis	GreatApes	itsafetish	WhereAreAllTheGoodMen
classic4chan	WhiteRights	GCdebatesQT	TheRedPill
CringeAnarchy	AntiPOZi	Gender_Critical	asktrp
circlejerk	european	GenderCriticalGuys	MGTOW2
TumblrInAction	OffensiveSpeech	truelesbians	Braincels
Discussion subreddits			
changemyview	hiphop	cars	relationships
PoliticalDiscussion	90sHipHop	Autos	AskWomen
bestof	rap	BMW	relationship_advice
TrueAskReddit	hiphop101	carporn	offmychest
explainlikeimfive	trapmuzik	AutoDetailing	TwoXChromosomes
OutOfTheLoop	makinghiphop	Justrolledintotheshop	sex

Table 6–3: 5 most similar subreddits to the subreddit on top. We find that subreddits cluster not only on topic but also on their polarity towards social groups.

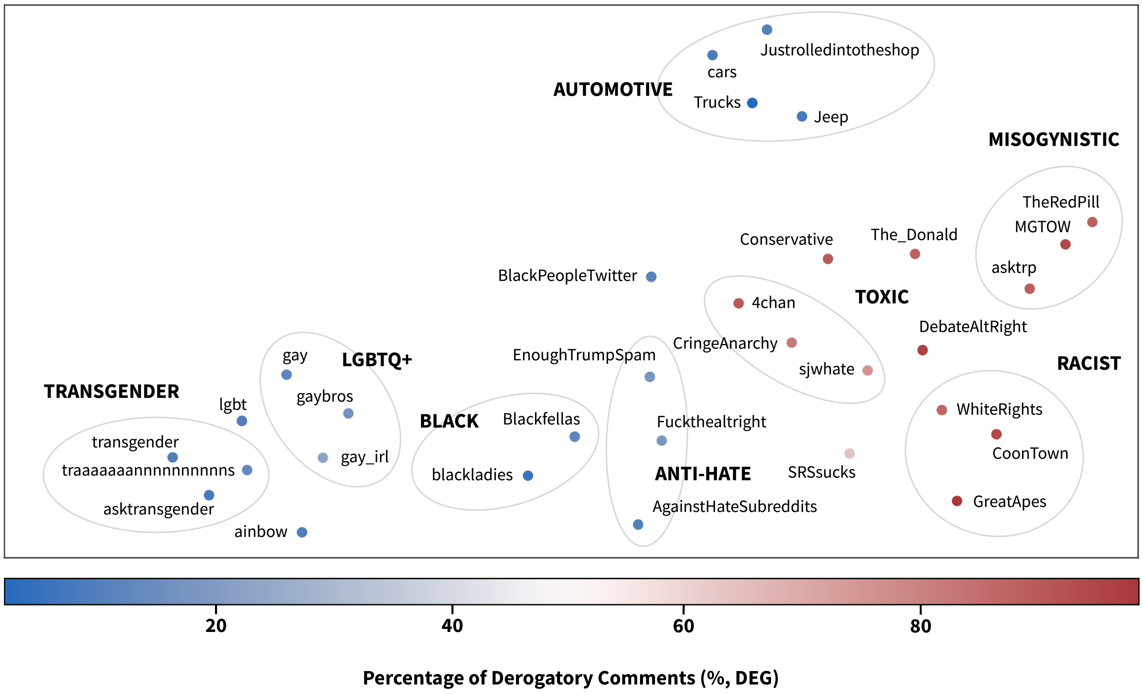


Figure 6–1: Clustering of prominent subreddits present in the `Slur-Corpus`. We find subredit clusters based on topic as well polarities towards social groups.

First, we notice that subreddits tend to cluster around topics. These topics can be interest-based, as seen in the case of `r/cars` and `r/hiphop`, or they can be utility-based, e.g. subreddits like `r/relationships` that provide advice. This observation is in line with previous work that establishes how subreddits cluster around discussion points like music, movies, and sports [129, 143].

Second, we observe that supportive subreddits are most similar to other supportive subreddits that cater towards the same minority community. For instance, the neighbourhood of `r/gaybros`, a subreddit built for the LGBTQ+ community, contains other subreddits based on pride and support: `r/askgaybros`, `r/gay`, `r/gaymers`,

	Perfromance				**% label classified as DEG**			
	Acc	**Pre**	**Rec**	**F1**	DEG	NDNA	APR	HOM
PERSPECTIVE	0.6132	0.6147	0.6102	0.6079	70.75%	53.10%	53.16%	10.71%
LOG-REG	0.8003	0.8009	0.7994	0.7997	82.85%	22.46%	61.30%	16.67%
LOG-REG-COMM	0.8002	0.8001	0.7999	0.8000	81.10%	20.53%	58.95%	15.67%
BERT	0.8856	0.8854	0.8857	0.8855	88.06%	10.26%	47.20%	6.31%
BERT-COMM	0.8905	0.8904	0.8908	0.8905	88.08%	9.38%	42.31%	5.36%
BERT-COMM-SEP	0.8930	0.8930	0.8934	0.8930	88.12%	8.95%	39.60%	5.11%
BERT-COMM-NGH	0.8923	0.8924	0.8928	0.8923	87.82%	8.80%	39.78%	4.75%

Table 6–4: The results of our models on the Slur-Corpus. In the left half of the table we report popular performance metrics. On the right half of the table, we report what percentage of each label was classified as derogatory. It indicates percentage true positives for DEG and percentage false positives for the other three labels.

r/lolgrindr, and r/gaybrosgonemild. A similar trend is observed with the neighbours of r/Blackfellas and r/trans.

Third, we see that antagonistic subreddits are most similar to other antagonistic subreddits. r/GenderCritical is contained in a cluster of anti-trans subreddits, r/MGTOW is near misogynistic subreddits, and r/CoonTown is surrounded by racist subreddits. Therefore polarizing communities cluster with other communities of similar polarity.

Figure 6–1 shows the embeddings of a sample of subreddits from Slur-Corpus plotted in two-dimensions using UMAP [147]. There are independent groups for misogynistic, racist, toxic, anti-hate, black, gay, trans, and automotive subreddits.

Subreddit Context Reduces False Positives

We present the results from our classification experiments in Table 6–4. We discuss the results from two lenses: (1) overall model performance and (2) model performance by label.

Overall, BERT-based models outperform classifiers based on Logistic Regression. This is unsurprising, given that Transformers are the current state of the art in NLP. However, LOG-REG achieves nearly 20% higher accuracy than PERSPECTIVE. While this performance gap is likely the result of the data used to train both models, it is concerning given that the Perspective API is widely used as a tool for toxicity detection with both commercial [2] and academic applications [57].

For both BERT and LOG-REG, the addition of subreddit context reduced the amount of false positives across all three non-derogatory labels. Performance on derogatory comments remained relatively unchanged. The highest increase in performance was seen with BERT-COMM-SEP, that uses the sentence entailment format, with each source subreddit concatenated with each comment and a [SEP] token in the middle. Adding subreddit context led to a significant improvement for appropriative text, within which the false positive rate went down by almost 8%. For example, "*Tr*nny* here, some of us are actually really cool." was misclassfied without context. Interestingly, BERT-COMM-NGH, our model containing subreddit neighbourhood context, shows little improvement over the baseline. While the identification of NDNA and HOM improves marginally, the false positive rate for appropriative language increases.

Context for Detecting Human Annotation Errors

We call a comment "context sensitive" if the addition of context changes its classification label. We examine the difference between BERT and BERT-COMM-SEP,

[2] Trusted partners include Reddit, The New York Times, The Financial Times, and the Wall Street Journal.

its community contextualized counterpart. The two models perform identically on the majority of the corpus: 94% of comments are context insensitive (see Table 6–5).

1364 of the total classification errors made by `BERT` were rectified when social context was included. These classifications represents > 3% of the actual corpus, but 56% of the context-sensitive comments. In Table 6–6, we present the top subreddits for both the *true positive* and *true negative* context sensitive comments, along with examples for each. The *true positive* comments largely belong to antagonistic subreddits, while the *true negative* ones belong to supportive subreddits. Generally speaking, community context helped `BERT-COMM-SEP` identify the antagonistic or supportive nature of a community.

On the other hand, context resulted in the misclassification of 1067 comments that were initally correctly labelled by basic `BERT`. *False positives* were concentrated across general discussion subreddits, like `r/pics`, `r/funny`, and `r/AskReddit`. Surprisingly, the top contributor to *false negative* classifications turned out to be supportive subreddits like `r/transgendercirclejerk`, `r/askgaybros`, and `r/rupaulsdragrace`. These comments were identified as derogatory by both human annotators and `BERT` but classified as non-derogatory when context was added to the model.

		`BERT`	∩	`BERT-COMM-SEP`	
False Positives		765	1339	480	
True Positives	1067{	587	17492	599	}1364
True Negatives		480	16696	765	
False Negatives		599	1853	587	
	2.68%	6.11%	93.89%	6.11%	3.43%

Table 6–5: Comparing performance difference from social priming for `BERT`. Column ∩ are comments that were classified in the same manner by both context sensitive and insensitive model and the other two are comments with dissimilar classification.

True Positives	
`CringeAnarchy`	I am genuinly surpised at a suicidal _tr*nny_
`4chan`	This is basically everyday in Atlanta. It's _n*gger/sp*c_ central. Give a useful warning next time.
`The_Donald`	Yeah they spit in your face then say don't hit me I'm a _tr*nny_ thats hate crime! Degenerate scumbags.
`ImGoingToHellForThis`	I'm a trans_n*gger_ as well. It also gives me, a physically white male the privilege of saying _n*gger_
True Negatives	
`transgendercirclejerk`	uj/ this is a circle jerk sub its a hyperbolic representation of how cis think trans people act (I assume) rj/ uh god there ((((they)))) go again being transphobic against a poor _tr*nny_ who isnt suicidal
`rupaulsdragrace`	Soooooo.... what about _Tr*nny_ Chaser and Ladyboy? Are those gonna be removed too?
`BlackPeopleTwitter`	Shit Britney rides for us too, idk if you seen when she was about to let the hands fly on some dude for calling her security a _n*gger_
`askgaybros`	Masc bear here. Twinks are my favorite and _f*ggot_ is a pretty funny word :b

Table 6–6: Top subreddits in the corrections made by `BERT-COMM-SEP` along with examples. We find socially priming improves classification for antagonistic and supportive subreddits.

To investigate this further, we had an expert re-annotate 50 random comments from each of the *true positive*, *false positive*, *true negative*, and *false negative* classes. The annotator had full access to context, which is a significant difference from the annotation method described in the original corpus. We observe that for 60% of *false negatives*, the expert labels do not match the gold labels reported in the `Slur-Corpus` (see Table 6–7). In these cases, the *false negatives* appear not to be derogatory and

	True	**False**
Positive	6/50	6/50
Negative	0/50	30/50

Table 6–7: Number of comments for which the re-annotated label did not match the original label. *False negative* comments were mislabelled in the original corpus at a much higher rate.

thus were likely mislabelled during human annotation. These findings underline the importance of community context in human and systematic classification procedures.

6.1.6 Discussion

Our analysis points to three key insights that would benefit future abusive language research.

Subreddit embeddings promote community sampling.. Systems for abuse detection should reliably identify variations of abuse (e.g. sexism, racism, etc.), while still exhibiting sensitivity towards non-derogatory comments (e.g. appropriation, reclamation, etc.). One way to achieve this is to ensure content diversity in training data, and Kurrek, Saleem, and Ruths [130] specifically use community sampling to achieve diversity. The authors collect comments from various Reddit communities, but their work is limited by the absence of resources that identify and consolidate supportive or antagonistic subreddits. Instead, they rely on manual data exploration. There are several issues with this approach. First, knowing which communities to look for (and how to find them) requires a high degree of domain knowledge. Second, manual comment analysis is an expensive task, which makes it difficult to scale or reuse as new communities form. Third, this method is prone to overlooking smaller, niche subreddits that would otherwise have been found using a neighborhood exploration of community clusters. We propose the use of subreddit embeddings in future research to further extend efforts on diverse and representative content collection.

Community context protects productive conversations.. One of our primary research objectives was to ensure that detection frameworks do not mistakenly classify productive conversations as abusive. Community contextualized models, based on Logistic Regression as well as `BERT`, perform better at identifying non-derogatory comments than their context-free counterparts. Context is particularly helpful for identifying appropriative language, resulting in an 8% increase in accuracy with the addition of the subreddit name. Appropriation is a tool used by marginalized communities to counteract oppression. When abuse detection frameworks misclassify it, they censor the empowerment tools of the very communities that they are installed to protect. Our analysis of the `Slur-Corpus` suggests that productive conversations, of which empowerment is a subset, tend to happen in safe and supportive social spaces. It is therefore crucial that these spaces be considered for a more nuanced classification of abuse.

Human annotation procedures are better informed when they include access to context.. We observe a significant discrepancy between reported labels in the `Slur-Corpus` and expert labels obtained through re-annotation with context (see Section 4.3). This discrepancy is largely associated with *false negatives.* While some disagreement may be attributed to the subjectivity of abusive language (particularly when that language is assessed by different annotators), the degree to which that disagreement was observed is more likely to be caused by the absence of contextual information during coding [130]. It is true that not all comments require context. For example: “I have no interest in arguing with someone that resorts to name calling. Especially those who call others ‘*f*ggots*’. I’m more appalled that bigots like you support the team.” is clearly not abusive. Likewise, “U jew *n*gger* supporter” is clearly abusive. However, there do exist comments which are closer to the decision boundary, and additional information is crucial for their classification. Consider “Breaking News: Some folks still hate *n*ggers*! (er... thugs) And now over to our urban youth pundit and renowned Hip Hop expert: Geraldo!”. The intent of the author is unclear, but knowing that the text was posted in `r/blackfellas`, a community that supports “All Black Lives”, helps us recognize that it is unlikely to be abusive. The same is true for comments “That’s the one problem with these 4wd Monaros, the *tr*nny* is a dog.” and “My friend and I are going to start a channel called *tr*nny* trash and its gonna be hella amazing”, sourced from `r/cars` and `r/traaaaaaannnnnnnnnns`, respectively. Integrating community context is therefore crucial for improving algorithmic classifications as well as human annotations.

6.1.7 Conclusion and Future Work

The subjective nature of abuse makes it challenging to annotate and detect reliably. One method for making the problem of subjectivity tractable is to position online conversations within the larger context that they occur. This paper is an exploration of one type of contextual information: community identity. We find that the context derived from community identity can help in the collection, annotation, and classification of abusive language. We therefore believe that social context is integral to all the stages of abusive language research. For future work, we envision the integration of community information into a single, ensemble detection framework.

[2] Swati Agarwal and Ashish Sureka. “But i did not mean it!—intent classification of racist posts on tumblr”. In: *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE. 2016, pp. 124–127.

[5] Pedro Alonso, Rajkumar Saini, and György Kovács. “Hate speech detection using transformer ensembles on the hasoc dataset”. In: *International Conference on Speech and Computer*. Springer. 2020, pp. 13–21.

[10] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. “Stereotypical bias removal for hate speech detection task using knowledge-based generalizations”. In: *The World Wide Web Conference*. 2019, pp. 49–59.

[11] Koray Balci and Albert Ali Salah. “Automatic analysis and identification of verbal aggression and abusive behaviors for online social games”. In: *Computers in Human Behavior* 53 (2015), pp. 517–526.

[15] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[17] Emily M Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.

[18] Susan Benesch et al. “Counterspeech on Twitter: A field study”. In: *A report for Public Safety Canada under the Kanishka Project* (2016).

[22] Sravan Bodapati et al. “Neural Word Decomposition Models for Abusive Language Detection”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 135–145.

[29] Judith Bridges. “Gendering metapragmatics in online discourse:“Mansplaining man gonna mansplain. . . ”” In: *Discourse, Context & Media* 20 (2017), pp. 94–102.

