The Influence of Commercial and Community Platforms on Product Reviews

Stefan Dimitrov

Master of Science

School of Computer Science

McGill University

Montreal,Quebec

2016-07-11

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

DEDICATION

I would like to dedicate this work to my family who supported me throughout the challenging times as a graduate student. The results of this research wouldn't be possible without their continuous encouragement.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Prof. Derek Ruths, and co-supervisor Prof. Andrew Piper, for their guidance. I have learnt a lot while working with them and the lessons they taught me opened the door to the vast and interesting world of digital humanities. Furthermore, I am grateful to Faiyaz Zamal and Edward Newell for sharing with me their expertise in research methodologies.

ABSTRACT

Online reviews form an integral part of product discovery and purchasing decisions today. The numerous platforms eliciting user generated content can be broadly categorized into commercial (whose primary revenue is derived from product sales) and community (often funded through advertising and dependent on metrics such as number of visitors). In this thesis we consider Amazon and Goodreads reviews discussing the same set of books in order to understand whether the implicit goals of these two platforms are reflected in the content users generate. We offer a comparative analysis of two datasets: all biography books with reviews on Goodreads and Amazon, and New York Times bestsellers listed between January 3, 2010 and January 3, 2015. Through statistical metrics, content analysis and rating comparisons we demonstrate that reviews for the same books differ significantly between the bookseller and the social network for readers. To quantify these differences we train a SVM classifier to separate ensembles of reviews for a given book between Goodreads and Amazon. Our classifier achieves over 90% F1 score indicating that, when taken together, reviews for the same book exhibit highly distinct characteristics on Goodreads and Amazon. We also look into the promotion mechanisms available on the two platforms: "like" and "helpful"/"not helpful" buttons. Through a controlled crowdsourcing experiment we show that these concepts are perceived differently by users. The promotion mechanisms may partially explain why reviewers write differently on Amazon than they do on Goodreads.

ABRÉGÉ

Les revues en ligne font partie intégrante de la découverte de produits et de la décision d'achat. On peut classifier les nombreuses plateformes générant du contenu par l'utilisateur comme commercial (dont le principal revenu provient de la vente de produits) ou communautaire (souvent financés par la publicité et dépendant de mesures telles que le nombre de visiteurs). Dans cette thèse, nous considérons des évaluations publiés sur Amazon et Goodreads qui discutent un même livre, afin de comprendre si le contenu généré par des utilisateurs reflète les objectifs implicites de ces deux plateformes. Nous proposons une analyse comparative de deux ensembles de données, l'un comprenant tous les livres biographiques avec des avis sur Goodreads et Amazon, et l'autre comprenant les meilleurs livres vendus entre le 3 Janvier 2010 et le 3 Janvier 2015 selon le New York Times. En utilisant des mesures statistiques, une analyse du contenu et une comparaison des évaluations entre les utilisateurs, nous démontrons que, pour un même livre, les évaluations diffèrent sensiblement entre le libraire et le réseau social auquel les lecteurs appartiennent. Pour quantifier ces différences, nous appliquons un algorithme de machine à vecteurs de support pour séparer des ensembles d'évaluations écrites pour un même livre sur Goodreads et Amazon. Notre classificateur réalise un score F1 de plus de 90% indiquant que, lorsqu'elles sont prises dans leur ensemble, les évaluations pour un même livre ont des caractéristiques très distinctes sur Goodreads et Amazon. Nous examinons également les mécanismes de promotion disponibles sur les deux plateformes tels que les boutons "j'aime" et "je trouve utile"/ "je ne trouve pas utile".

A l'aide d'une expérience contrôlée de crowdsourcing, nous montrons que les utilisateurs perçoivent ces concepts différemment. Les mécanismes de promotion peuvent expliquer en partie pourquoi les utilisateurs écrivent différemment sur Amazon et sur Goodreads.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

ABSTRACT iv

ABRÉGÉ v

LIST OF TABLES ix

LIST OF FIGURES x

1 Introduction 1

2 Related Works 10

2.1 Effects of Reviews on the Purchasing Decision 10
2.2 Moderation, Priming and Conformity 14
2.3 Helpful and Unhelpful Reviews 17
2.4 Review Ratings 19

3 Platforms and Dataset 21

3.1 Design Differences 21
3.2 Dataset 27
3.2.1 Biography Books 27
3.2.2 New York Times Bestsellers 29

4 Results and Discussion 32

4.1 Biography Results 32
4.1.1 Statistical Analysis 32
4.1.2 Book Ratings 34
4.1.3 Sentiment Analysis 36
4.1.4 Content Analysis 38

4.2 Bestsellers Results 39
4.2.1 Buying Experience 40
4.2.2 Linguistic analysis 41
4.2.3 Star Rating Analysis 43
4.2.4 Review Classification 46
4.2.5 Analysis of Review Promotion 52

5 Conclusion 65

References 68

LIST OF TABLES

Table page

3–1 Metadata we collected for each review. 31

4–1 Statistical comparison of Goodreads and Amazon biography book reviews. Statistics annotated with a * have statistically significant differences ($p < 0.00001$). Positive difference indicates higher value on Goodreads. 33

4–2 Frequency of linguistic features (expressed as % of words belonging to each dictionary) in reviews, by platform. All statistics shown are significant with $p < 0.001$. Relative difference is calculated with respect to Amazon, where positive numbers indicate higher frequency on Amazon and negative indicate lower frequency on Amazon. 42

4–3 Features used for the classification of review ensembles, by level of granularity. 48

4–4 Description of each feature used for classification 49

4–5 F1 scores for classification of review ensembles when using each of the feature sets individually. 52

4–6 Percentage of reasons given for applying each label to reviews. Each column sums to 100%. 60

4–7 Overlap between categories expressed as percentage of reviews from each category as marked by all workers. 60

LIST OF FIGURES

Figure		page
3–1	Amazon shows the book cover with a description and buying options.	22
3–2	Goodreads shows the book cover with a description and links to libraries / bookstores.	22
3–3	Amazon offers professional reviews to set the stage for user contributed content.	24
3–5	Amazon offers "customer reviews" featuring a title, a link to comments associated with the review, the choice of marking the review as either "helpful" or "not helpful" and the option to report the review as abuse of the reviewing guidelines.	24
3–6	Goodreads provides "community reviews" with the avatar (profile picture) prominently featured to the left of the review and a convenient option to "like" a review, but not to "dislike" it.	25
3–4	Amazon provides a convenient basket management panel to the side of the book description for easy purchase. Multiple purchase options are offered.	27
3–7	The distribution of books across New York Times Bestsellers categories between January 3, 2010 and January 3, 2015.	29
3–8	The distribution of the 3,381 New York Times Bestseller books we were able to match between Goodreads and Amazon by category.	30
4–1	The distribution of stars over all ratings for biography books	34
4–2	The distribution of biography books by average rating	34
4–3	The distribution of stars over a sample of ratings for NYT bestsellers	44
4–4	The distribution of New York Times bestsellers by average rating	44

4–5 Distribution of Amazon reviews into "helpful", "not helpful" and "not ranked" categories, and of Goodreads reviews into "liked" and "not liked" categories. The numbers add up to 200% because they include reviews from both platforms. 53

4–6 The Crowdflower task coding interface as seen by a worker. The justifications are visible because the review was marked as "helpful". 59

4–7 Agreement between Crowdflower workers and Amazon/Goodreads visitors expressed as % of reviews from each category unanimously labeled by all workers. 62

CHAPTER 1
Introduction

In the age of social networks and social media, user generated content is paramount to the success of online platforms. It can take the form of comments, reviews or discussion posts. In this study we focus on one of these forms – product reviews. Reviews are becoming increasingly popular on web shopping platforms and interest communities alike. The reason for their popularity can be attributed in part to the usefulness of reviews to all parties involved.

For platforms with overt commercial intent, reviews are a valuable asset in promoting products and enhancing the shopping experience. The ultimate goal of a web store is to increase the volume of sales, and the number and quality of online reviews have previously been shown to affect the purchasing decision [27] [8]. Hence, online stores benefit the most from the type of reviews which inform potential customers and contribute to making the right buying decision. Reviews on such platforms may also discuss aspects of the transaction itself: the shipping cost and speed, packaging and physical condition of the product.

On the other hand, interest-based communities – social platforms centered around visitors with similar interests towards a product or a type of products – greatly benefit from user reviews as a valuable tool in attracting visitors. However, for these platforms revenue is derived from advertising and is often correlated with the number of visitors and returning users. Thus, interest-based communities

benefit the most from the type of reviews liked by the majority of members and from controversial reviews which prompt visitors to participate in a discussion or to gain deeper understanding of the interest space. Product features and the user's experience with the product may be common topics of interest to the community. Reviewers may pose questions about the product, respond to previous reviews or compare the product to other related items.

Online shoppers leverage the experience shared in reviews by other buyers to save time and money for trying the product. Their incentive for writing a review is to help a good product reach more buyers, or to inform others about a potentially disappointing purchase. This is particularly true for books, which have few extrinsic features to facilitate a buying decision (for example: author, genre and theme) while their value is derived from the experience of the reader.

For book enthusiasts, reviews are a medium which enables the exchange of ideas. For instance, users of community oriented platforms might derive satisfaction from sharing their views and understanding, receiving feedback, and using reviews as a way to socialize with others who share similar interests by participating in a discussion. Similarly, reviewers on a commercial platform may be motivated by altruistic desire to help others in choosing a good book.

Last but not least, book authors and publishers benefit from reviews in two ways: as a feedback channel from consumers and as a marketing instrument.

By allowing users to freely express their opinions about a book, both the online bookseller and the community platform foster user engagement. Understandably, the two types of platforms are interested in attracting more reviews aligned with

their aims. In our comparative study we discover differences in the design, features, choice of wording and moderation mechanisms. Thus, understanding how context shapes user behaviour will inform future platform design.

It is to be expected then, that online reviews have attracted significant interest in the academic community, with researchers focusing on understanding the economic impact [11, 5] and helpfulness of reviews to book buyers [24], the social aspect of user generated content online [12, 22] and the effect of moderation [6]. Some works study the relationship of ratings (numeric summary of a user's opinion about a product) with reviews [15, 21], which is particularly interesting in the context of sentiment analysis of reviews [3].

While the extent of existing literature helps us understand reviews as a form of subjective expression and how they affect other processes, we found few works on what impacts reviews, or more precisely: do platforms with different explicit goals solicit substantially dissimilar reviews about the same product, and if they do, how can we explain this phenomenon?