[33] Peter Burnap and Matthew Leighton Williams. “Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making”. In: *Internet, Policy & Politics* (2014).

[38] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. “DeepHate: Hate speech detection via multi-faceted text representations”. In: *12th ACM Conference on Web Science*. 2020, pp. 11–20.

[40] Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. “Pay “attention” to your context when classifying abusive language”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 70–79.

[41] Despoina Chatzakou et al. “Mean birds: Detecting aggression and bullying on twitter”. In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 13–22.

[45] Patricia Chiril et al. “He said “who’s gonna take care of your children when you are at ACL?”: Reported Sexist Acts are Not Sexist”. In: *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics. 2020, pp. 4055–4066.

[46] Shivang Chopra et al. "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 386–393.

[47] Arijit Ghosh Chowdhury et al. "# YouToo? detection of personal recollections of sexual harassment on social media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2527–2537.

[48] Arijit Ghosh Chowdhury et al. "Speak up, fight back! detection of social media disclosures of sexual harassment". In: *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*. 2019, pp. 136–146.

[51] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". In: *Journal of communication* 64.2 (2014), pp. 317–332.

[57] Lana Cuthbertson et al. In: *Proceedings AI for Social Good workshop at NeurIPS*. 2019.

[58] Maral Dadvar et al. "Improving cyberbullying detection with user context". In: *Proceedings of the 35th European conference on Advances in Information Retrieval*. 2013, pp. 693–696.

[59] Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. "Developing a New Classifier for Automated Identification of Incivility in Social Media". In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 95–101.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. "Racial Bias in Hate Speech and Abusive Language Detection Datasets". In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019, pp. 25–35.

[61] Thomas Davidson et al. "Automated hate speech detection and the problem of offensive language". In: *Eleventh international aaai conference on web and social media*. 2017.

[62] Bonnie-Elene Deal et al. ""I Definitely Did Not Report It When I Was Raped...# WeBelieveChristine# MeToo": A Content Analysis of Disclosures of Sexual Assault on Twitter". In: *Social Media+ Society* 6 (2020).

[63] Morteza Dehghani et al. "Purity homophily in social networks." In: *Journal of Experimental Psychology: General* 145 (2016).

[64] Fabio Del Vigna et al. "Hate me, hate me not: Hate speech detection on facebook". In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC)*. 2017, pp. 86–95.

[65] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. 2019, pp. 4171–4186.

[67] Lucas Dixon et al. “Measuring and mitigating unintended bias in text classification”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 67–73.

[75] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. “Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter”. In: *Proceedings of the Association for Information Science and Technology* 57.1 (2020), e313.

[79] Antigoni Maria Founta et al. “A unified deep learning architecture for abuse detection”. In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 105–114.

[82] Patxi Galán-Garcıéa et al. “Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying”. In: *Logic Journal of the IGPL* 24.1 (2016), pp. 42–53.

[83] Björn Gambäck and Utpal Kumar Sikdar. “Using convolutional neural networks to classify hate-speech”. In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 85–90.

[84] Lei Gao and Ruihong Huang. “Detecting Online Hate Speech Using Context Aware Models”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., 2017, pp. 260–266.

[86] Jacqueline Garrick. “The humor of trauma survivors: Its application in a therapeutic milieu”. In: *Journal of aggression, maltreatment & trauma* 12.1-2 (2006), pp. 169–182.

[88] Debbie Ging. “Alphas, betas, and incels: Theorizing the masculinities of the manosphere”. In: *Men and Masculinities* 22.4 (2019), pp. 638–657.

[93] Edel Greevy. “Automatic text categorisation of racist webpages”. PhD thesis. Dublin City University, 2004.

[96] Yosh Halberstam and Brian Knight. “Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter”. In: *Journal of public economics* 143 (2016), pp. 73–88.

[101] Christopher Hom. “The semantics of racial epithets”. In: *The Journal of Philosophy* 105.8 (2008), pp. 416–440.

[105] Andrej Janchevski and Sonja Gievska. “A Study of Different Models for Subreddit Recommendation Based on User-Community Interaction”. In: *International Conference on ICT Innovations*. Springer. 2019, pp. 96–108.

[114] Srecko Joksimovic et al. “Automated Identification of Verbally Abusive Behaviors in Online Discussions”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 36–45.

[117] Mladen Karan and Jan Šnajder. “Preemptive toxic language detection in Wikipedia comments using thread-level context”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 129–134.

[118] Christos Karatsalos and Yannis Panagiotakis. “Attention-based method for categorizing different types of online harassment language”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 321–330.

[120] Brendan Kennedy et al. “Contextualizing Hate Speech Classifiers with Post-hoc Explanation”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2020.

[126] Anna Koufakou et al. “HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 2020, pp. 34–43.

[129] Srijan Kumar et al. “Community interaction and conflict on the web”. In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 933–943.

[130] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. “Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 138–149.

[131] Irene Kwok and Yuzhou Wang. “Locate the Hate: Detecting Tweets against Blacks.” In: *AAAI*. 2013.

[133] Ronan Le Bras et al. “Adversarial filters of dataset biases”. In: *International Conference on Machine Learning*. 2020, pp. 1078–1088.

[134] Omer Levy, Yoav Goldberg, and Ido Dagan. “Improving distributional similarity with lessons learned from word embeddings”. In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 211–225.

[135] Shuhua Liu and Thomas Forss. “Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification”. In: *Proceedings*

of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 1. 2014, pp. 530–537.

[136] Yingchi Liu et al. "Sexual harassment story classification and key information identification". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 2385–2388.

[137] Yingchi Liu et al. "Uncover Sexual Harassment Patterns from Personal Stories by Joint Key Element Extraction and Categorization". 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)*. 2019, pp. 2328–2337.

[138] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. 2018.

[142] Ilia Markov and Walter Daelemans. "Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate". In: *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2021.

[143] Trevor Martin. "community2vec: Vector representations of online communities encode semantic relationships". In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. 2017, pp. 27–31.

[144] Adrienne Massanari. "# Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures". In: *New media & society* 19.3 (2017), pp. 329–346.

[145] Puneet Mathur et al. "Detecting offensive tweets in hindi-english code-switched language". In: *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. 2018, pp. 18–26.

[147] Leland McInnes et al. "UMAP: Uniform Manifold Approximation and Projection". In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[149] Quinnehtukqut McLamore and Özden Melis Uluğ. "Social representations of sociopolitical groups on r/The_Donald and emergent conflict narratives: A qualitative content analysis". In: *Analyses of Social Issues and Public Policy* (2020).

[151] Johannes Skjeggestad Meyer and Björn Gambäck. "A platform agnostic dual-strand hate speech detector". In: *ACL 2019 The Third Workshop on Abusive Language Online Proceedings of the Workshop*. Association for Computational Linguistics. 2019.

[152] Fernando Miró-Llinares, Asier Moneva, and Miriam Esteve. "Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments". In: *Crime Science* 7.1 (2018), pp. 1–12.

[154] Pushkar Mishra et al. "Author profiling for abuse detection". In: *Proceedings of the 27th international conference on computational linguistics*. 2018, pp. 1088–1098.

[155] Sandip Modha, Prasenjit Majumder, and Thomas Mandl. "Filtering aggression from the multilingual social media feed". In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 2018, pp. 199–207.

[164] Kanika Narang and Chris Brew. “Abusive Language Detection using Syntactic Dependency Graphs”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 44–53.

[167] Chikashi Nobata et al. “Abusive language detection in online user content”. In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 145–153.

[171] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. “Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online”. In: *Sexuality & Culture* 25.2 (2021), pp. 700–732.

[172] Kadir Bulut Ozler et al. “Fine-tuning BERT for multi-domain and multi-label incivil language detection”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 28–33.

[174] Etienne Papegnies et al. “Graph-based features for automatic online abuse detection”. In: *International conference on statistical language and speech processing*. Springer. 2017, pp. 70–81.

[178] John Pavlopoulos et al. “Toxicity Detection: Does Context Really Matter?” In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[180] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[181] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[187] Jing Qian et al. “Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 2. 2018, pp. 118–123.

[191] Michael Ridenhour et al. “Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models”. In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer. 2020, pp. 202–212.

[196] Maarten Sap et al. “The risk of racial bias in hate speech detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1668–1678.

[197] Carla Schieb and Mike Preuss. “Governing hate speech by means of counterspeech on Facebook”. In: *66th International Communication Association Annual Conference*. 2016, pp. 1–23.

[198] Melani Schröter and Petra Storjohann. “Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009”. In: *Pragmatics and Society* 6.1 (2015), pp. 43–66.

[201] Wendel Silva et al. “A methodology for community detection in Twitter”. In: *Proceedings of the International Conference on Web Intelligence*. 2017, pp. 1006–1009.

[203] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. “A characterization of political communities on reddit”. In: *Proceedings of the 30th ACM conference on hypertext and Social Media*. 2019, pp. 259–263.

[211] Leo G. Stewart and Emma S. Spiro. “Nobody Puts Redditor in a Binary: Digital Demography, Collective Identities, and Gender in a Subreddit Network”. In: *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2021.

[212] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. “Generating Counter Narratives against Online Hate Speech: Data and Strategies”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 1177–1190.

[215] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. “Textual aggression detection through deep learning”. In: *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*. 2018, pp. 177–187.

[216] Thanh Tran et al. “HABERTOR: An Efficient and Effective Deep Hatespeech Detector”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 7486–7502.

[218] Stéphan Tulkens et al. “A Dictionary-based Approach to Racism Detection in Dutch Social Media”. In: *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety.* LREC, 2016, p. 11.

[220] Elise Fehn Unsvåg and Björn Gambäck. “The Effects of User Features on Twitter Hate Speech Detection”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2).* 2018, pp. 75–85.

[221] Cynthia Van Hee et al. “Automatic detection of cyberbullying in social media text”. In: *PloS one* 13 (2018).

[222] Bertie Vidgen et al. “Challenges and frontiers in abusive content detection”. In: *Proceedings of the Third Workshop on Abusive Language Online.* 2019, pp. 80–93.

[223] Bertie Vidgen et al. “Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.* Association for Computational Linguistics, 2021.

[225] Isaac Waller and Ashton Anderson. “Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms”. In: *The World Wide Web Conference.* 2019, pp. 1954–1964.

[227] William Warner and Julia Hirschberg. “Detecting hate speech on the world wide web”. In: *Proceedings of the Second Workshop on Language in Social Media.* Association for Computational Linguistics. 2012, pp. 19–26.

[229] Zeerak Waseem and Dirk Hovy. “Hateful symbols or hateful people? predictive features for hate speech detection on twitter”. In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[233] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. “Detection of abusive language: the problem of biased datasets”. 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)*. 2019, pp. 602–608.

[235] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[236] Lucas Wright et al. “Vectors for Counterspeech on Twitter”. In: *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, 2017, pp. 57–62.

[237] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex machina: Personal attacks seen at scale”. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 1391–1399.

[238] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. “Demoting Racial Bias in Hate Speech Detection”. In: *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 2020, pp. 7–14.

[243] Marcos Zampieri et al. "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)". In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, 2020, pp. 1425–1447.

[244] Savvas Zannettou et al. "Measuring and characterizing hate speech on news websites". In: *12th ACM Conference on Web Science*. 2020, pp. 125–134.

[245] Guanhua Zhang et al. "Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4134–4145.

[246] Xiang Zhang et al. "Cyberbullying detection with a pronunciation based convolutional neural network". In: *2016 15th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2016, pp. 740–745.

[247] Ziqi Zhang, David Robinson, and Jonathan Tepper. "Detecting hate speech on twitter using a convolution-gru based deep neural network". In: *European semantic web conference*. Springer. 2018, pp. 745–760.

[248] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. "Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 808–819.

[END OF MANUSCRIPT]

For abusive language detection, I primarily focus on the issue of false-positives. False positive errors can have significant consequences, especially when they facilitate the inadvertent censoring of conversations from marginalized populations. In such cases, abuse detection frameworks can bring further harm the victims of abuse. It is imperative that the research community works towards reducing false positive errors, especially for reclamatory language. I provide evidence that contextualizing conversations within the social spaces they occur in, can help detection frameworks improve their performance by reducing misclassification of non-derogatory content. Furthermore, community context can also assist human annotation process with added clarification. Overall community context is an encouraging avenue for research to ensure robust detection of abusive language.

CHAPTER 7
Banning Antagonistic Communities for Abuse Intervention

This chapter is a case-study of an essential tool for content moderation - banning antagonistic communities.

Antagonistic communities not only condone abusive content, they explicitly promote it. There are multiple examples of such communities across the internet that range from stand alone websites (for example Stormfront [27] and ShitSkin [113]) to sub communities within large social media platforms such as Reddit[1] and Facebook[2] .

At the outset, it might seem that the most obvious course of action in combating abusive content is to disband these spaces. However such an intervention does not guarantee a favourable outcome and can lead to undesirable consequences such a platform migration or community brigading (when members of a community purposely flood another community to harass or cause turmoil). Therefore, it is important to analyse user behaviour to fully understand the aftermath of such a large-scale platform intervention.

[1] `www.adl.org/news/press-releases/adl-praises-reddit-for-removal-of-nazi-and-white-supremacist-content`

[2] `www.adl.org/blog/hateful-and-conspiratorial-groups-on-facebook`

In the summer of 2015, Reddit admins introduced a new policy to ban subreddits that systematically harass other users, which led to the shutdown of multiple communities, most prominent of which was `r/fatpeoplehate`. At the time of the ban, platform maintainers had no way of knowing how the affected users would react. This provided an opportunity to observe user behaviour amidst the fallout from the community banning. I study user engagement with the platform before and after the ban, alongside their efforts to infiltrate other communities and circumvent the intervention.

7.1 Manuscript 5: The Aftermath of Disbanding an Online Hateful Community

Authors: Haji Mohammad Saleem and Derek Ruths

Available at arXiv:1804.07354, 2018.

7.1.1 Abstract

Harassing and hateful speech in online spaces has become a common problem for platform maintainers and their users. The toxicity created by such content can discourage user participation and engagement. Therefore, it is crucial for and a common goal of platform managers to diminish hateful and harmful content. Over the last year, Reddit, a major online platform, enacted a policy of banning sub-communities (subreddits) that they deem harassing, with the goal of diminishing such activities. We studied the effects of banning the largest hateful subreddit (`r/fatpeoplehate` or FPH) on the users and other subreddits that were associated with it. We found that, while a number of outcomes were possible — in this case the subreddit ban led to a sustained reduced interaction of its members (FPH users) with the Reddit platform. We also found that the many counter-measures taken by FPH users were short-lived and promptly neutralized by both Reddit administrators and the admins of individual subreddits. Our findings show that forum-banning can be an effective means by which to diminish objectionable content. Moreover, our detailed analysis of the post-banning behavior of FPH users highlights a number of the behavioral patterns that banning can create.

7.1.2 Introduction

Hate speech has become widespread across the Internet, especially affecting online forums and social media platforms due to their user-generated content. Online hate speech can take on a number of forms: for example, a microblog that verbally abuses a black actor based on her race [3] or a homophobic video on a popular video sharing website [4] . However, it is not just the social media platforms, but news websites and in-game chat rooms of major MMOGs (Massively Multiplayer Online Games) are also affected by hate speech.

From the perspective of platform managers, hateful speech creates negative user experiences that can hurt user engagement and drive desirable users away, ultimately hurting the growth of the platform. At the user level, the effects of online hateful speech can extend far beyond the virtual world, causing serious mental and enduring psychological distress and even physical harm by inciting violence against the targets of the hate. As a result, diminishing hateful speech in online social spaces has become a major objective of both platform maintainers and many of the users who use them.

In an effort to combat such objectionable content, platforms employ a number of different strategies. Despite the widespread use of techniques ranging from community-based flagging to platform-level banning, little is known about the implications of these policies: in other words, even though platforms routinely deploy countermeasures against hate speech and their producers, very little is known about

[3] `www.huffingtonpost.ca/2016/07/19/leslie-jones-twitter_n_11069228.html`

[4] `www.independent.co.uk/voices/comment/youtubes-hall-of-homophobia-shame-8201996.html`

the direct and indirect effects of such actions - on the hateful content and on the communities affected by it.