Some works consider the impact of external factors, such as the weather [4] and aspects of the platform [33] but we weren't able to find a truly comparative study between two platforms. Thus, we focus on comparing book reviews on Amazon.com and Goodreads. Amazon is an e-commerce website which started as an on-line book seller in 1995 [31], and as of 2015 has an active customer base of over 244 million people globally[1] . It has presence in multiple markets around the world but for the

[1] http://www.amazon.com/b?ie=UTF8&node=8445211011

purpose of this study we restrict our dataset to the English-language version. Similarly, Goodreads, which launched in 2007 and claims to have 40 million members, offers English-language book reviews, but does not sell books. Instead, it enhances the reading experience by connecting readers in virtual communities around books they like and allowing them to catalogue books they have read, or would like to read, into shelves organized by user-assigned themes. We consider books to be an excellent product to center our study around, firstly because they are perceived subjectively, secondly because they are hard to describe in terms of features and, last but not least, because the book market remains one of the largest in the cultural sphere, with sales in the USA valued at over 27 billion dollars [2] compared to only 4.5 billion dollars for music [16]. Both Amazon and Goodreads allow users to express their opinions about a given book: in the "customer reviews" section on Amazon or as part of the "community reviews" on Goodreads, accompanied by a numeric rating between 1 and 5 stars. Visitors can also react to others' reviews: by marking them as helpful or unhelpful on Amazon, or by liking them on Goodreads, or by writing comments. Goodreads highlights the identity of the reviewer by featuring their chosen profile picture (avatar) beside their review or comment. Based on the data we collected we see that Amazon moderates reviews while Goodreads gives reviewers the freedom to express their opinion even if it involves profanity. While the two platforms are similar in their reviewing tools, one aspect remains substantially different: the goal of the platform. Thus, in this thesis we study the impact of a platform's purpose on the reviews users generate and the ratings they give, to the same book. It is worth mentioning that Amazon acquired Goodreads in 2013 [10] but at the time of writing

the two platforms remain largely independent. We approach this problem from two sides. First we study a large dataset of reviews for a large number of books all taken from the same genre: biography. In the second part of our work we opt for a sample of reviews from a broader selection of books belonging to a variety of genres, and listed as New York Times best sellers. The reason for taking two approaches was primarily to analyse the population behaviour overall for a given genre on each of the platforms, and then to confirm that our findings hold across genres. We chose the biography genre for the first part of our study because it is relatively well-defined, particularly in light of the user-sourced list of "shelves" on Goodreads which often cross the traditional genre boundaries. As part of our first study we present a statistical comparison between Goodreads and Amazon in terms of review length, number of reviews per book/user, vocabulary, and rating distributions. We also apply sentiment analysis in an attempt to understand the remarkably different rating distributions between the two platforms. As a result, with our first study we outline some of the major characteristics of Goodreads - highly engaged user base, more critical ratings, and Amazon - vocabulary intended to inform the buying decision, U-shaped rating distribution not necessarily reflected in the review sentiment. For our second study we gather a dataset consisting of 195,195 Amazon and 189,329 Goodreads reviews for books which were listed as New York Times bestsellers between January 2010 and January 2015. Based on our findings for the biography genre, we expect to find differences in vocabulary, star ratings and the use of review promotion tools provided by the platforms. We test whether our discoveries with respect to review content

and star ratings for biography books hold for a larger variety of genres. To quantify the differences in content we apply Linguistic Inquiry and Word Count analysis [29], which shows the prevalence of words from specific categories and the use of punctuation in reviews from each of the two platforms. If the platform design and purpose have an impact on reviewers then this should be reflected in the content of reviews, because text is written for a purpose. We identified some striking differences between the two platforms: significantly higher frequency of the pronoun "you" on Amazon, as well as, words from the certainty and positive emotions dictionaries; in contrast, Goodreads reviews more often contain words from the negative emotions, tentative and swear dictionaries. While the discrepancies between the two platforms are statistically significant when considering all reviews, are they noticeable at the individual review level? To answer this question we train an SVM classifier on individual reviews from each of the two platforms. We find that the classifier performs rather poorly (less than 60% F1 score) suggesting that individual reviews from the two platforms may not be clearly separable. Nonetheless, both platforms show 10-30 reviews on every book page, exposing visitors to an ensemble of reviews. This observation motivated us to repeat the SVM classification experiment but considering a sample of reviews for the same book from each platform as an instance in the dataset. The setup resulted a remarkable performance improvement: 95% F_1 for fiction and 92% F_1 for non-fiction books. Therefore, our study offers quantitative evidence of the content differences between reviews on a commercial and a community-oriented platform for the same product.

We also perform a crowdsourcing experiment to analyze how platform-specific tools such as "like" and "helpful" buttons are perceived by users with respect to reviews. By default Amazon sorts reviews by the number of "helpful" votes, making reviews considered "helpful" bubble up to the top and reviews marked as "not helpful" buried on the last pages. Similarly, Goodreads uses a proprietary sorting algorithm which includes the number of "likes" as a factor. The very existence of a self-moderation tool (i.e. "not helpful" button) on Amazon is a major difference in design from Goodreads where users can "like" but can't "dislike" reviews. With our experiment we aim to understand whether users are driven by different reasons when they choose to mark a review as "helpful" (a more utilitarian concept) as opposed to when they decide to "like" it (indicating personal preference). The perception of these concepts coupled with the sorting algorithms in the two platforms can lead to different kind of review content being showcased. Our results support the claim that the two concepts ("like" and "helpful") are used differently, as evidenced by a test of marginal homogeneity. Are reviewers aware of what makes a review receive a large number of "likes" vs. a large number of "helpful" votes, and are they influenced by how their review will be perceived on a given platform, is a question we leave for a future study.

Finally, we compare the rating distributions on Goodreads and Amazon for the New York Times bestsellers dataset. We show that our original findings for biography books: more extreme, U-shaped, rating distribution on Amazon with 5-star ratings most common, in contrast to more moderate ratings on Goodreads with 3 and 4-star ratings prevalent, hold true for a larger variety of genres as well. These results

are in agreement with the content analysis, namely tentative and negative emotions vocabulary can be associated with lower ratings while positive emotions and words expressing certainty are accompanied by higher rating as one may expect.