In this study, we aim to understand one common practice among online forums: community banning. We specifically consider this in the context of Reddit and its 2016 policy of banning "subreddits that allow their communities to use the subreddit as a platform to harass individuals when moderators don't take action" [159].

Reddit is a social media content aggregation and discussion website, which works on the model of multiple niche communities, called subreddits. Subreddits cover a wide variety of topics, and users can subscribe to the subreddits that interest them or create their own to personalize their Reddit experience. Up until the summer of 2015, Reddit users were free to create subreddits dedicated to hateful themes. `r/fatpeopelhate` (FPH) was one such subreddit that was created to denigrate plus-sized people. The FPH subreddit was highly popular, with approximately 150,000 subscribers at its peak (Figure 7–1). Last year, Reddit introduced a new policy to ban harassing subreddits and `r/fatpeopelhate` was the most prominent subreddit banned under this policy. This study investigates the after-effects of banning this large self-identifying hateful subreddit on the users that participated in it (FPH users) and other related subreddits.

A priori banning a subreddit hardly guarantees that the kind of hateful content produced by the community will go away. A number of outcomes are possible. One outcome involves the community simply creating a new subreddit under a different name. Another outcome can be that hateful speakers post their content in diverse subreddits, making the hateful content much harder to expunge and possibly directly exposing members of target communities to the hateful content.

Our research produced strong evidence that all of these negative outcomes happened - but were very short-lived. Ultimately, we found that the banning of the FPH subreddit reduced the engagement of active FPH contributors with the Reddit platform and dramatically decreased the volume of such content being posted to Reddit. Not only did the FPH users comment less after the ban, a larger portion of the users stopped their commenting activity entirely. The initial response of the banned users (creating new subreddits and posting their hateful content to other platforms) was quickly neutralised (within 1-2 days) by either Reddit administrators or the moderators of specific subreddits.

Overall, our findings support the conclusion that banning of objectionable subreddits does diminish that kind of content on the platform. Notably, achieving this end involves swift, independent, and decisive action on the part of both administrators and moderators.

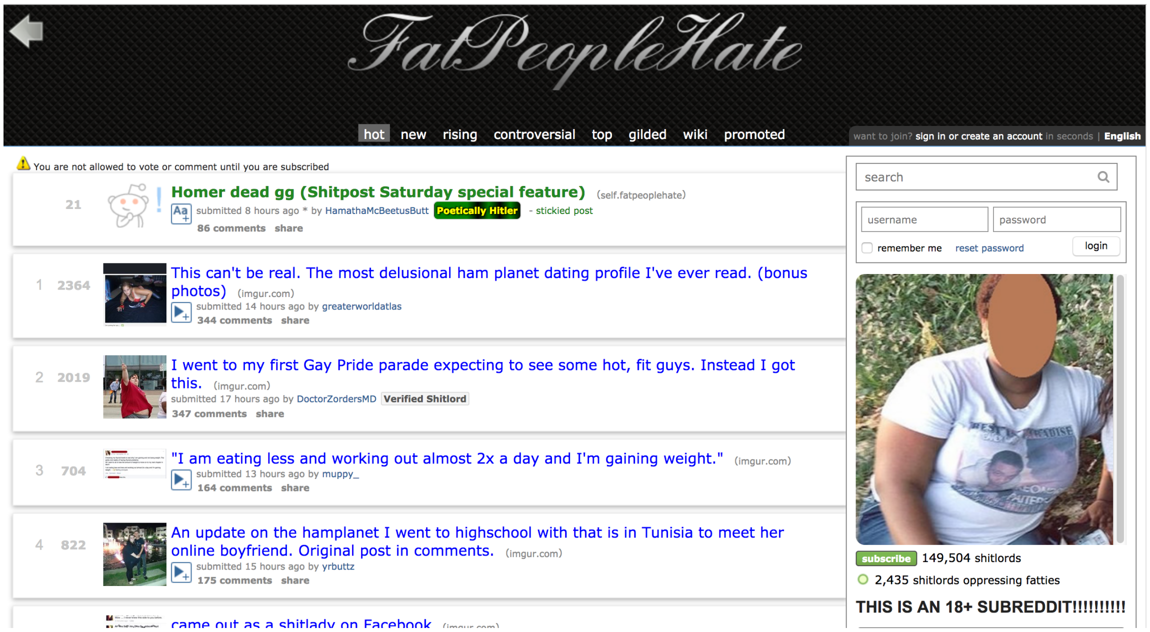
7.1.3 Background

Hate speech

At the outset, establishing precisely what constitutes hate speech is an unsolved and important problem. In order to address often-contested definitions of hate speech and remove ambiguity we use the term "hateful speech"[194] which is defined as "speech which contains an expression of hatred on the part of the speaker/author, against a person or people, based on their group identity". This definition removes the ambiguity in the term "hate" itself, which might refer to the speaker/author's hatred, or his/her desire to make the targets of the speech feel hated, or their desire to make others hate the target(s), or the apparent capacity of their speech to increase hatred. Therefore, in this article we use the term "hate speech" in a general sense and the term "hateful speech" specifically as defined above.

Banning of `r/FatPeopleHate`

`r/FatPeopleHate` (FPH) was a subreddit that was created to mock and denigrate plus-sized people. The users would often post pictures of plus-sized individuals and ridicule them along with the usage of slurs, such as "*hamplanet*", "*landwhale*", "*beetus*". One of the rules for contributing to the subreddit included having "*absolutely no fat sympathy*". While the community admitted that it may seem that "*all*



You are not allowed to vote or comment until you are subscribed

21 **Homer dead gg (Shitpost Saturday special feature)** (self.fatpeoplehate)
submitted 8 hours ago * by HamathaMcBeetusButt **Poetically Hitler** - stickied post
86 comments **share**

1 2364 This can't be real. The most delusional ham planet dating profile I've ever read. (bonus photos) (imgur.com)
submitted 14 hours ago by greaterworldatlas
344 comments **share**

2 2019 I went to my first Gay Pride parade expecting to see some hot, fit guys. Instead I got this. (imgur.com)
submitted 17 hours ago by DoctorZordersMD **Verified Shitlord**
347 comments **share**

3 704 "I am eating less and working out almost 2x a day and I'm gaining weight." (imgur.com)
submitted 13 hours ago by muppy_
164 comments **share**

4 822 An update on the hamplanet I went to highschool with that is in Tunisia to meet her online boyfriend. Original post in comments. (imgur.com)
submitted 15 hours ago by yrbuttz
175 comments **share**

came out as a shitlady on Facebook (imgur.com)

149,504 shitlords

○ 2,435 shitlords oppressing fatties

THIS IS AN 18+ SUBREDDIT!!!!!!!!!!

Figure 7–1: A screen shot of `r/fatpeoplehate` obtained through WayBack Machine dated June 7, 2015. It shows how users would link images of plus-sized individuals accompanied with mocking titles. It also shows the number of subscribers at that moment: 149,504.

we do is aimlessly bully and ridicule people", they maintained that their actions have a deeper meaning. They go on to state that obesity is a choice that indicates selfishness, lack of discipline and other mental weaknesses. They believed that creates an expensive toll on the society with disproportionate burdens on health care systems.

The subreddit grew to become very popular and had almost 150,000 subscribers at its peak. A screen shot of the subreddit is provided in Figure 7–1, obtained through WayBack Machine[5] . On June 10, 2015, Reddit admins announced that they were going to remove harassing subreddits [159]. They clarified that "*we will ban subreddits that allow their communities to use the subreddit as a platform to harass individuals when moderators don't take action. We're banning behavior, not ideas.*". This announcement marked the implementation of new community policy that allowed Reddit to ban or quarantine objectionable subreddits. This policy received mixed response.

In this paper, we present the reaction of the users affected by this ban and try to gauge the success of Reddit's community policy in controlling FPH-specific hateful speech on their platform.

7.1.4 Data

Jason Baumgartner, under the Reddit user name `Stuck_In_the_Matrix`[6] , has made available large data dumps that contain a majority of the content (posts and

[5] `web.archive.org/web/20150607141552/www.reddit.com/r/fatpeoplehate/`

[6] `www.reddit.com/user/Stuck_In_the_Matrix/`

comments) generated on Reddit[7] . The data dumps, collected using the Reddit API, are organized by month and year dating back to 2006. This resource is regularly updated with new monthly content.

For our research, we focused on June 10, 2015 as the pivot point since it was on this day that Reddit announced the banning of harassing subreddits [159], the most prominent of which was `r/fatpeoplehate`. At the time of its banning, the subreddit had almost 150,000 subscribers (Figure 7–1). Since we were interested in understanding the effects of banning such a large community, we studied the activity of users primarily associated with the FPH subreddit, over a four-month period, two months before and two months after it was banned. Thus, we restricted our analysis to the time period between the days of April 10, 2015 and August 10, 2015.

FPH users

We aggregated all users that commented in `r/fatpeoplehate` over our period of interest, resulting in 42,354 unique usernames. From this list we removed common bot accounts, such as `AutoModerator, autotldr, Mentioned_Videos`, etc. When users delete their accounts or the moderators remove comments, the authors of such comments would be displayed as `[deleted]`. We removed these from our list as well, since there is no way to link such comments to their original authors. In addition, this initial sample presumably included some (small) population of users who do not condone the FPH community's views but still comment in the subreddit for the sake of discussion or argument. In order to remove such accounts, we only considered

[7] `files.pushshift.io/reddit/`

Sample	**# of users**	**# of comments**
FPH	13,916	878,276
Random	13,916	585,632
`loseit`	-	159,250
`fatlogic`	-	266,456
`fatpeoplestories`	-	45,484

Table 7–1: Details on each of our data sample. Note that the three communities we investigate also differ in size, which can induce different behaviours.

users who had FPH as their most commented subreddit. This left us with 13,916 users, whom we refer to as FPH users in this paper. For each FPH user, we collected all the comments associated with them in the period of interest and call it the FPH sample.

Random users

We sought to compare the activity of FPH users with an equal-sized systematic random sample of Reddit users. To this end, we curated a list of all the Reddit users that posted a comment during the period of interest. After removing the major bot accounts and FPH participants, we had a list of 4,677,759 users. To obtain an equally sized subset we perform systematic sampling. To do so, we first sort our list of users by the number of comments they made during this period. We selected an initial user randomly from the list and the remaining users at regular interval from the last one, to obtain a sample of 13,916 random users. Again, for each random user, we collected all the associated comments in the relevant time period and call it the random sample. The details on the two samples are provided in Table 7–1.

Subreddit-specific comments

We were also interested in studying the effects of the ban on other communities, especially the ones with themes related to the banned community. Accordingly, we collected all the comments generated in:

`r/loseit`: a subreddit that supports people who want to lose weight.

`r/fatlogic`: a subreddit that presents itself as against the "*fat way*" of thinking. They describe themselves as a place for "*anything to do with fatlogic or anything else related to fatty logic. This is not a hate sub.* `r/fatpeoplehate` *is the place for that.*"

`r/fatpeoplestories`: a subreddit to share the stories about fat people. They clarify that "*We are NOT FatPeopleHate or FatPeopleObservations or StoryWithAFatPersonInIt. It'll get removed regardless of how long you spent typing it up. Repeat offenders will be banned.*"

Further details are provided in Table 7–1.

7.1.5 Study Design

At a high-level, we seek to characterize the effects of banning a prominent community on Reddit in two ways:

1. at the user level, we wish to study the effects on the active members of the banned community, and
2. at the community level, we want to study if the ban created any counter effects on other communities, specifically those that operate in a similar sphere as the banned community.

Banning and FPH users

We analyzed multiple aspects of FPH user population behavior, contrasting it with behavior exhibited by the random user group:

1. the difference in overall commenting behaviour between the two samples over our period of analysis;
2. the difference in the commenting behaviour of the two user groups before and after the ban;
3. the retention of users by the platform after the banning of the subreddit;
4. the effect on bad user behaviour in the form of downvoted content;
5. the change in user exploration patterns of new subreddits; and
6. the search for a an alternative subreddit for the disbanded community.

Banning and other communities

We also studied the direct effects of the ban on other communities that are relevant to the topic of the banned subreddit. Specifically, we studied `r/loseit, r/fatlogic and r/fatpeoplestories`. For each of these subreddits, we studied the downvoted comment volume, the deleted comment volume, and the overall comment volume during our four-month period of interest. As we discuss later, these metrics were chosen as proxies for the presence of FPH-specific negative content in other subreddits before and after the ban.

7.1.6 Results

Effect on FPH users

The overall commenting behaviour of the FPH sample is significantly different from that of the random sample. Figure 7–2 presents the difference in

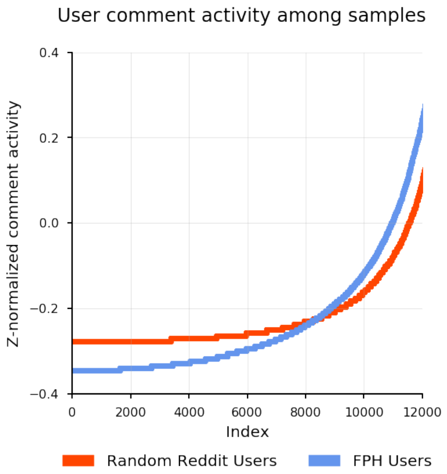


Figure 7–2: Z-normalized comment activity of the two samples. FPH and random users are ranked based on their total comment activity. Then the Z-normalized comment activity is plotted at each rank with FPH users indicated in blue and random users in orange. This figure contains the first 12,000 users for a clearer representation.

overall commenting behaviour between the two samples during the period of analysis. The two samples follow each other very closely, which is backed by the Pearson's correlation coefficient of 0.97. Comparing the two distributions by the Kolmogorov-Smirnov goodness-of-fit test returns a p-value of ¡0.0001. This indicates that the null hypothesis that the two independent samples are drawn from the same continuous distribution can be rejected, and that there is a significant difference in the overall behaviour of users in our two samples.

The comment activity of the FPH users declined after the ban, when compared to the random users. Figure 7–3 presents the effect of the community ban on the two samples individually through scatter plots and their respective regression lines. Figure 7–3c presents the regression lines together to emphasize the difference in the proportion of comment activity before and after the ban between the two user groups. From this figure we can infer that the FPH users commented more before the ban, than after, when compared to the random users. This shows that the ban caused a decrease in the Reddit interaction of FPH users.

To check the significance of this claim, we compare the difference in activity of the two samples. The distribution of the difference in depicted in Figure 7–3d. Since the distribution is normal, we can perform a paired t-test. For the random sample, the t-test returns a p-value of 0.7, and we fail to reject the null hypothesis that the two samples (pre-ban and post-ban comment activity) are similar. However, for the FPH sample, the paired t-test returns a p-value of 8.7e-197, suggesting that we can reject the null hypothesis and that there is a significant difference in the comment activity before and after the ban for FPH users.

A higher proportion of FPH users completely stopped comment engagement in the post-ban period. Figure 7–4a shows the overall trend of Reddit comment activity in the two samples after the ban. We can clearly observe that almost 1.75 times more FPH users became inactive after the banning of the subreddit, when compared to the random sample. Therefore, banning of the subreddit led to higher than average number of accounts with no direct comment engagement.

To delve deeper into this result, we can study the user-engagement of the FPH sample in their subreddit. In Figure 7–4b, we present the distribution of overall engagement in `r/fatpeoplehate` by the FPH sample. It is the distribution of users with $x\%$ of their overall comments having been generated in `r/fatpeoplehate`. From the cumulative distribution, we can assert that more than half the user base made

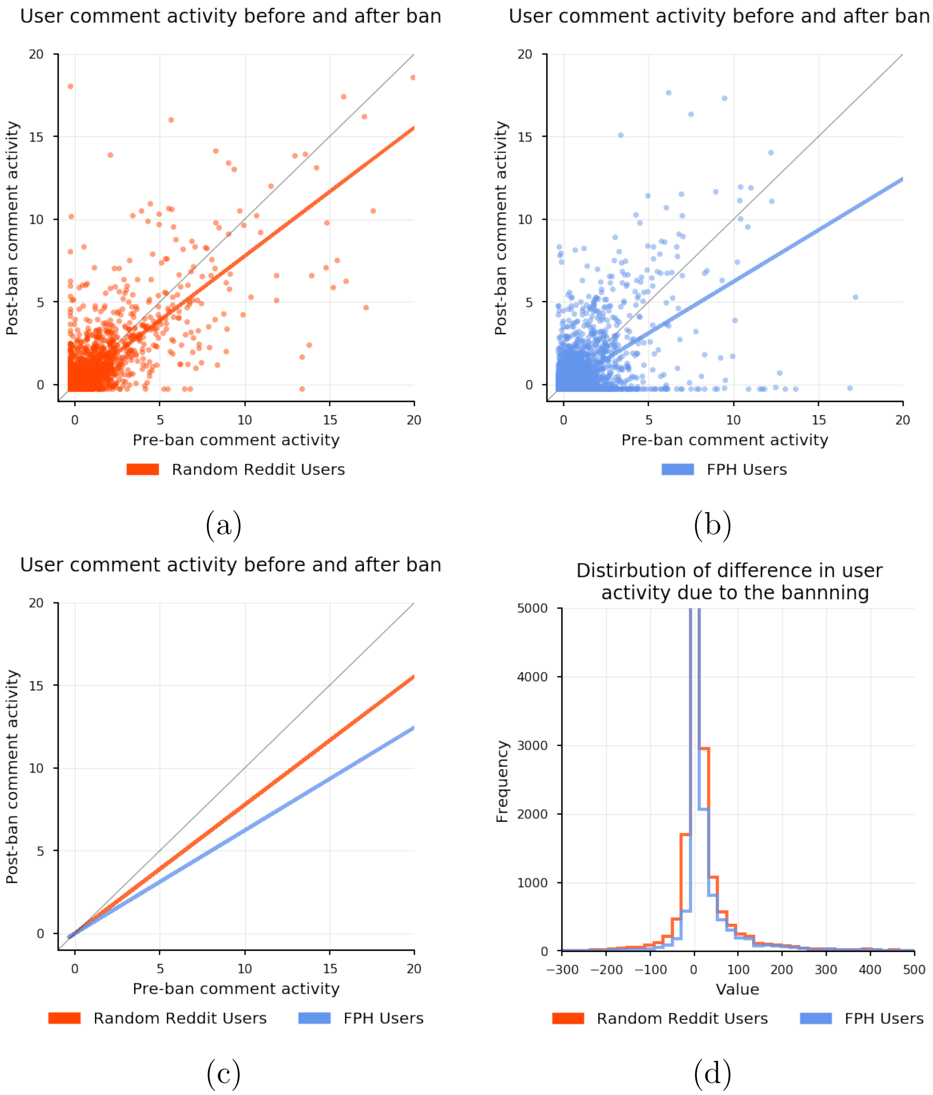
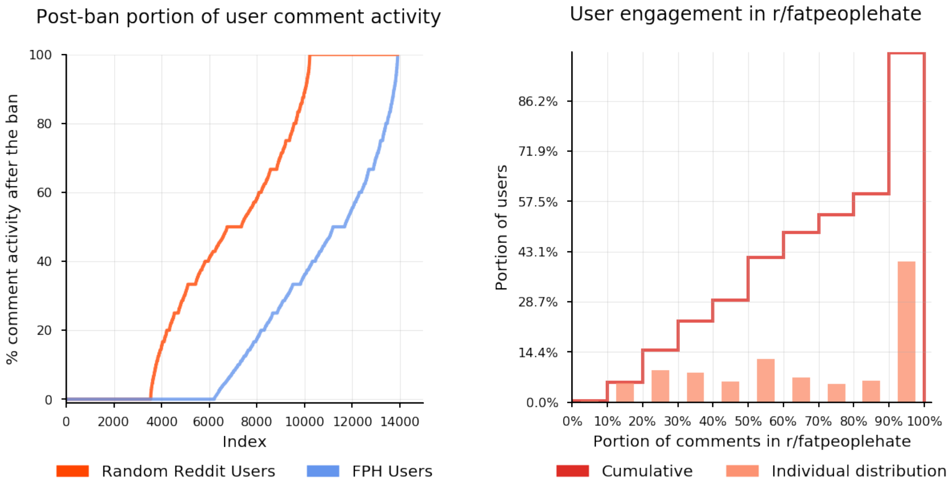


Figure 7–3: The comment activity of FPH and random users before and after the ban. Scatter plots of pre-ban vs post-ban comment activity of the two user groups with a linear model regression line. A 45 degree regression line indicates equal activity before and after the ban. Distributions of differences in user comment activity is provided in (d).

at least half of their comments in `r/fatpeoplehate`. This shows that our sample of FPH users were highly engaged in `r/fatpeoplehate` before the ban. Consequently,



(a) (b)

Figure 7–4: (a)Percentage of total comment activity after the ban. FPH and random users are ranked based on their total comment activity after the ban. At each rank the percentage of post-ban comment activity account out of the total activity is depicted. (b) Individual and cumulative distribution of FPH user engagement.

the ban led to a negative feedback towards this engagement. Hence, we see a steeper decline in the FPH user group, with more FPH users showing no comment engagement than random users.

This effect is also presented in Figure 7–3b. Many FPH users have high pre-ban comment activity and almost null post-ban activity. The scattered data points at the bottom of the y-axis and along the length of the x-axis are reflective of the portion of high-activity FPH users who stopped commenting after the ban. Notice a lack of similar pattern of data points in Figure 7–3a.

A smaller portion of FPH users exhibited negative user behaviour. Before we present these results, let us preface this by understanding negative user behaviour in context of r/fatpeoplehate. In this subreddit, users indulged in mocking plus-sized people. While malicious, this behaviour was appreciated and promoted within the subreddit. However, same behaviour would be considered negative outside of the subreddit.

Reddit introduced the banning of the subreddits, on the grounds of user harassment, to control such negative behaviour. While we do not have a measure for gauging hate speech across the entire platform, we can however, study the users that engaged in the harassing subreddit and observe if they continued their actions in other subreddits.

For this experiment, we use a Reddit feature called 'downvote'. Reddit is community-centric and the content generated in the community is monitored by the community as well. Positive user-behaviour is upvoted by the members while negative user behaviour is downvoted. Unfortunately, the downvotes are not reserved solely for harassing behaviour, but can also contain irrelevant or wrong content. Because it is difficult to separate these sub-categories, we inevitably study the overall negative behaviour by FPH users.

Figure 7–5 depicts a histogram of downvoted user activity by FPH users, before and after the ban. It represents that $y\%$ of the users had $x\%$ of their comments downvoted by others. We can make two meaningful observations from this plot:

1. For a vast majority of the FPH users, only a small portion of their comments are downvoted.
2. After the ban, the portion of users with less than 5% of downvoted comments further increases.

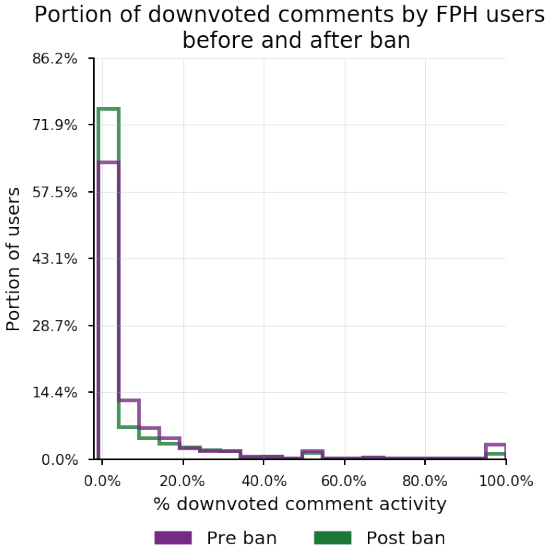


Figure 7–5: Distribution of proportion of downvoted comments in the FPH sample. It represents that $y\%$ of the users had $x\%$ of their comments downvoted by others.

Assuming that had the FPH users exhibited similar behaviour outside of the community, their comments would have been downvoted, the first observation suggests that majority of the FPH users indulged in fat-shaming behaviour, largely within the FPH subreddit. Therefore, we can assert that the `r/fatpeoplehate` was rather self-contained. The second observation may stem from the fact that a decrease in user comment activity after the ban would also reduce the number of downvoted comments and, therefore, increase the portion of users with low downvoted comment percentage.

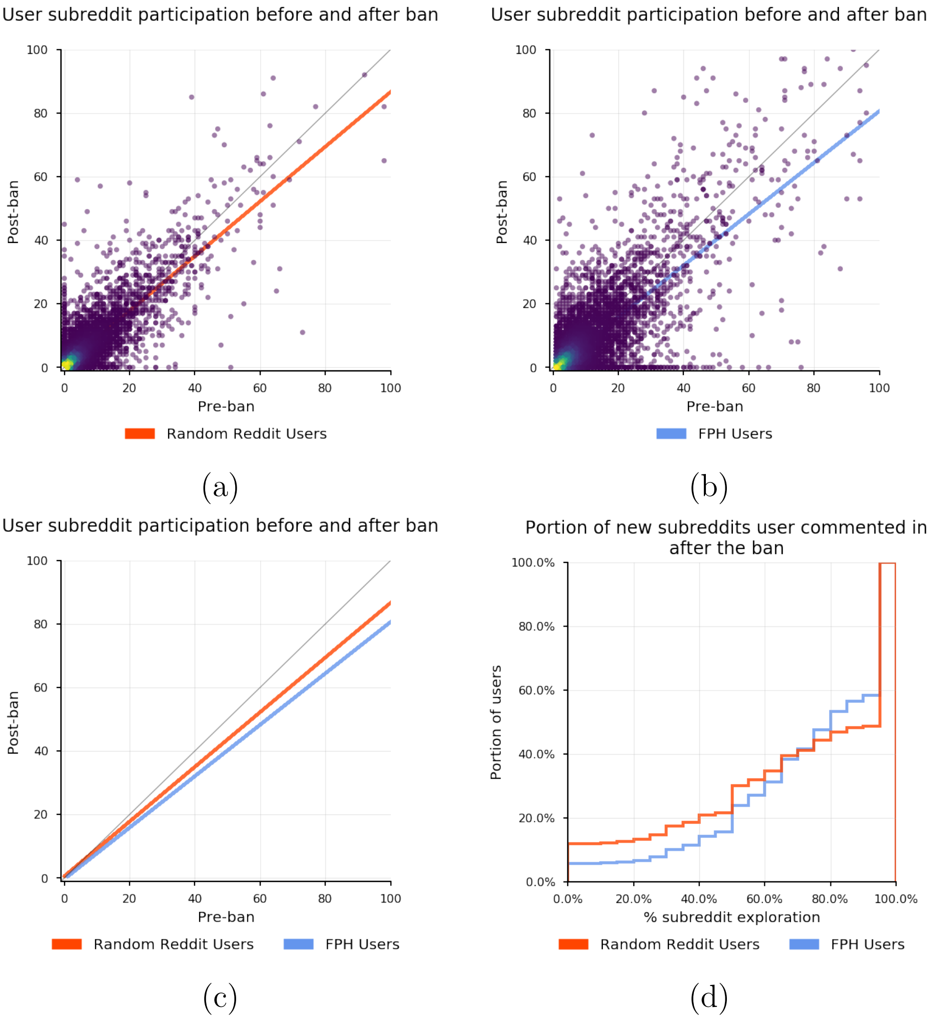
Since downvoted comments do not directly suggest harassing / hateful speech, we manually labelled 100 downvoted FPH user comments from before and after the ban. 100 FPH users were randomly selected for this purpose. We found that while 13 of the 100 comments exhibited FatPeopleHate-like behaviour before the ban, the number rose to 25 after the ban. It should be noted that both samples contained comments from subreddits other than `r/fatpeoplehate` and were manually labelled by an expert.

Therefore, while the overall negative behaviour declined, the hateful behaviour of FPH users was spilling more in other subreddits after the ban.

Reddit users explored other subreddits more post-ban, compared to the random sample. Users in the FPH sample were highly engaged and associated with `r/fatpeoplehate`. Banning of the subreddit breaks this association, which can cause users to look for other venues to continue being engaged with the theme of the subreddit. It can also cause users to completely disengage if the subreddit was one of the primary reasons for their use of the platform or continue the engagement with the rest of the platform as before, if they are still interested in rest of the content.

We wanted to study if the ban on their primary subreddit affected multi-subreddit participation of FPH users. Figure 7–6 presents how many subreddits users participated in, before and after banning of `r/fatpeoplehate`, for both samples. The general trends in the form of linear regression lines are presented together in Figure 7–6c. The behaviour for the two samples is quite similar. Therefore, the banning of `r/fatpeoplehate` did not produce a major difference in how many subreddits FPH users participated in.

While the FPH users continued to participate in an average number of subreddits, we investigated if they were exploring Reddit by participating in subreddits they were not participating in, before the ban, or were they less likely to participate in newer subreddits due to the loss of a subreddit they strongly associated with.



(a)

(b)

(c)

(d)

Figure 7–6: User subreddit participation. The number of subreddits that users commented in before and after the ban. A linear fit is also generated. A 45 degree line indicates equal number of subreddits before and after the ban. (d) New subreddit exploration - percentage of users with atmost x% of new subreddits after the ban.

In Figure 7–6d, we observe that 80% of FPH users were now participating in at least 50% new subreddits after the ban, as compared to 40% for the random sample. Further more, the FPH sample had a larger portion of users with high subreddit exploration than the random sample. Therefore, FPH users were participating in more new subreddits after the ban.

FPH sample continued to be interested in the subreddits it was previously active in. In Table 7–2, we present the subreddits that were popular amongst the users of FPH user sample. From the two lists, we can ascertain that FPH users were still highly interested in the subreddits they used to frequent before. So, even through they were exploring new subreddits, the group as a whole was still actively participating in the subreddits they used to frequent before the ban. However, it is important to adress the fact that these subreddits are popular across Reddit and therefore remain popular after the ban as well.

Pre-ban	**Post-ban**
`fatpeoplehate`	`AdviceAnimals`
`WTF`	`WTF`
`AdviceAnimals`	`fatlogic`
`fatlogic`	`TumblrInAction`
`TalesofFatHate`	`pcmasterrace`
`TumblrInAction`	`BlackPeopleTwitter`
`pcmasterrace`	`trashy`
`BlackPeopleTwitter`	`punchablefaces`
`trashy`	`trees`
`trees`	`technology`
`ImGoingToHellForThis`	`ImGoingToHellForThis`
`cringepics`	`cringepics`
`4chan`	`KotakuInAction`
`fatpeoplestories`	`relationships`
`punchablefaces`	`fatpeoplestories`
`thebutton`	`4chan`
`FitshionVSFatshion`	`politics`
`atheism`	`conspiracy`
`AdiposeAmigos`	`SubredditDrama`
`relationships`	`Tinder`

Table 7–2: Popular subreddits amongst the FPH users before and after the ban.