Overall, both our studies demonstrate that the difference in design, choice of language and explicit purpose of two platforms can result substantially different user generated content, discussing exactly the same book. This is a fundamental finding which forms the basis for future studies into user behavior online. It would be very interesting to identify whether the differences can be attributed to population selection or if the same users behave differently on each of the two platform, but such study would require internal data which is not accessible to independent researchers.

The rest of this thesis is organized as follows: next chapter is dedicated to related work, we offer a profound analysis of previous studies on product reviews and user generated online content. We then describe our data collection methods and characterize the datasets on which our two studies are based. Following, we provide a statistical comparison between reviews on the two platforms in terms of ratio of reviews per book and user, length, vocabulary and punctuation, both for biography books and New York Times Bestsellers. We then describe the SVM classification experiment for bestseller reviews and show our results. Next, we compare the rating distributions on the two platforms for each of the two datasets. Finally, we discuss the "like" vs. "helpful" crowdsourcing experiment and offer results from statistical tests which show how these two concepts differ. We conclude with an in-depth discussion of our overall findings both for biography book reviews and for New York Times Bestsellers reviews; we show how our study can form the basis for future works

on the growing importance of attracting the desired type of user generated content online.

CHAPTER 2
Related Works

A number of previous works have studied user generated content online such as commentary and product reviews. In this chapter we present some of the most relevant works upon which we built our experiments.

2.1 Effects of Reviews on the Purchasing Decision

One of the most utilitarian benefits of online reviews is that they inform the purchasing decision. Not surprisingly then, a large volume of academic literature is motivated primarily by the effects of reviews on the reader. These studies help us understand what kind of reviews would be most useful in promoting the goals of a commercial platform such as Amazon.

Forman et al. [11] collect a total of 175,714 reviews from Amazon for 786 books listed as Amazon "purchase circle" bestsellers in at least one US city between April 2005 and January 2006. Through regression models they correlate review rating (valence), reviewer identity disclosure (in the form of Amazon verified real name) and location disclosure, with monthly book sales data nationally and per city, as well as with helpful votes by other reviewers. They find significant (at the 1% level) positive association between identity disclosure and both sales and number of helpful votes. The effect of this relationship is amplified when limited to the state where the reviewer resides. Conversely, upon examining the relationship between average review rating and change in sales they don't find a significant correlation

[11]. The method applied in this work is important because it attempts to correct for confounding factors between review rating and sales: i.e. a high quality book may be rated highly because of its quality and may also sell well for the same reason. However, Forman et al. consider the change in sales as a function of the change in average rating, thus controlling for book quality. Still, as the authors disclose, their results may be affected by external events (such as promotions) which may influence change in sales. Thus, based on this study a commercial platform can benefit from segregating reviews by target market. Amazon already does this by having international websites for United Kingdom, Germany, France, Japan, Canada, China, Italy, Spain, and Brazil [31] as opposed to Goodreads which has a globally unique website.

Park et al. [27] conduct an experiment to measure the effect of review quality (with high quality reviews defined as those providing factual and objective information about the product, also referred to as "informant" reviews in the study, while low quality reviews were those expressing primarily emotions, subjective opinion or interjections, referred to as "recommenders") and review quantity on two groups including a total of 352 college students (high involvement - instructed to make a purchasing decision for a business, and low involvement - asked to treat the experiment as a browsing session). The authors find that both review quantity (indicating popularity of the product) and review quality impact the purchasing decision, but the extent to which they influence the subjects depends on the level of involvement. Namely, both low and high involvement subjects were affected by review quantity, however for the high involvement group the effect was significantly amplified by the

quality of reviews [27]. This study is relevant to Amazon and Goodreads, because one of the platforms is a bookseller (where people may go with the intent to purchase a book) where the other is a community (which may attract a greater number of low involvement users who are not explicitly looking to place an order). Thus, one of the platforms may have an incentive to focus on review quality while the other may choose to prioritize quantity.

In a study on the effect of reviews and comments on movie sales, as well as the feedback loop of movie screenings (i.e. popularity) on online user generated content about the movie, Duan et al. [8] find that there exists an interdependence relationship where the volume of reviews is positively correlated with box office sales for the movie and vice-versa. The authors show the dual role of reviews as a precursor and outcome to sales by comparing the traditional Ordinary Least Squares (OLS) model and a Three-stage Least Squares (3SLS) model while into account extrinsic factors such as cast, genre and critical reception of the movies [8]. Movies are similar to books in that both are experience goods, difficult to evaluate based on objective criteria.

To gain a better understanding of the relationship between book reviews and sales we consider a study by Judith A. Chevalier and Dina Mayzlin which compares book reviews and book sales on two online bookstores: Amazon and Barnes & Noble [5]. Similar to this thesis, the authors choose books because they enable direct comparison of online viva voce discussing exactly the same product. By controlling for book and considering the difference in sales and reviews on each of the two platforms, Chevalier and Mayzlin's model excludes external factors, such as offline

promotions by the publisher, which could contribute to a simultaneous change in both reviews and sales, but there is no reason why they would affect one platform and not the other. However, this study makes the assumption that a potential buyer reads reviews only on the platform where they make a purchase. For instance, if a bad review is posted on Amazon the study assumes that only people deciding to buy the book on Amazon will read it, and not those who may be searching for reviews about the same book but prefer to order from Barnes & Noble. This limitation is understandable due to the lack of publicly available data about individual users on each of the platforms. The first interesting finding presented in this study is that Amazon reviews are on average longer than reviews on Barnes & Noble, with 2-star, 3-star and 4-star reviews longer on both platforms than the extreme 1-star and 5-star. The authors also show that positive reviews prevail on Amazon, as well as, on Barnes & Noble. Lastly, both the number of reviews and their star rating have a causal relationship with sales rank [5].