FPH users tried to actively create alternative subreddits for their banned community. The FPH community did not readily accept the banning of their subreddit. They actively tried to circumvent the banning by creating new subreddits to act as an alternative to `r/fatpeoplehate`. In Table 7–3, we present 99 subreddits that were created as reaction to the banning of the initial subreddit. However, Reddit admins were able to control this surge and banned a majority of these alternatives as well. The ones that were not banned are inactive, most likely due to not being known to the community. Therefore, while FPH community reacted by creating a

```
CandidDietPolice
FatPeopleDislike
fatpeopledislike1
fatpeoplehate1
fatpeoplehate10
fatpeoplehate100
FatPeopleHate1000
FatPeopleHate10000
fatpeoplehate101
fatpeoplehate102
fatpeoplehate11
fatpeoplehate12
fatpeoplehate13
fatpeoplehate14
fatpeoplehate15
fatpeoplehate16
fatpeoplehate17
fatpeoplehate18
fatpeoplehate19
fatpeoplehate2
fatpeoplehate20
Fatpeoplehate200
fatpeoplehate21
fatpeoplehate22
fatpeoplehate23
fatpeoplehate24
FatPeopleHate25
fatpeoplehate26
fatpeoplehate27
Fatpeoplehate28
fatpeoplehate29
Fatpeoplehate3
fatpeoplehate30
FatPeopleHate300
fatpeoplehate31
FatPeopleHate314
FatPeopleHate32
fatpeoplehate33
fatpeoplehate34
fatpeoplehate35
fatpeoplehate36
Fatpeoplehate37
Fatpeoplehate38
FatPeopleHate39
fatpeoplehate4
fatpeoplehate40
fatpeoplehate41
fatpeoplehate42
fatpeoplehate420
fatpeoplehate43
fatpeoplehate44
fatpeoplehate442
Fatpeoplehate45
fatpeoplehate46
fatpeoplehate47
fatpeoplehate48
fatpeoplehate49
fatpeoplehate5
fatpeoplehate50
fatpeoplehate51
fatpeoplehate52
fatpeoplehate54
fatpeoplehate55
fatpeoplehate58
fatpeoplehate59
fatpeoplehate60
fatpeoplehate61
fatpeoplehate62
fatpeoplehate64
fatpeoplehate68
FatPeopleHate69
fatpeoplehate7
FatPeopleHate70
fatpeoplehate73
fatpeoplehate75
FatPeopleHate77
fatpeoplehate8
Fatpeoplehate80
fatpeoplehate88
Fatpeoplehate9
fatpeoplehate90
fatpeoplehate9000
Fatpeoplehate97
fatpeoplehate98
fatpeoplehate99
fatpeoplehateFFFFFFFF
FatpeoplehateX
fatpersonhate
fatpersonhate1
fattypeople
landwhaledistaste
landwhaleh8
ObesePeopleDislike
obesepeoplehate
ObesityAnger
thinpersonlove
wedislikefatpeople
WellInsulatedPeople
whalepeoplehate
```

Table 7–3: A list of subreddits created as alternatives to `r/fatpeoplehate`

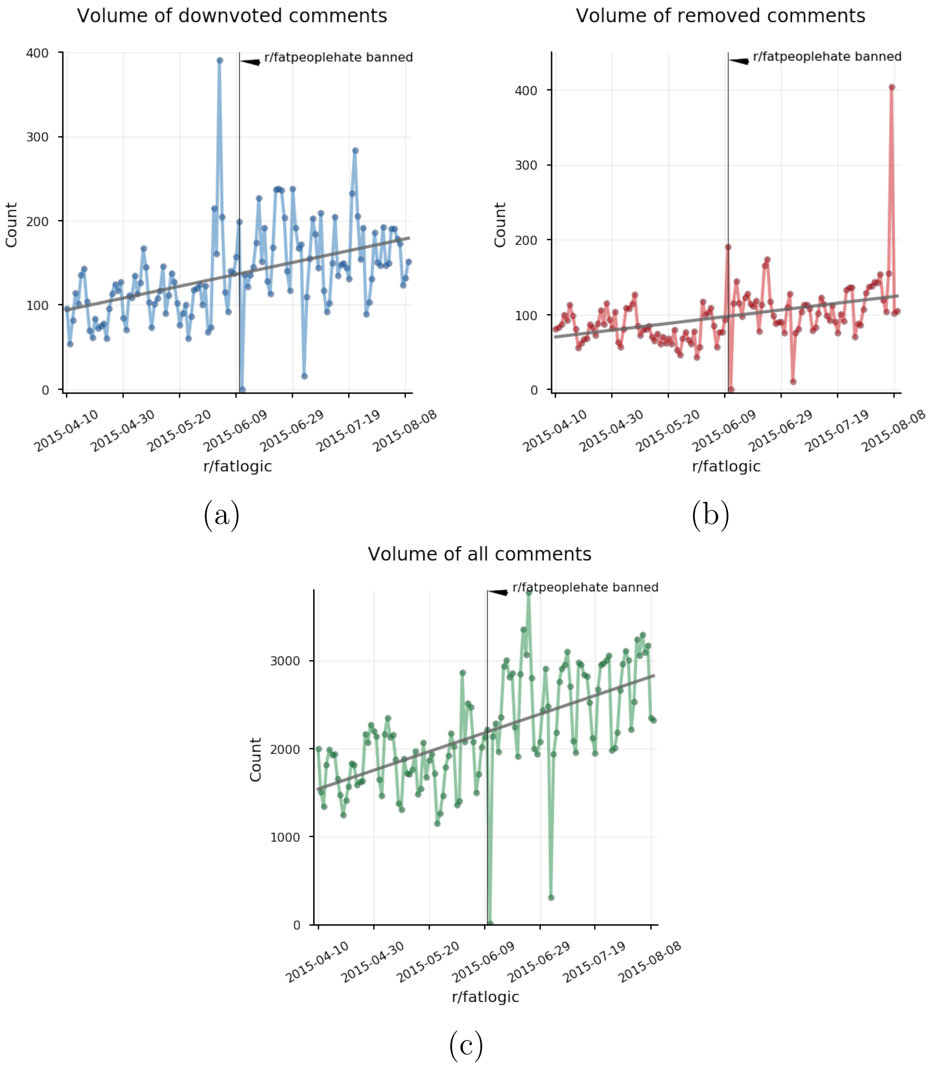
multitude of alternatives, Reddit administration was comprehensive in their policy and banned a majority of them. It is important to note that it is likely that other alternatives were also created and are not present in the list. In case it did happen, the created subreddit is not common knowledge.

Effect on weight-related communities

`r/fatlogic` and `r/fatpeoplestories` are also popular subreddits themes similar to `r/fatpeoplehate`. While both subreddits make it clear that they are not associated with `r/fatpeoplehate`, they encourage content that mocks plus-sized people. On the other end of the spectrum is `r/loseit`, a weight loss subreddit where fat people are supported. For each, we study the volume of downvoted comments, deleted/removed comments and all comments, during the period of analysis.

`r/fatlogic` In Figure 7–7), we notice zero activity following the FPH ban. On further investigation, we find that moderators of the subreddit made it private [146]. In a post made by the moderators, they explain how in wake of the banning, their subreddit was flooded with users from FPH. Being overwhelmed by the traffic, and not wanting to be banned for mistaken association, the mods temporarily made the subreddit inaccessible. An excerpt from the post is provided below:

> "*And then, on the morning of June 10, 2015, fatpeoplehate was banned by reddit, We were given no warning and we immediately were overcome with posts from indignant fph subscribers, posts worrying if we were next and it became way too much work for a our small mod team to handle so we went private for a few days.* "



(a)

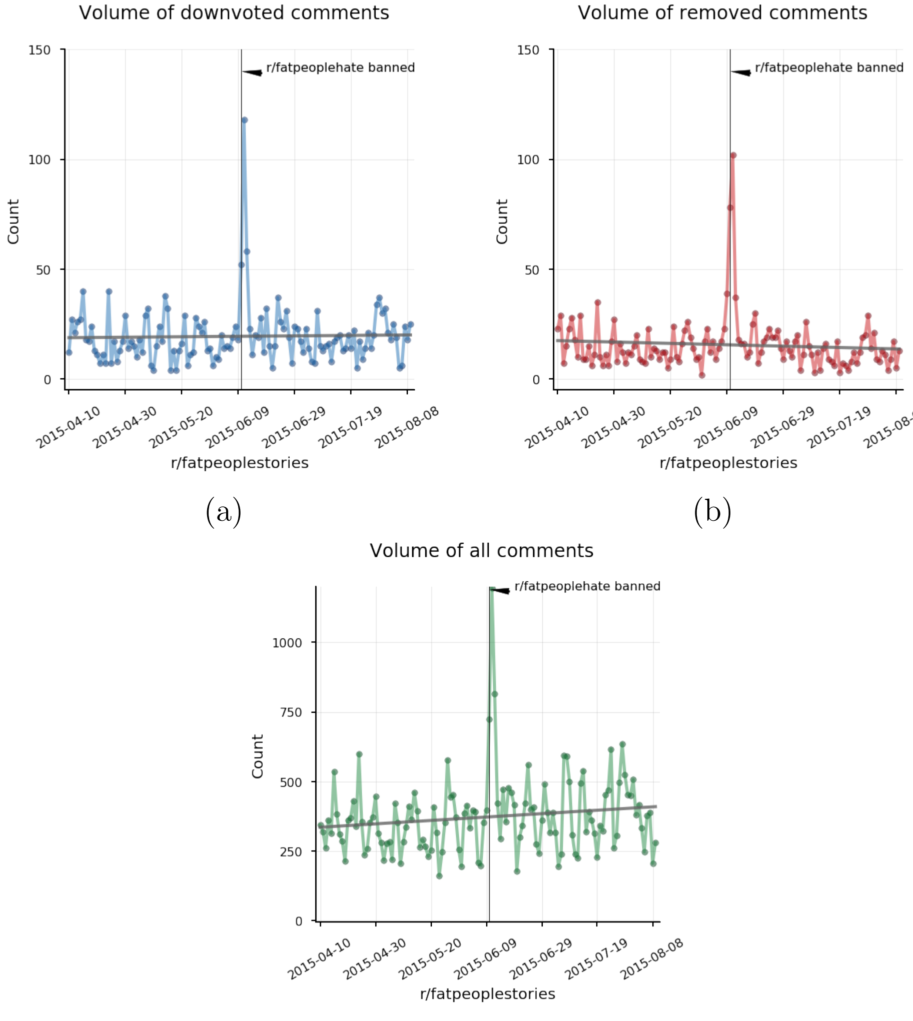
(b)

(c)

Figure 7–7: Volume analysis of `r/fatlogic`. There is a dip in volume right after the ban.

Post-ban, FPH users surged into `r/fatlogic`. This migration was checked by the moderators by making the subreddit private.

`r/fatpeoplestories` remained public (Figure 7–8). Although, we do witness a spike in the total number of comments posted on the subreddit, following the ban, it is short-lived and normal traffic patterns resume soon after. We see a spike in both the number of downvoted comments and the number of deleted comments. We can assert that the community actively checked the surge in traffic by downvoting more than


Volume of downvoted comments

Volume of removed comments

Volume of all comments

(a)

(b)

(c)

Figure 7–8: Volume analysis of `r/fatpeoplestories`. The volume parameters resume normal values after the surge.

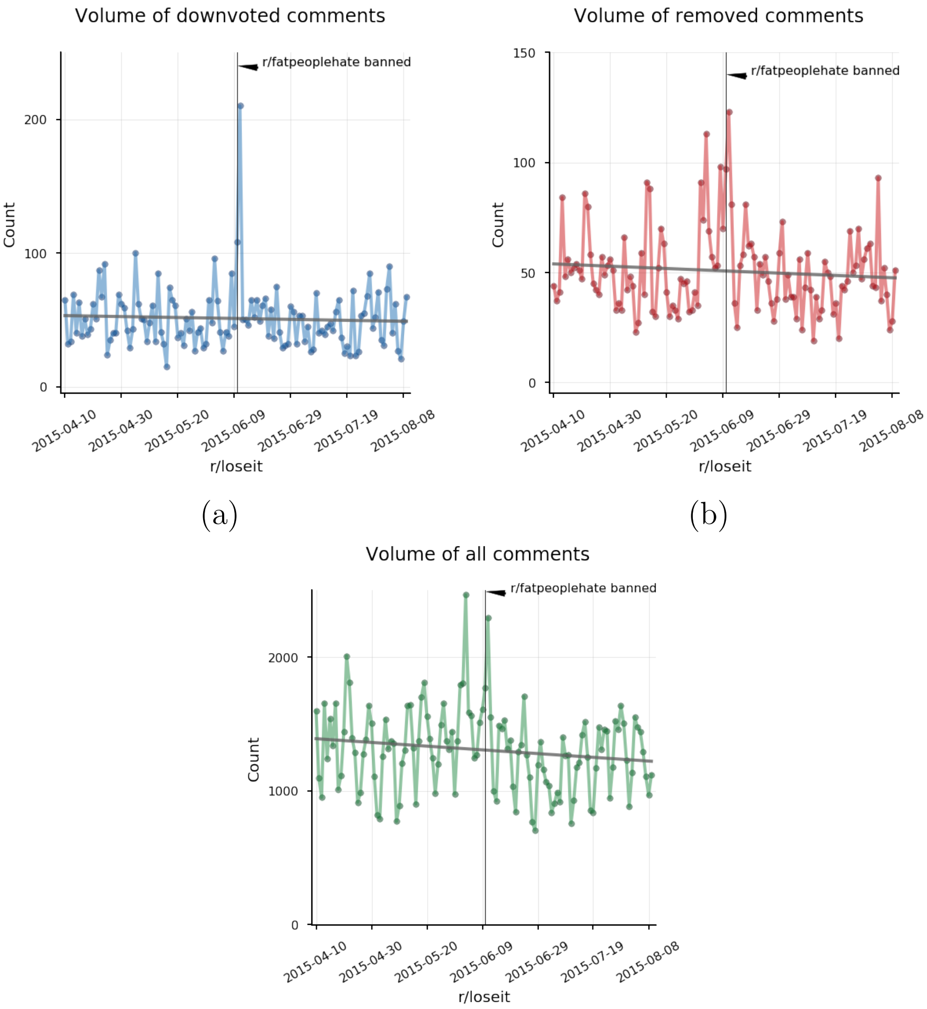
usual. Similarly, moderators also actively kept the migration in check by deleting comments that went against the subreddit.

`r/loseit` In Figure 7–9 we observe that a higher than normal number of comments were posted on the subreddit following the FPH ban. However, since the demographic of the community is plus-sized people, these comments also include the discussion on the banning of a subreddit that was abusive to them. Nonetheless, we also notice that a higher than average number of comments were deleted and that there is a major spike in the volume of downvoted comments. Therefore, we can say that `r/loseit` did face increased negative user behaviour following the ban of FPH. Yet, in this case as well, the community and the moderators were actively checking for such behaviour and soon after, normal characteristics resumed.

7.1.7 Discussion

In this research, we studied (1) how the banning of a large hateful subreddit affected its active members and (2) whether the ban drove hateful content into other subreddits.

We discovered that the banning had the intended effect on FPH-specific activity on the platform. Not only did we observe a significant decrease in the user comment activity of FPH users after the ban, we also found that a larger portion of the users completely stopped engagement. Though Reddit was able to sustain this disengagement, it is not necessarily a positive outcome in every scenario. For new platforms or niche platforms with limited user base, user disengagement, if in large enough numbers, can lead to the demise of the platform.



(a)

(b)

(c)

Figure 7–9: Volume analysis of `r/loseit`. There is a higher than usual count of downvoted and deleted comments right after the ban.

While we cannot directly state if FPH users generated less hate speech after their community was banned, we can say that the community of users was less active on Reddit after it.

Furthermore, with the reduced user activity post-ban, in the long-term, the overall negative user activity (in the form of comments that were downvoted by other users) also decreased. However, hand-labelling of a selected set of comments suggested that the amount of hateful speech in the downvoted comments increased. That is, while users were producing fewer comments with downvotes, these comments contained more hateful speech.

It is important to note that any user behaviour is positive or negative within the context of a community. So, the same comment that is applauded in one subreddit can be removed from another. In the context of FatPeopleHate, all the hate speech generated in the subreddit was positive behaviour for the community and would have been upvoted. After the banning of the subreddit, users could not garner similar behaviour in other subreddits, which is why we do not have instances of upvoted hate speech. Therefore, even though the portion of hate speech in downvoted comments increased, counter-intuitively the overall volume of hate speech decreased.

As for the effect on other communities, we observed that FPH users continued to be interested in other subreddits in which they used to be active. Even so, they flooded `r/fatlogic` since it is among the subreddits most closely related to the theme of FatPeopleHate. But their moderators were not keen with this surge and made the subreddit private. `r/fatpeopelstories` and `r/loseit` also witnessed a spike in user activity. The spike was accompanied with spikes in the number of comments that were downvoted by the community and the number of comments that were deleted, in part, by moderators. In short, the FPH users surged into other subreddits post FPH ban. Although, these communities were able to neutralise this migration through a combined action of the members (in the form of downvoting negative behaviour) and by the moderators (by deleting comments or restricting access to the subreddit itself). So while there was a strong immediate reaction to the banning of FatPeopleHate, other communities were able to withstand it and normal behaviour resumed soon after.

Furthermore, FPH users did not just surge into existing communities. They also actively counteracted by trying to create alternative subreddits for their banned community. However, Reddit administration was comprehensive and banned a large number of these offspring subreddits. Even though some of these alternatives managed to escape the ban, they remain inactive. Reddit was, thus, able to thwart a possible fallout from the banning of `r/fatpeoplehate` by a combined action of its users, its moderators and its administrators.

Overall, our findings confirm that banning a community can be an effective strategy for diminishing hateful content, if most of the hateful content is confined within that particular community. Of course, the story might be quite different for harassing communities whose negative behavior has already infiltrated other communities. The latter can prove more difficult to curb at the administrative level but might be restrained at the user/moderator level through regular action, in view to the fact that other subreddits demonstrated a remarkable resistance to the incursion of FPH content. This suggests a banning strategy needs to be accompanied by strong moderation of other communities and negative reinforcement of hateful-content by other users for it to be a success.

7.1.8 Conclusion

In this article, we studied the outcome of a common practice among online forums of banning unsavoury activities. While it is more common for group moderators to ban unwanted users, we specifically studied the banning on a larger scale, that of an entire sub-community. Banning of a large and popular sub-group is not a risk-free endeavour. A priori banning hardly guarantees that the content produced by the community will go away. Nevertheless, we observed that it does discourage users from interacting with the platform, especially if the negative reinforcement lingers on. Since sub-communities promote associated sub-culture of the online forum, banning it takes away a major association that users had with the platform itself.

In this case, Reddit was successful in banning FatPeopleHate. Other platforms with sub-communities can also implement similar policies for better control over how the general user experiences their platform. However, depending on the scale of the platform, it is going to be harder for the platform maintainers to control the creation of offspring sub-communities. For example, Facebook deals with a much larger user base and banning of a popular page would result in the creation of many more alternative pages. It would also be harder to control groups that have off-platform / real-life connections since they can persist. Nevertheless, banning the community can provide significant control to platform maintainers.

[146] maybesaydie. *A very belated introduction to the entire /t/fatlogic mod team.* 2015. URL: `www.reddit.com/r/fatlogic/comments/3mt2w4/a_very_belated_introduction_to_the_entire/`.

[159] Jessica Moreno, Ellen Pao, and Alexis Ohanian. *Removing harassing subreddits*. 2015. URL: `www.reddit.com/r/announcements/comments/39bpam/removing_harassing_subreddits/`.

[194] Haji Mohammad Saleem et al. "A web of hate: Tackling hateful speech in online social spaces". In: 2016.

[END OF MANSCRIPT]

The previous manuscript does not attempt to detect abusive content but rather observers the aftermath of a large scale platform intervention that focuses on antagonistic communities. The case study establishes that the banning of `r/fatpeoplehate` was an overall successful operation for the Reddit admins.

However, the full picture is a more nuanced. While the ban decreased objectionable behaviour, it also reduced engagement from the affected user base. In fact, a portion of the users ceased all interaction. In a separate study, I observe users that were unhappy platform interventions chose to migrate to other platforms with lax moderation [165]. The affected users also spilled over and disrupted adjacent communities, which required intervention from moderators to either restrict subreddit access or remove a large volume of unwelcome comments. The active users from these communities also voted down the surge of comments and Reddit admins banned multiple additional subreddits that mushroomed to replace `r/fatpeoplehate`.

The ban therefore required multiple decisive actions from Reddit admins, moderators and users for it to be successful. Overall it pushed the users of an antagonistic community onto other platforms.

CHAPTER 8
Conclusion

Detection of abusive language is absolutely critical for maintaining safe online spaces. However, it is not an easy task. In Chapter 2, I present the many challenges that abusive language research faces. I believe that the inherent diversity in abuse is one of the biggest hurdles in creating reliable detection frameworks. These frameworks largely rely on supervised machine learning and therefore require representative training resources. However, it is challenging to collect abusive language in a manner that fully represents its diversity. Much of this thesis is therefore dedicated to explicitly introducing diversity in abusive language research.

This thesis begins by tackling representative collection of abusive language. Through comparison of online communities that self-identify as either antagonistic or supportive of marginalized populations, I have shown that abusive and supportive language share vocabulary which leaves keyword detection largely ineffective. However, these communities contain diverse perspectives towards marginalised populations. I have further shown that the language from these communities is cohesive, even across different platforms but only within a particular target group. These self-identifying communities can and should be leveraged to collect diverse data.

While I was able to identify a potential solution for aggregating diverse data, I lacked a formal framework to precisely address it. Building on pejorative expressions

I have constructed a taxonomy that distinguishes such language in 4 major categories and 12 minor categories. This taxonomy adds nuance and allows for a deeper engagement with abusive language. I strategically sampled comments from a range of online communities - both supportive and antagonistic - to ensure diversity of content. I then assembled and trained a demographically diverse cohort to annotate the aggregated data. Community sampling was successful and resulted in an even split between derogatory and non-derogatory comments, even though all comments contain pejorative expressions.

The taxonomy as well as community sampling helped in assembling a diverse corpus. However, this corpus did not address all questions regarding the construction of such resources. I followed up with a much broader analysis in which I collected data form different platforms with different slurs and then annotated them using the 12 fine categories from the taxonomy. By aggregating the results, I have shown that both filtering keywords and source platforms play defining roles in the resulting corpus, where not only do they control for the amount but also the type of language captured. The heterogeneous nature of the collected data further illustrated the diversity in abusive language.

With the corpus assembled, next I focused on the detection of abusive language. During that time, an emerging theme in the research community was the discovery of biased frameworks that mis-classify marginalized perspectives as abusive. I tackled the problem of false-positive labels by investigating a new source of contextual information - that of the online community in which a comment was generated. I have shown that basic as well as complex language models benefit from access to community context. By utilizing the taxonomy labels in the error analysis, I have shown that context improves performance on appropriative language and also assists humans to make better judgements. Community context is a promising avenue for future research.

Finally, during the course of this work, I was provided with an opportunity to study large-scale effort in abuse moderation. In 2015, Reddit banned a large communities - `r/fatpeoplehate` for harassment. This made it possible to observe the reaction of users that heavily engaged with a 'deplatformed' community. My analysis revealed that community banning reduced the engagement of affected users but the ones that remained tried to infiltrate similar communities or create another space for themselves - both of these actions eventually failed due to additional actions taken by Reddit users, moderators, and administrators.

Looking forward, the future of abusive language research lies in accepting and engaging with its diverse nature. This thesis supports the construction and utilization of diverse datasets. Community sampling allows the aggregation of perspectives from people who condone and celebrate abusive behaviour as well as from those who condemn and abhor it. Furthermore, viewing abusive language through the presented taxonomy promotes integration of such diversity in its analysis. The taxonomy helps examine the composition of different datasets and also assists fine-grained evaluation of detection frameworks. Combined, these tools are the first steps towards and a principled engagement with abusive language as a phenomenon.

The contributions of this thesis suggest several avenues to further advance abusive language research. First, instead of trying to find one solution that fits all forms of abusive language, a more fruitful approach would be to build modules that excel at identifying abuse against a particular marginalized group. As I have shown in Chapter 3, abusive language detection fails when tested out of domain. Building specialized modules, for example, a module that only detects cases of racism against black people, or a module that only focuses on transphobia, would help limit the scope of the problem, control for diversity, and allow customization of detection frameworks. Limiting to a singular target adds nuance where we are able to understand the different forms of abuse that affect such a marginalized population in more detail.

Second, context plays a critical role in enhancing the credibility of detection frameworks. Contextualizing language reduces the ambiguity in its interpretation. In Chapter 6, I highlight how community context can help protect the perspectives of marginalized populations. However, it is only a singular facet in the large context around online conversations. Other researchers have provided evidence towards the value of profiling authors and accounting for neighbouring comments in detection of abusive language. However, an untapped frontier in this body of research is amalgamation of different sources of contextual information. We require broad research that analyzes different aspects of context and how they inform abuse individually and collectively.

Finally, abusive language is an evolving phenomenon. The definition of abuse is derived from the prevailing social norms. Commonly used phrases can be deemed unacceptable as we socially progress. For example, Brazil nuts were sometimes referred to as *n*gger* toes but gradually fell out of use as the slur became socially unacceptable. Furthermore, propagators of abuse generate new ways to derogate their targets. For example, abusive communities created proxy-labels in the form of popular internet brands to refer to marginalized populations. Words such as 'Google', 'Yahoo', 'Skype', and 'Bing' were used to refer to black, Mexican, Jewish, and Chinese people respectively. However such euphemisms are not one off. `Hatebase`, a collection of hateful vocabulary continues to add new pejorative terms to its database. The term *MooSlime* was recently added, which refers to Muslims in a derogatory manner. With this constant evolution, abusive language research cannot rely on static resources. Instead, we need language resources to evolve alongside the abuse they represent.

Understanding and integrating the complete diversity of abusive language is a difficult problem that will require a great deal more research. However, the work presented in this thesis materially advances abusive language research by providing the means to ensure and address diversity in abuse.

Bibliography

[1] Aseel Addawood et al. “Telling apart tweets associated with controversial versus non-controversial topics”. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. 2017, pp. 32–41.

[2] Swati Agarwal and Ashish Sureka. “But i did not mean it!—intent classification of racist posts on tumblr”. In: *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE. 2016, pp. 124–127.

[3] Betty van Aken et al. “Challenges for Toxic Comment Classification: An In-Depth Error Analysis”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 33–42.

[4] James Allan. “The Harm in Hate Speech”. In: *Constitutional Commentary* 29.1 (2013), pp. 59–80.

[5] Pedro Alonso, Rajkumar Saini, and György Kovács. “Hate speech detection using transformer ensembles on the hasoc dataset”. In: *International Conference on Speech and Computer*. Springer. 2020, pp. 13–21.

[6] Luvell Anderson and Ernie Lepore. “Slurring words”. In: *Noûs* 47.1 (2013), pp. 25–48.

[7] Lora Aroyo et al. “Crowdsourcing subjective tasks: the case study of understanding toxicity in online discussions”. In: *Companion proceedings of the 2019 world wide web conference*. 2019, pp. 1100–1105.

[8] Zahra Ashktorab and Jessica Vitak. “Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers”. In: *Proceedings of the 2016 CHI conference on human factors in computing systems.* 2016, pp. 3895–3905.

[9] Lauren Ashwell. “Gendered slurs”. In: *Social Theory and Practice* 42.2 (2016), pp. 228–239.

[10] Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. “Stereotypical bias removal for hate speech detection task using knowledge-based generalizations”. In: *The World Wide Web Conference.* 2019, pp. 49–59.

[11] Koray Balci and Albert Ali Salah. “Automatic analysis and identification of verbal aggression and abusive behaviors for online social games”. In: *Computers in Human Behavior* 53 (2015), pp. 517–526.

[12] Michele Banko, Brendon MacKeen, and Laurie Ray. “A unified taxonomy of harmful content”. In: *Proceedings of the fourth workshop on online abuse and harms.* 2020, pp. 125–137.

[13] Jamie Bartlett et al. “Anti-social media”. In: *Demos* (2014), pp. 1–51.

[14] Valerio Basile et al. “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation.* Association for Computational Linguistics, 2019, pp. 54–63.

[15] Jason Baumgartner et al. “The pushshift reddit dataset”. In: *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 14. 2020, pp. 830–839.

[16] Anat Ben-David and Ariadna Matamoros Fernández. “Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain”. In: *International Journal of Communication* 10 (2016), p. 27.

[17] Emily M Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.

[18] Susan Benesch et al. “Counterspeech on Twitter: A field study”. In: *A report for Public Safety Canada under the Kanishka Project* (2016).

[19] Michael S Bernstein et al. “4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.” In: *ICWSM*. 2011, pp. 50–57.

[20] Claudia Bianchi. “Slurs and appropriation: An echoic account”. In: *Journal of Pragmatics* 66 (2014), pp. 35–44.

[21] Michał Bilewicz et al. “Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment”. In: *Aggressive behavior* 47.3 (2021), pp. 260–266.

[22] Sravan Bodapati et al. “Neural Word Decomposition Models for Abusive Language Detection”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 135–145.

[23] Robert J Boeckmann and Jeffrey Liew. “Hate speech: Asian American students’ justice judgments and psychological responses”. In: *Journal of Social Issues* 58.2 (2002), pp. 363–381.

[24] Aditya Bohra et al. “A dataset of hindi-english code-mixed social media text for hate speech detection”. In: *Proceedings of the second workshop on computational modeling of people’s opinions, personality, and emotions in social media.* 2018, pp. 36–41.

[25] Renée Jorgensen Bolinger. “The pragmatics of slurs”. In: *Noûs* 51.3 (2017), pp. 439–462.

[26] Tolga Bolukbasi et al. “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”. In: *Advances in neural information processing systems.* 2016, pp. 4349–4357.

[27] Lorraine Bowman-Grieve. “Exploring “Stormfront”: A virtual community of the radical right”. In: *Studies in conflict & terrorism* 32.11 (2009), pp. 989–1007.

[28] Johannes Breuer. “Hate speech in online games”. In: *Online Hate Speech. Perspektiven auf eine neue Form des Hasses. Kopaed: Düsseldorf* (2017), pp. 107–112.

[29] Judith Bridges. “Gendering metapragmatics in online discourse:“Mansplaining man gonna mansplain. . . ”” In: *Discourse, Context & Media* 20 (2017), pp. 94–102.

[30] Mary Bucholtz and Kira Hall. “Identity and interaction: A sociocultural linguistic approach”. In: *Discourse studies* 7.4-5 (2005), pp. 585–614.

[31] Leah Burch. “‘You are a parasite on the productive classes’: online disablist hate speech in austere times”. In: *Disability & Society* 33.3 (2018), pp. 392–415.

[32] Pete Burnap and Matthew L Williams. “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making”. In: *Policy & internet* 7.2 (2015), pp. 223–242.

[33] Peter Burnap and Matthew Leighton Williams. “Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making”. In: *Internet, Policy & Politics* (2014).

[34] Isabel Cachola et al. “Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media”. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 2927–2938.

[35] Justin Caffier. “Here Are Reddit’s Whiniest, Most Low-Key Toxic Subreddits”. In: *Vice.com* (2017). URL: www.vice.com/en_us/article/dyz377/trump-supporters-at-his-tulsa-rally-wrote-a-terrible-trump-2020-reelection-song.

[36] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. “Semantics derived automatically from language corpora contain human-like biases”. In: *Science* 356.6334 (2017), pp. 183–186.

[37] Elisabeth Camp. “Sarcasm, pretense, and the semantics/pragmatics distinction”. In: *Noûs* 46.4 (2012), pp. 587–634.

[38] Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. “DeepHate: Hate speech detection via multi-faceted text representations”. In: *12th ACM Conference on Web Science*. 2020, pp. 11–20.

[39] Allison T Casar. “Queer Stance: Metalinguistic attitudes towards slur reclamation among LGBTQ young adults”. In: *Lavender Languages AND Linguistic Conference* (2021).