The effect of online word of mouth on sales has been shown to be significant not only for movies and books as discussed above, but also for much more expensive subjective experiences such as travel and hotel stays. Ye et al. study the impact of review valence and variance of ratings on room bookings for 1639 hotels in China. They find that higher average rating is positively correlated with room bookings, but rating variance does not result fewer bookings [35]. Similarly, Beverley A. Sparks and Victoria Browning conduct an experiment with 554 participants asked to read reviews on simulated hotel booking websites in order to evaluate the response of users to review valence (positive, negative or neutral), framing (ordering of reviews

with respect to valence) and rating (1.5, 3 or 4.5 in agreement with the valence of the review). Their findings confirm the expected result that both review valence and framing are positively correlated with booking likelihood (i.e. subjects were more likely to book a hotel after reading positive reviews or positively framed reviews than otherwise). Furthermore, the authors show that negative framing and negative valence both have greater effect amplitude than their positive counterparts. Similarly, rating was found to affect the booking decision in unison with framing. A very important result of this study is that positive framing increases the chance of making a booking even when the reviews are negative overall [32]. The relevance of these studies is twofold: on one hand the authors show that the impact of reviews on sales is applicable to a wide range of experience goods from books to hotels, on the other hand the findings with respect to framing can be used to further an online platform's goal through design and review ordering.

2.2 Moderation, Priming and Conformity

Understanding the factors that stimulate users to produce quality contributions may be important for platforms such as Amazon, based on previous findings [27] which identify the role of quality in facilitating the purchasing decision. We offer a review of related works concerned with the behaviour of users when generating content online. Our foremost interest is in understanding the means by which a platform can influence the content users generate.

Nicholas Diakopoulos and Mor Naaman study priming and moderation with respect to news comments by collecting a dataset of 54,540 comments and conducting a surveys with 390 visitors. They look at two moderation strategies: pre- and

post-moderation, where pre-moderation requires the approval of a moderator prior to publishing and post-moderation relies on crowdsourced feedback such as flagging or downvoting. Both approaches are found to be imperfect with pre-moderation relying on the journalistic competence of moderators to make unbiased judgement while flagging is subject to abuse by users for the purpose of silencing certain viewpoints [6]. Based on their reviewing guidelines we can say that Amazon applies a combination of pre- and post- moderation (a review may be "rejected or removed" [1]) while Goodreads relies entirely on the community to promote or demote reviews. The results of Diakopoulos and Naaman's study suggest that readers perceive comment quality differently, depending on their personal motivations. Namely, users motivated by individual-centric reasons (to gain more information, to validate personal opinion or to seek entertainment) are more likely to find comments offensive than individuals driven by social interaction motives [6]. If we relate these user types to the explicit purpose of Amazon and Goodreads we can see why one platform may be more interested in rejecting profanity than the other.

Sukumaran et. al conduct two experiments to understand the stimuli for thoughtfulness of user generated content online. First they look at the influence of existing content on new user contributions and then they show that the same effect can be achieved by modifying the design of the platform itself [33]. The finding of Sukumaran et. al that users can be induced to follow a pre-existing standard based on the content they see is highly applicable to our study: Amazon reviews are sorted by helpfulness (a measure of adhering to the community standard) and above user content we see professional editorial reviews. In contrast, Goodreads shows only

community reviews sorted by a complex proprietary algorithm. Other works have also studied the effect of existing content on future reviews. For instance, Loizos Michael and Jahna Otterbacher look at 1,023,753 reviews on TripAdvisor in order to understand how the frequency of 12 stylistic features (use of first person, all capital letters, punctuation, emoticons, etc.) varies depending on the context a review appears in. They find that the increase in probability of a review experiencing similar frequency as its preceding reviews, also known as herding behaviour, is statistically significant for all features [22]. This result is important for our study of Goodreads and Amazon because the two platforms are very similar to TripAdvisor in allowing users to review the same product, read previous reviews, but not communicate privately with each other. Eric Gilbert and Karrie Karahalios label this phenomenon of reviews and reviewers resembling existing ones "deja reviews" and "deja reviewers". They study a dataset of 98,191 Amazon reviews and find that 10 – 15% of all reviews can be classified as "deja reviews". The authors interviewed 20 reviewers who had written reviews most similar to existing ones based on cosine similarity. They identified two clearly separable clusters of reviewers: amateurs and pros. While both groups contribute "deja reviews" they are motivated by very different reasons: for inexperienced reviewers the primary goal is to express their spontaneous reaction to the product, without necessarily being aware of prior reviews. In contrast, professional reviewers intentionally avoid reading existing reviews in an attempt to build their own unique "brand" and follow their personal agenda, uninfluenced by previous contributions [12]. Gilbert and Karahalios conclude their study with a suggestion for

platforms to apply a form of "social navigation" to alleviate repetitiveness by using a measure of review helpfulness.

2.3 Helpful and Unhelpful Reviews

What makes a review helpful? To answer this question we must first define review helpfulness. It is often understood as a measure of the influence a review has on the purchasing decision. Based on this definition Susan M. Mudambi and David Schuff look at 1,587 Amazon reviews to understand the impact of review extremity (star rating), review depth (word count) and product type (search or experience goods) on the percentage of helpful votes a review receives. Experience goods are those which depend primarily on subjective attributes, for example books, movies or music. The authors find that for reviews of experience goods (a song and a video game) extreme ratings are associated with lower percentage of helpful votes, while review length has a positive effect on helpfulness [24]. Lionel Martin and Pearl Pu train Support Vector Machine (SVM), Random Forest and Naive Bayes models to classify reviews as helpfulness or not helpful based on their helpful ratings. The datasets consist of reviews from three platforms: Amazon, TripAdvisor and Yelp, where only Amazon allows for negative (i.e. "unhelpful") votes. The authors aim to show that the number of words associated with emotions are a good predictor for review helpfulness. The results indicate that while the count of such words improves the accuracy of the classifiers by up to 9%, the truly important feature for helpful reviews classification is a vector of Part Of Speech (POS) tag counts (0.86, 0.89 and 0.97 F1 scores accordingly for the Amazon, TripAdvisor and Yelp datasets) [20]. We

build upon some of the feature selection techniques presented in this paper for our Amazon vs. Goodreads reviews classifier.