[40] Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. “Pay “attention” to your context when classifying abusive language”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 70–79.

[41] Despoina Chatzakou et al. “Mean birds: Detecting aggression and bullying on twitter”. In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 13–22.

[42] Hao Chen, Susan Mckeever, and Sarah Jane Delany. “Presenting a labelled dataset for real-time detection of abusive user posts”. In: *Proceedings of the International Conference on Web Intelligence*. ACM. 2017, pp. 884–890.

[43] Mengtong Chen, Anne Shann Yue Cheung, and Ko Ling Chan. “Doxing: What adolescents look for and their intentions”. In: *International journal of environmental research and public health* 16.2 (2019), p. 218.

[44] Sapna Cheryan and Galen V Bodenhausen. “When positive stereotypes threaten intellectual performance: The psychological hazards of “model minority” status”. In: *Psychological Science* 11.5 (2000), pp. 399–402.

[45] Patricia Chiril et al. “He said “who’s gonna take care of your children when you are at ACL?”: Reported Sexist Acts are Not Sexist”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4055–4066.

[46] Shivang Chopra et al. "Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 386–393.

[47] Arijit Ghosh Chowdhury et al. "# YouToo? detection of personal recollections of sexual harassment on social media". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2527–2537.

[48] Arijit Ghosh Chowdhury et al. "Speak up, fight back! detection of social media disclosures of sexual harassment". In: *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*. 2019, pp. 136–146.

[49] Yi-Ling Chung et al. "CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2819–2829.

[50] Tuba Ciftci et al. "Hate speech on Facebook". In: *Proceedings of the 4th European Conference on Social Media, ECSM 2017*. 2017, pp. 425–433.

[51] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data". In: *Journal of communication* 64.2 (2014), pp. 317–332.

[52] James Cook. "Amazon scraps 'sexist AI' recruiting tool that showed bias against women". In: *The Telegraph* (2018). URL: `www.telegraph.co.uk/`

`technology/2018/10/10/amazon-scraps-sexist-ai-recruiting-tool-showed-bias-against/`.

[53] Susana Raquel Costa et al. “Playing Against Hate Speech”. In: *Journal of Digital Media & Interaction* 3.6 (2020), pp. 34–52.

[54] Adam M Croom. “How to do things with slurs: Studies in the way of derogatory words”. In: *Language & Communication* 33.3 (2013), pp. 177–204.

[55] Adam M Croom. “Slurs”. In: *Language Sciences* 33.3 (2011), pp. 343–358.

[56] Adam M Croom. “The semantics of slurs: A refutation of coreferentialism”. In: *Ampersand* 2 (2015), pp. 30–38.

[57] Lana Cuthbertson et al. In: *Proceedings AI for Social Good workshop at NeurIPS*. 2019.

[58] Maral Dadvar et al. “Improving cyberbullying detection with user context”. In: *Proceedings of the 35th European conference on Advances in Information Retrieval*. 2013, pp. 693–696.

[59] Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. “Developing a New Classifier for Automated Identification of Incivility in Social Media”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 95–101.

[60] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. “Racial Bias in Hate Speech and Abusive Language Detection Datasets”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019, pp. 25–35.

[61] Thomas Davidson et al. “Automated hate speech detection and the problem of offensive language”. In: *Eleventh international aaai conference on web and social media*. 2017.

[62] Bonnie-Elene Deal et al. ““I Definitely Did Not Report It When I Was Raped...# WeBelieveChristine# MeToo”: A Content Analysis of Disclosures of Sexual Assault on Twitter”. In: *Social Media+ Society* 6 (2020).

[63] Morteza Dehghani et al. “Purity homophily in social networks.” In: *Journal of Experimental Psychology: General* 145 (2016).

[64] Fabio Del Vigna et al. “Hate me, hate me not: Hate speech detection on facebook”. In: *Proceedings of the First Italian Conference on Cybersecurity (ITASEC)*. 2017, pp. 86–95.

[65] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Vol. 1. 2019, pp. 4171–4186.

[66] Karthik Dinakar, Roi Reichart, and Henry Lieberman. “Modeling the detection of textual cyberbullying”. In: *fifth international AAAI conference on weblogs and social media*. 2011.

[67] Lucas Dixon et al. “Measuring and mitigating unintended bias in text classification”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 67–73.

[68] Nicola Döring and M Rohangis Mohseni. “Fail videos and related video comments on YouTube: a case of sexualization of women and gendered hate speech?” In: *Communication Research Reports* 36.3 (2019), pp. 254–264.

[69] Nicola Döring and M Rohangis Mohseni. “Gendered hate speech in YouTube and YouNow comments: Results of two content analyses”. In: *SCM Studies in Communication and Media* 9.1 (2020), pp. 62–88.

[70] Maeve Duggan. *Online harassment*. Pew Research Center, 2014.

[71] Megan Duncan et al. “Staying silent and speaking out in online comment sections: The influence of spiral of silence and corrective action in reaction to news”. In: *Computers in Human Behavior* 102 (2020), pp. 192–205.

[72] Mai ElSherief et al. “Hate lingo: A target-based linguistic analysis of hate speech in social media”. In: *Twelfth International AAAI Conference on Web and Social Media*. 2018.

[73] Karmen Erjavec and Melita Poler Kovačič. ““You Don’t Understand, This is a New War!” Analysis of Hate Speech in News Web Sites’ Comments”. In: *Mass Communication and Society* 15.6 (2012), pp. 899–920.

[74] Christian Ezeibe. “Hate speech and election violence in Nigeria”. In: *Journal of Asian and African Studies* 56.4 (2021), pp. 919–935.

[75] Lizhou Fan, Huizi Yu, and Zhanyuan Yin. “Stigmatization in social media: Documenting and analyzing hate speech for COVID-19 on Twitter”. In: *Proceedings of the Association for Information Science and Technology* 57.1 (2020), e313.

[76] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. “Overview of the Task on Automatic Misogyny Identification at IberEval 2018.” In: *IberEval@ SEPLN*. 2018, pp. 214–228.

[77] *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018.

[78] Antigoni Founta et al. “Large scale crowdsourcing and characterization of twitter abusive behavior”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 2018.

[79] Antigoni Maria Founta et al. “A unified deep learning architecture for abuse detection”. In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 105–114.

[80] Jesse Fox and Wai Yen Tang. “Sexism in online video games: The role of conformity to masculine norms and social dominance orientation”. In: *Computers in Human Behavior* 33 (2014), pp. 314–320.

[81] Rolf Fredheim, Alfred Moore, and John Naughton. “Anonymity and online commenting: The broken windows effect and the end of drive-by commenting”. In: *Proceedings of the ACM web science conference*. 2015, pp. 1–8.

[82] Patxi Galán-Garcıéa et al. “Supervised machine learning for the detection of troll profiles in twitter social network: application to a real case of cyberbullying”. In: *Logic Journal of the IGPL* 24.1 (2016), pp. 42–53.

[83] Björn Gambäck and Utpal Kumar Sikdar. “Using convolutional neural networks to classify hate-speech”. In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 85–90.

[84] Lei Gao and Ruihong Huang. “Detecting Online Hate Speech Using Context Aware Models”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., 2017, pp. 260–266.

[85] Nikhil Garg et al. “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* 115.16 (2018), E3635–E3644.

[86] Jacqueline Garrick. “The humor of trauma survivors: Its application in a therapeutic milieu”. In: *Journal of aggression, maltreatment & trauma* 12.1-2 (2006), pp. 169–182.

[87] Ona de Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018, pp. 11–20.

[88] Debbie Ging. “Alphas, betas, and incels: Theorizing the masculinities of the manosphere”. In: *Men and Masculinities* 22.4 (2019), pp. 638–657.

[89] Jennifer Golbeck et al. “A large labeled corpus for online harassment research”. In: *Proceedings of the 2017 ACM on web science conference*. 2017, pp. 229–233.

[90] Viktor Golem, Mladen Karan, and Jan Šnajder. “Combining Shallow and Deep Learning for Aggressive Text Detection”. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 2018, pp. 188–198.

[91] Hila Gonen and Yoav Goldberg. “Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them”. 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)*. 2019, pp. 609–614.

[92] Philip B Gove. ““Noun Often Attributive” and “Adjective””. In: *American Speech* 39.3 (1964), pp. 163–175.

[93] Edel Greevy. “Automatic text categorisation of racist webpages”. PhD thesis. Dublin City University, 2004.

[94] John J Gumperz. “The speech community”. In: *Linguistic anthropology: A reader* 1 (2009), p. 66.

[95] Hatem Haddad, Hala Mulki, and Asma Oueslati. “T-HSAB: A Tunisian Hate Speech and Abusive Dataset”. In: *International Conference on Arabic Language Processing*. Springer. 2019, pp. 251–263.

[96] Yosh Halberstam and Brian Knight. “Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter”. In: *Journal of public economics* 143 (2016), pp. 73–88.

[97] Hugo L Hammer et al. “THREAT: A Large Annotated Corpus for Detection of Violent Threats”. In: *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*. IEEE. 2019, pp. 1–5.

[98] Amanda Haynes and Jennifer Schweppe. “STAD: Stop Transphobia and Discrimination Report”. In: *Transgender Equality Network Ireland* (2017).

[99] Eric Holgate et al. “Why Swear? Analyzing and Inferring the Intentions of Vulgar Expressions”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018, pp. 4405–4414.

[100] Christopher Hom. “Pejoratives”. In: *Philosophy compass* 5.2 (2010), pp. 164–185.

[101] Christopher Hom. “The semantics of racial epithets”. In: *The Journal of Philosophy* 105.8 (2008), pp. 416–440.

[102] Hossein Hosseini et al. “Deceiving google’s perspective api built for detecting toxic comments”. In: *arXiv preprint arXiv:1702.08138* (2017).

[103] Muhammad Okky Ibrohim and Indra Budi. “A dataset and preliminaries study for abusive language detection in Indonesian social media”. In: *Procedia Computer Science* 135 (2018), pp. 222–229.

[104] Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. “Brevity is the soul of Twitter: The constraint affordance and political discussion”. In: *Journal of Communication* 69.4 (2019), pp. 345–372.

[105] Andrej Janchevski and Sonja Gievska. “A Study of Different Models for Subreddit Recommendation Based on User-Community Interaction”. In: *International Conference on ICT Innovations*. Springer. 2019, pp. 96–108.

[106] Myungha Jang and James Allan. “Explaining Controversy on Social Media via Stance Summarization”. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM. 2018, pp. 1221–1224.

[107] Henry Jenkins. *Confronting the challenges of participatory culture: Media education for the 21st century.* Mit Press, 2009.

[108] Robin Jeshion. "Expressivism and the Offensiveness of Slurs". In: *Philosophical Perspectives* 27.1 (2013), pp. 231–259.

[109] Robin Jeshion. "Pride and Prejudiced: on the Reclamation of Slurs". In: *Grazer Philosophische Studien* 97.1 (2020), pp. 106–137.

[110] Robin Jeshion. "Slurs and stereotypes". In: *Analytic Philosophy* 54.3 (2013), pp. 314–329.

[111] Akshita Jha and Radhika Mamidi. "When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data". In: *Proceedings of the second workshop on NLP and computational social science.* 2017, pp. 7–16.

[112] Jialun'Aaron' Jiang et al. "Characterizing Community Guidelines on Social Media Platforms". In: *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing.* 2020, pp. 287–291.

[113] Evan Johnson. "But the Crowd was not Satisfied: Blackface Minstrelsy and Lynching as Fandoms of the Remediated Black Body". PhD thesis. University of Texas at Dallas, 2017.

[114] Srecko Joksimovic et al. "Automated Identification of Verbally Abusive Behaviors in Online Discussions". In: *Proceedings of the Third Workshop on Abusive Language Online.* 2019, pp. 36–45.

[115] David Kaplan. “The meaning of ouch and oops. explorations in the theory of meaning as use”. 1999.

[116] Mladen Karan and Jan Šnajder. “Cross-Domain Detection of Abusive Language Online”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 132–137.

[117] Mladen Karan and Jan Šnajder. “Preemptive toxic language detection in Wikipedia comments using thread-level context”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 129–134.

[118] Christos Karatsalos and Yannis Panagiotakis. “Attention-based method for categorizing different types of online harassment language”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 321–330.

[119] Aaron C Kay et al. “The insidious (and ironic) effects of positive stereotypes”. In: *Journal of Experimental Social Psychology* 49.2 (2013), pp. 287–291.

[120] Brendan Kennedy et al. “Contextualizing Hate Speech Classifiers with Post-hoc Explanation”. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2020.

[121] George Kennedy et al. “Technology solutions to combat online harassment”. In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 73–77.

[122] Randall Kennedy. *Nigger: The strange career of a troublesome word*. Vintage, 2008.

[123] Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. "A Large Self-Annotated Corpus for Sarcasm". In: *Proceedings of the Linguistic Resource and Evaluation Conference (LREC)*. 2018.

[124] Hannah Kia, Kinnon Ross MacKinnon, and Melissa Marie Legge. "In pursuit of change: Conceptualizing the social work response to LGBTQ microaggressions in health settings". In: *Social work in health care* 55.10 (2016), pp. 806–825.

[125] Olivier Klein, Russell Spears, and Stephen Reicher. "Social identity performance: Extending the strategic side of SIDE". In: *Personality and Social Psychology Review* 11.1 (2007), pp. 28–45.

[126] Anna Koufakou et al. "HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language". In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. Association for Computational Linguistics, 2020, pp. 34–43.

[127] Rohan Kshirsagar et al. "Predictive Embeddings for Hate Speech Detection on Twitter". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 26–32.

[128] Ritesh Kumar et al. "Aggression-annotated Corpus of Hindi-English Code-mixed Data". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[129] Srijan Kumar et al. "Community interaction and conflict on the web". In: *Proceedings of the 2018 world wide web conference*. 2018, pp. 933–943.

[130] Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. "Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage". In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 138–149.

[131] Irene Kwok and Yuzhou Wang. "Locate the Hate: Detecting Tweets against Blacks." In: *AAAI*. 2013.

[132] Wan Shun Eva Lam. "Language socialization in online communities". In: *Encyclopedia of language and education* 8.301 (2008), p. 11.

[133] Ronan Le Bras et al. "Adversarial filters of dataset biases". In: *International Conference on Machine Learning*. 2020, pp. 1078–1088.

[134] Omer Levy, Yoav Goldberg, and Ido Dagan. "Improving distributional similarity with lessons learned from word embeddings". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 211–225.

[135] Shuhua Liu and Thomas Forss. "Combining N-gram based Similarity Analysis with Sentiment Analysis in Web Content Classification". In: *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-Volume 1*. 2014, pp. 530–537.

[136] Yingchi Liu et al. "Sexual harassment story classification and key information identification". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 2385–2388.

[137] Yingchi Liu et al. "Uncover Sexual Harassment Patterns from Personal Stories by Joint Key Element Extraction and Categorization". 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). 2019, pp. 2328–2337.

[138] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2018.

[139] Rijul Magu, Kshitij Joshi, and Jiebo Luo. “Detecting the hate code on social media”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017.

[140] Rijul Magu and Jiebo Luo. “Determining code words in euphemistic hate speech using word embedding networks”. In: *Proceedings of the 2nd workshop on abusive language online (ALW2)*. 2018, pp. 93–100.

[141] Suman Kalyan Maity et al. “Opinion conflicts: An effective route to detect incivility in Twitter”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–27.

[142] Ilia Markov and Walter Daelemans. “Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate”. In: *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. 2021.

[143] Trevor Martin. “community2vec: Vector representations of online communities encode semantic relationships”. In: *Proceedings of the Second Workshop on NLP and Computational Social Science*. 2017, pp. 27–31.

[144] Adrienne Massanari. “# Gamergate and The Fappening: How Reddit’s algorithm, governance, and culture support toxic technocultures”. In: *New media & society* 19.3 (2017), pp. 329–346.