Jahna Otterbacher investigates the review helpfulness correlation with 17 quality metrics loaded into 5 factors: relevancy, reputation (of reviewer), representation (ease of understanding metrics), believability (difference from existing reviews), and objectivity (similarity to product description). She analyses a dataset of 68,393 Amazon reviews and finds that less surprising reviews (higher believability value) receive more helpful votes, followed by relevancy, reputation, representation and objectivity in decreasing order of positive correlation with helpfulness. Her results also show that newer reviews are more likely to receive helpful votes than older ones, which can be explained with review ordering mechanisms on the platform [26]. This study is important because it allows us to gain an understanding of what Amazon visitors might mean when they mark a review as "helpful".

Jingjing Liu et al. look at helpful votes for Amazon reviews. Surprisingly they find that earlier reviews receive more helpful votes than later ones ("early bird" bias), which is contrary to Otterbacher's result presented above, but can be explained with different review ordering on the platform. Similarly Liu et al. show that reviews which already have helpful votes attract a disproportionately large number of new helpful votes, a phenomenon the authors define as "winner circle" bias. Finally, based on their analysis of 23,141 reviews Liu et al. find that Amazon users are much more likely to rate a review as "helpful" than "not helpful" with half of all reviews receiving more than 90% "helpful" votes. The propensity of users to express their

positive opinion of a review but not their disagreement is defined as "imbalance vote" bias [18].

Goodreads also offers a user-sourced promotion mechanism for reviews exposed through a "like" instead of "helpful" button. Vasconcelos et al. study the popularity of tips (a type of micro-reviews) on Foursquare by means of the "likes" they receive [9]. Foursquare tips share similarities with some of the Goodreads reviews we looked at: subjective and informal language, shorter length; and, the two platforms offer the same "liking" feature. The dataset collected by Vasconcelos et al. consists of more than 10 million tips, written by 13,5 million users and awarded a total of 9 million likes. The authors evaluate the extent to which the rich-get-richer effect (defined similarly as the "winner circle" bias on Amazon studied by Liu et al. [18]) can be observed for likes given to Foursquare tips. They find a weak correlation between the current number of likes a tip has and its future popularity, suggesting that there are other features which may help predict the number of likes a tip will get such as the author, the venue for which it's written and the content [9]. We expect similar forces to drive likes on Goodreads where the review author and the book are prominently featured on the same page.

2.4 Review Ratings

Ratings accompany reviews on both Goodreads and Amazon. They provide a succinct summary of the reviewer's sentiment towards a book in the form of 1 to 5 stars. Nan Hu et al. [15] analyse a dataset of Amazon reviews for 4,000 books in order to better understand the impact of ratings and review sentiment on sales. Their dataset consists of up to five most recent reviews and up to five most helpful

reviews for each book. The results show that there is no significant direct impact of review ratings on book sales, but there is a significant ($p < .01$) indirect relationship through average review sentiment. The authors calculate review sentiment by using a dictionary of words with preassigned sentiment value and converting the average review sentiment to a scale of 1 to 5, which can be compared to the rating scale. Review content sentiment has stronger impact on sales than title sentiment, as does moderate sentiment and not extremely negative or extremely positive sentiment. The authors suggest that review rating is used sequentially with sentiment when making a buying decisions: users may filter reviews by rating first, to reduce the volume of content they need to read. This explanation is also supported by an experimental study with 156 students which finds that ratings are important primarily during the search and awareness phase of the buying process. [15] These results are important with respect to Goodreads and Amazon which experience significant discrepancy in review rating, thus rendering one platform potentially more helpful than the other in informing the purchasing decision.

Julian McAuley and Jure Leskovec study the relationship between ratings and reviews by performing LDA topic modelling on the review text. In doing so they uncover hidden dimensions in the rating with corresponding hidden topics in the review content, which helps relate a rating given by a reviewer to the product category or subcategory aligned with their preference. [21]

CHAPTER 3
Platforms and Dataset

3.1 Design Differences

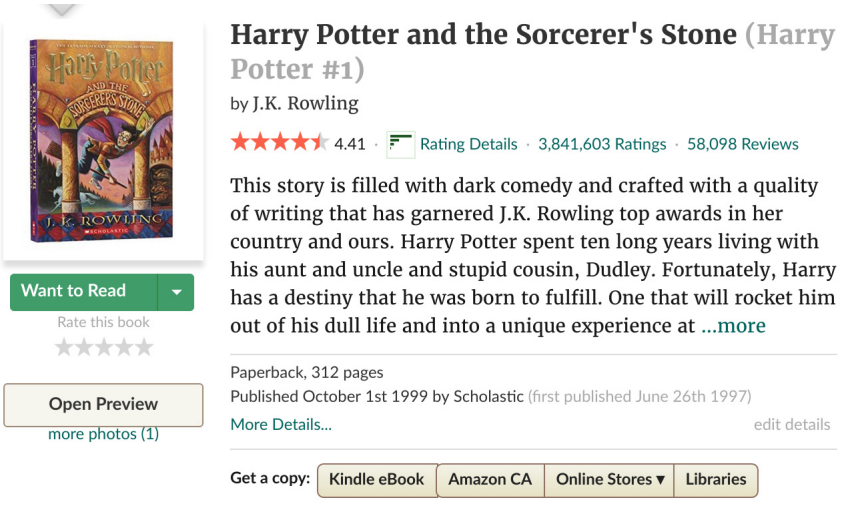
Amazon is a web store which sells books among a large variety of items. In contrast, Goodreads is a community-oriented platform centered around books with the mission to "help people find and share books that they love" [13]. To achieve its mission Goodreads offers a number of features around organizing book collections and connecting with other readers. Both platforms provide book pages with user contributed reviews. On Amazon these reviews may be anonymous (with the user identified as "Amazon Customer"), but not on Goodreads. Before diving into analysing the datasets we collected, we offer an overview of the similarities and differences in design between the two platforms. Our hypothesis is that the choice of design and wording is intentional, it reflects the explicit purpose of each of the platforms and influences user generated content as demonstrated by Sukumaran et. al [33].

At first glance the book pages on both Amazon (Fig. 3–1) and Goodreads (Fig. 3–2) are very similar: they offer a picture of the book cover and a short description. However, even here we see Amazon emphasizing the commercial aspect of its service with multiple pricing options prominently displayed below the description.

As we explore the pages further the differences become more explicit. For instance, to the right of the book description Amazon offers a "shopping basket" ready

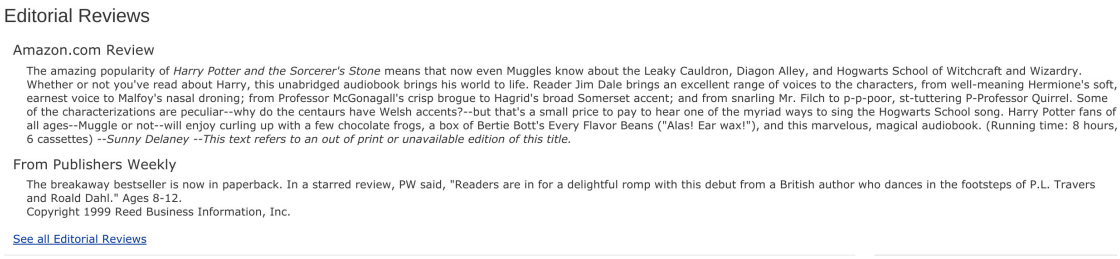
Figure 3–1: Amazon shows the book cover with a description and buying options.

Figure 3–2: Goodreads shows the book cover with a description and links to libraries / bookstores.



for the user to purchase the book with one click (Fig. 3–4) with multiple shipping options likely to meet any shopper's demands. Goodreads offers no such facility. Further on the Amazon page we find professional editorial reviews (Fig. 3–3) written at a high standard and with clear purpose in mind: to help a buyer decide whether the books is the right choice for their reading needs. Again, Goodreads offers no such content, instead the platform showcases community reviews (Fig. 3–6). On the other hand, Amazon features "customer reviews" (Fig. 3–5). While the format of reviews is very similar: rating from 1 to 5 stars, author and review date, followed by review content of arbitrary length, there are some substantial differences. Firstly, Goodreads features a reviewer avatar (or picture) allowing community members to express their individuality (Fig. 3–6); Amazon provides no such facility. On the contrary, Amazon invites reviewers to summarize their review in the form of a short title, which makes it easier for visitors to quickly scan a large number of different opinions (Fig. 3–5). May be the most substantial difference between the review presentation on Amazon and Goodreads is the promotion mechanism. Amazon asks users if the review they just read was helpful or not helpful (Fig. 3–5), thus giving them the option to promote the review (by marking it as helpful), to demote the review (by voting "no") or to keep the status quo (by expressing no choice). In contrast, Goodreads offers a singular review promotion facility - a "like" button, there is no option for users to "dislike" or demote a review. In the next chapter we present the results of an experiment comparing the perceived meaning of these two seemingly very similar features ("helpful" and "like").

Figure 3–3: Amazon offers professional reviews to set the stage for user contributed content.



Editorial Reviews

Amazon.com Review

The amazing popularity of *Harry Potter and the Sorcerer's Stone* means that now even Muggles know about the Leaky Cauldron, Diagon Alley, and Hogwarts School of Witchcraft and Wizardry. Whether or not you've read about Harry, this unabridged audiobook brings his world to life. Reader Jim Dale brings an excellent range of voices to the characters, from well-meaning Hermione's soft, earnest voice to Malfoy's nasal droning; from Professor McGonagall's crisp brogue to Hagrid's broad Somerset accent; and from snarling Mr. Filch to p-p-poor, st-tuttering P-Professor Quirrel. Some of the characterizations are peculiar--why do the centaurs have Welsh accents?--but that's a small price to pay to hear one of the myriad ways to sing the Hogwarts School song. Harry Potter fans of all ages--Muggle or not--will enjoy curling up with a few chocolate frogs, a box of Bertie Bott's Every Flavor Beans ("Alas! Ear wax!"), and this marvelous, magical audiobook. (Running time: 8 hours, 6 cassettes) --*Sunny Delaney* --*This text refers to an out of print or unavailable edition of this title.*

From Publishers Weekly

The breakaway bestseller is now in paperback. In a starred review, PW said, "Readers are in for a delightful romp with this debut from a British author who dances in the footsteps of P.L. Travers and Roald Dahl." Ages 8-12.

See all Editorial Reviews

Figure 3–5: Amazon offers "customer reviews" featuring a title, a link to comments associated with the review, the choice of marking the review as either "helpful" or "not helpful" and the option to report the review as abuse of the reviewing guidelines.