[145] Puneet Mathur et al. “Detecting offensive tweets in hindi-english code-switched language”. In: *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. 2018, pp. 18–26.

[146] maybesaydie. *A very belated introduction to the entire /t/fatlogic mod team.* 2015. URL: `www.reddit.com/r/fatlogic/comments/3mt2w4/a_very_belated_introduction_to_the_entire/`.

[147] Leland McInnes et al. “UMAP: Uniform Manifold Approximation and Projection”. In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[148] Peggy McIntosh. *White privilege: Unpacking the invisible knapsack.* 1988.

[149] Quinnehtukqut McLamore and Özden Melis Uluğ. “Social representations of sociopolitical groups on r/The_Donald and emergent conflict narratives: A qualitative content analysis”. In: *Analyses of Social Issues and Public Policy* (2020).

[150] Toby Mendel, M Herz, and P Molnar. “Does International Law Provide for Consistent Rules on Hate Speech?” In: *The content and context of hate speech: Rethinking regulation and responses* (2012), pp. 417–429.

[151] Johannes Skjeggestad Meyer and Björn Gambäck. “A platform agnostic dual-strand hate speech detector”. In: *ACL 2019 The Third Workshop on Abusive Language Online Proceedings of the Workshop*. Association for Computational Linguistics. 2019.

[152] Fernando Miró-Llinares, Asier Moneva, and Miriam Esteve. “Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments”. In: *Crime Science* 7.1 (2018), pp. 1–12.

[153] Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. “Neural Character-based Composition Models for Abuse Detection”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 1–10.

[154] Pushkar Mishra et al. “Author profiling for abuse detection”. In: *Proceedings of the 27th international conference on computational linguistics*. 2018, pp. 1088–1098.

[155] Sandip Modha, Prasenjit Majumder, and Thomas Mandl. “Filtering aggression from the multilingual social media feed”. In: *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. 2018, pp. 199–207.

[156] Shruthi Mohan et al. “The impact of toxic language on the health of reddit communities”. In: *Canadian Conference on Artificial Intelligence*. Springer. 2017, pp. 51–56.

[157] Mairead Eastin Moloney and Tony P Love. “Assessing online misogyny: Perspectives from sociology and feminist media studies”. In: *Sociology Compass* 12.5 (2018), e12577.

[158] Michael J Moore et al. “Anonymity and roles associated with aggressive posts in an online forum”. In: *Computers in Human Behavior* 28.3 (2012), pp. 861–867.

[159] Jessica Moreno, Ellen Pao, and Alexis Ohanian. *Removing harassing subreddits*. 2015. URL: `www.reddit.com/r/announcements/comments/39bpam/removing_harassing_subreddits/`.

[160] Hala Mulki et al. “L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language”. In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 111–118.

[161] Lisa Nakamura. “Don’t hate the player, hate the game: The racialization of labor in World of Warcraft”. In: *Critical Studies in Media Communication* 26.2 (2009), pp. 128–144.

[162] Hiroki Nakayama et al. *doccano: Text Annotation Tool for Human*. 2018. URL: `github.com/doccano/doccano`.

[163] Courtney Napoles, Aasish Pappu, and Joel Tetreault. “Automatically identifying good conversations online (yes, they do exist!)” In: *Eleventh International AAAI Conference on Web and Social Media*. 2017.

[164] Kanika Narang and Chris Brew. “Abusive Language Detection using Syntactic Dependency Graphs”. In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 44–53.

[165] Edward Newell et al. “User migration in online social networks: A case study on reddit during a period of community unrest”. In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[166] Dong Nguyen and Carolyn Rose. “Language use as a reflection of socialization in online communities”. In: *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. 2011, pp. 76–85.

[167] Chikashi Nobata et al. “Abusive language detection in online user content”. In: *Proceedings of the 25th international conference on world wide web*. 2016, pp. 145–153.

[168] Julie Norman and Drew Mikhael. "Youth radicalization is on the rise. Here's what we know about why." In: *The Washington Post* (2017). URL: `www.washingtonpost.com/news/monkey-cage/wp/2017/08/25/youth-radicalization-is-on-the-rise-heres-what-we-know-about-why/`.

[169] Tanya Notley. "Young people, online networks, and social inclusion". In: *Journal of Computer-Mediated Communication* 14.4 (2009), pp. 1208–1227.

[170] Mariam Nouh, Jason RC Nurse, and Michael Goldsmith. "Understanding the radical mind: Identifying signals to detect extremist content on twitter". In: *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. IEEE. 2019, pp. 98–103.

[171] Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. "Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online". In: *Sexuality & Culture* 25.2 (2021), pp. 700–732.

[172] Kadir Bulut Ozler et al. "Fine-tuning BERT for multi-domain and multi-label incivil language detection". In: *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 2020, pp. 28–33.

[173] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. "Do You Really Want to Hurt Me? Predicting Abusive Swearing in Social Media". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 6237–6246.

[174] Etienne Papegnies et al. "Graph-based features for automatic online abuse detection". In: *International conference on statistical language and speech processing*. Springer. 2017, pp. 70–81.

[175] Ji Ho Park and Pascale Fung. "One-step and Two-step Classification for Abusive Language Detection on Twitter". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 41–45.

[176] Ji Ho Park, Jamin Shin, and Pascale Fung. "Reducing Gender Bias in Abusive Language Detection". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 2799–2804.

[177] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. "Deep Learning for User Comment Moderation". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 25–35.

[178] John Pavlopoulos et al. "Toxicity Detection: Does Context Really Matter?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

[179] Sai Teja Peddinti, Keith W Ross, and Justin Cappos. "User Anonymity on Twitter". In: *IEEE Security & Privacy* 15.3 (2017), pp. 84–87.

[180] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[181] Jeffrey Pennington, Richard Socher, and Christopher D Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[182] Quang Anh Phan and Vanessa Tan. “Play with bad words: A content analysis of profanity in video games”. In: *Acta Ludica-International Journal of Game Studies* 1.1 (2017), pp. 7–30.

[183] Xuan-Hieu Phan and Cam-Tu Nguyen. “Jgibblda: A java implementation of latent dirichlet allocation (lda) using gibbs sampling for parameter estimation and inference”. In: (2006).

[184] Mihaela Popa-Wyatt. “Not All Slurs are Equal”. In: *Phenomenology and Mind* 11 (2016), pp. 150–157.

[185] Mihaela Popa-Wyatt and Jeremy L Wyatt. “Slurs, roles and power”. In: *Philosophical Studies* 175.11 (2018), pp. 2879–2906.

[186] Jing Qian et al. “A Benchmark Dataset for Learning to Intervene in Online Hate Speech”. 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).* Association for Computational Linguistics, 2019, pp. 4757–4766.

[187] Jing Qian et al. “Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Vol. 2. 2018, pp. 118–123.

[188] Rahat Ibn Rafiq et al. “Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network”. In: *Social Network Analysis and Mining* 6.1 (2016), p. 88.

[189] Stephen D Reicher, Russell Spears, and Tom Postmes. “A social identity model of deindividuation phenomena”. In: *European review of social psychology* 6.1 (1995), pp. 161–198.

[190] Mohammadreza Rezvan et al. “A quality type-aware annotated corpus and lexicon for harassment research”. In: *Proceedings of the 10th ACM Conference on Web Science*. 2018, pp. 33–36.

[191] Michael Ridenhour et al. “Detecting Online Hate Speech: Approaches Using Weak Supervision and Network Embedding Models”. In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer. 2020, pp. 202–212.

[192] Katherine Ritchie. “Social Identity, Indexicality, and the Appropriation of Slurs”. In: *Croatian Journal of Philosophy* 17.2 (50) (2017), pp. 155–180.

[193] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. “Prevalence and psychological effects of hateful speech in online college communities”. In: *Proceedings of the 10th ACM Conference on Web Science*. 2019, pp. 255–264.

[194] Haji Mohammad Saleem et al. “A web of hate: Tackling hateful speech in online social spaces”. In: 2016.

[195] Manuela Sanguinetti et al. “An italian twitter corpus of hate speech against immigrants”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[196] Maarten Sap et al. “The risk of racial bias in hate speech detection”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 1668–1678.

[197] Carla Schieb and Mike Preuss. “Governing hate speech by means of counterspeech on Facebook”. In: *66th International Communication Association Annual Conference*. 2016, pp. 1–23.

[198] Melani Schröter and Petra Storjohann. “Patterns of discourse semantics: A corpus-assisted study of financial crisis in British newspaper discourse in 2009”. In: *Pragmatics and Society* 6.1 (2015), pp. 43–66.

[199] Yaye Nabo Sène. “Hate speech exacerbating societal, racial tensions with ‘deadly consequences around the world’, say UN experts”. In: *UN News* (2019). URL: `news.un.org/en/story/2019/09/1047102`.

[200] Leandro Silva et al. “Analyzing the targets of hate in online social media”. In: *Tenth International AAAI Conference on Web and Social Media*. 2016.

[201] Wendel Silva et al. “A methodology for community detection in Twitter”. In: *Proceedings of the International Conference on Web Intelligence*. 2017, pp. 1006–1009.

[202] John Oliver Siy and Sapna Cheryan. “When compliments fail to flatter: American individualism and responses to positive stereotypes.” In: *Journal of personality and social psychology* 104.1 (2013), p. 87.

[203] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. “A characterization of political communities on reddit”. In: *Proceedings of the 30th ACM conference on hypertext and Social Media*. 2019, pp. 259–263.

[204] Devin Soni and Vivek K Singh. "See no evil, hear no evil: Audio-visual-textual cyberbullying detection". In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–26.

[205] Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. "Automatic identification of personal insults on social news sites". In: *Journal of the American Society for Information Science and Technology* 63.2 (2012), pp. 270–285.

[206] Wiktor Soral, Michał Bilewicz, and Mikołaj Winiewski. "Exposure to hate speech increases prejudice through desensitization". In: *Aggressive behavior* 44.2 (2018), pp. 136–146.

[207] Russell Spears and Martin Lea. "Panacea or panopticon? The hidden power in computer-mediated communication". In: *Communication Research* 21.4 (1994), pp. 427–459.

[208] Russell Spears and Martin Lea. *Social influence and the influence of the'social'in computer-mediated communication.* Harvester Wheatsheaf, 1992.

[209] Rachele Sprugnoli et al. "Creating a whatsapp dataset to study pre-teen cyberbullying". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 51–59.

[210] Kameron Johnston St Clare et al. "Linguistic Disarmament: On How Hate Speech Functions, the Way Hate Words Can Be Reclaimed, and Why We Must Pursue Their Reclamation". In: *Linguistic and Philosophical Investigations* 17 (2018), pp. 79–109.

[211] Leo G. Stewart and Emma S. Spiro. "Nobody Puts Redditor in a Binary: Digital Demography, Collective Identities, and Gender in a Subreddit Network".

In: *Proceedings of the 24th ACM Conference on Computer-Supported Cooperative Work and Social Computing*. Association for Computing Machinery, 2021.

[212] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. "Generating Counter Narratives against Online Hate Speech: Data and Strategies". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 1177–1190.

[213] I-Hsien Ting et al. "An approach for hate groups detection in facebook". In: *The 3rd International Workshop on Intelligent Data Analysis and Management*. Springer. 2013, pp. 101–106.

[214] Robert S Tokunaga. "Following you home from school: A critical review and synthesis of research on cyberbullying victimization". In: *Computers in human behavior* 26.3 (2010), pp. 277–287.

[215] Antonela Tommasel, Juan Manuel Rodriguez, and Daniela Godoy. "Textual aggression detection through deep learning". In: *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*. 2018, pp. 177–187.

[216] Thanh Tran et al. "HABERTOR: An Efficient and Effective Deep Hatespeech Detector". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 7486–7502.

[217] Robert Truswell. "Attributive adjectives and the nominals they modify". PhD thesis. University of Oxford, 2004.

[218] Stéphan Tulkens et al. "A Dictionary-based Approach to Racism Detection in Dutch Social Media". In: *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety*. LREC, 2016, p. 11.

[219] Brendesha M Tynes et al. "Online racial discrimination and psychological adjustment among adolescents". In: *Journal of adolescent health* 43.6 (2008), pp. 565–569.

[220] Elise Fehn Unsvåg and Björn Gambäck. "The Effects of User Features on Twitter Hate Speech Detection". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 75–85.

[221] Cynthia Van Hee et al. "Automatic detection of cyberbullying in social media text". In: *PloS one* 13 (2018).

[222] Bertie Vidgen et al. "Challenges and frontiers in abusive content detection". In: *Proceedings of the Third Workshop on Abusive Language Online*. 2019, pp. 80–93.

[223] Bertie Vidgen et al. "Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021.

[224] James Vincent. "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day". In: *The Verge* 24 (2016). URL: `www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist`.

[225] Isaac Waller and Ashton Anderson. "Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms". In: *The World Wide Web Conference*. 2019, pp. 1954–1964.

[226] Cindy Wang. "Interpreting Neural Network Hate Speech Classifiers". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 86–92.

[227] William Warner and Julia Hirschberg. "Detecting hate speech on the world wide web". In: *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics. 2012, pp. 19–26.

[228] Zeerak Waseem. "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter". In: *Proceedings of the first workshop on NLP and computational social science*. 2016, pp. 138–142.

[229] Zeerak Waseem and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". In: *Proceedings of the NAACL student research workshop*. 2016, pp. 88–93.

[230] Zeerak Waseem et al. "Understanding Abuse: A Typology of Abusive Language Detection Subtasks". In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 78–84.

[231] Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection". In: *IEEE access* 6 (2018), pp. 13825–13835.

[232] Daniel Whiting. “It’s not what you said, it’s the way you said it: slurs and conventional implicatures”. In: *Analytic Philosophy* 54.3 (2013), pp. 364–377.

[233] Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. “Detection of abusive language: the problem of biased datasets”. 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)*. 2019, pp. 602–608.

[234] Deirdre Wilson. “The pragmatics of verbal irony: Echo or pretence?” In: *Lingua* 116.10 (2006), pp. 1722–1743.

[235] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020, pp. 38–45.

[236] Lucas Wright et al. “Vectors for Counterspeech on Twitter”. In: *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, 2017, pp. 57–62.

[237] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. “Ex machina: Personal attacks seen at scale”. In: *Proceedings of the 26th International Conference on World Wide Web*. 2017, pp. 1391–1399.

[238] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. “Demoting Racial Bias in Hate Speech Detection”. In: *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, 2020, pp. 7–14.

[239] Guang Xiang et al. “Detecting offensive tweets via topical feature discovery over a large scale twitter corpus”. In: *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM. 2012, pp. 1980–1984.

[240] Dawei Yin et al. “Detection of harassment on web 2.0”. In: *Proceedings of the Content Analysis in the WEB* 2 (2009), pp. 1–7.

[241] Li-Yin Young. “The effect of moderator bots on abusive language use”. In: *Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*. 2018, pp. 133–137.

[242] Marcos Zampieri et al. “Predicting the Type and Target of Offensive Posts in Social Media”. 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)*. Association for Computational Linguistics, 2019, pp. 1415–1420.

[243] Marcos Zampieri et al. “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, 2020, pp. 1425–1447.

[244] Savvas Zannettou et al. “Measuring and characterizing hate speech on news websites”. In: *12th ACM Conference on Web Science*. 2020, pp. 125–134.

[245] Guanhua Zhang et al. “Demographics Should Not Be the Reason of Toxicity: Mitigating Discrimination in Text Classifications with Instance Weighting”.

In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4134–4145.

[246] Xiang Zhang et al. "Cyberbullying detection with a pronunciation based convolutional neural network". In: *2016 15th IEEE international conference on machine learning and applications (ICMLA)*. IEEE. 2016, pp. 740–745.

[247] Ziqi Zhang, David Robinson, and Jonathan Tepper. "Detecting hate speech on twitter using a convolution-gru based deep neural network". In: *European semantic web conference*. Springer. 2018, pp. 745–760.

[248] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. "Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 808–819.