Customer Reviews

11,935

4.8 out of 5 stars

5 star		86%
4 star		9%
3 star		3%
2 star		1%
1 star		1%

Share your thoughts with other customers

Write a customer review

See all 11,935 customer reviews

Top Customer Reviews

Three Harry Potter Books in Three Days!

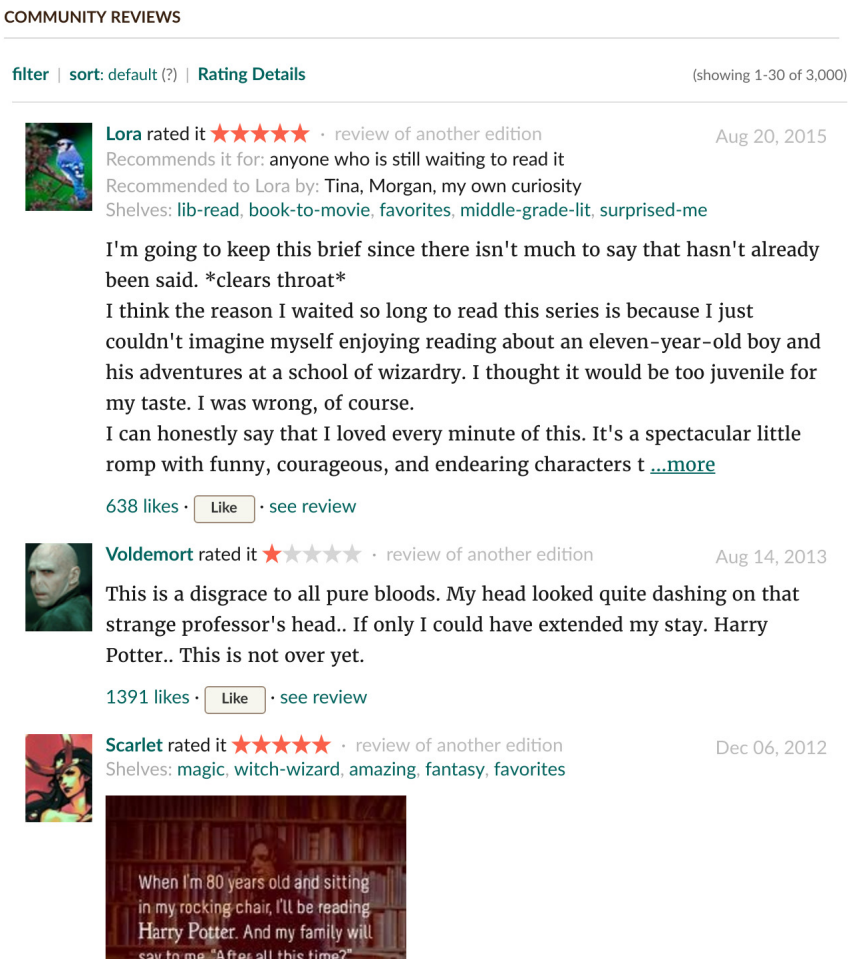
By Don Halpern on July 1, 2000

Format: Hardcover

An adult friend (age 49)loaned me three Harry Potter books for the summer. Wednesday evening I began the first book and I finished the third today, Saturday morning. I am writing this review before I order the fourth Potter book. Will my friend be surprised to get 4 books back! The author's imagination is vividly presented in a cast of almost believable characters attending a school we all wish we could attend. Classes like "Defense Against Dark Arts", "Divination", "Transfiguration", "Arithmancy" and "Care of Magical Creatures" are written as if the author actually attended them and certainly enjoyed every minute of class. More than can be said for most of the classes I have attended. Each book in the series encompasses one year of Harry's fascinating life. The Potter books are written in a way that can charm any age reader. I am 64.

9 Comments | 373 people found this helpful. Was this review helpful to you? Yes No Report abuse

Figure 3–6: Goodreads provides "community reviews" with the avatar (profile picture) prominently featured to the left of the review and a convenient option to "like" a review, but not to "dislike" it.



Another difference between the two platforms is that Amazon reviews are fully textual while Goodreads allows users to enhance their expression with embedded pictures (Fig. 3–6). Both platforms allow registered users to comment on existing reviews. Amazon shows the number of comments below the reviews and allows visitors to expand them on the same page (Fig. 3–5). On Goodreads a visitor needs to navigate away from the book page onto a specific review's page in order to see comments from other users. Due to these inconsistencies we did not consider review comments for the current study. Reviews on Amazon are sorted by the number of helpful votes by default while Goodreads uses a proprietary sorting algorithm which considers the number of likes, review length and age. Amazon also distinguishes reviews from verified buyers through a "verified purchase" badge. This features contributes to the trustworthiness of the review, an important factor for the purchasing decision.

Amazon actively discourages reviews in foreign languages on its platform by classifying them as "Inappropriate Content" in its Reviewing Guidelines [1] resulting significantly fewer non-English reviews for the English-language editions of the books we looked at compared to Goodreads which has no such requirement and favours a variety of community content.

Both platforms may display third-party ads, significantly more prominent on Goodreads where they have traditionally been the main source of revenue. We found the number of ads to be inconsistent and dependent on the user profile with Goodreads prominently displaying 2 or 3 ads to the right of the book description and

above reviews. In contrast, Amazon showed at most one ad usually placed further down on the book page.

3.2 Dataset

3.2.1 Biography Books

We crawled book reviews from the Amazon and Goodreads publicly facing websites. We started this study by crawling the Goodreads.com website. A Python script downloaded the list of all genres (also known as "shelves" on the Goodreads platform). For every genre, the script downloaded the list of books and corresponding book pages. We found that Goodreads uses AJAX requests accompanied by an authentication token in order to present reviews on the book pages. Thus, our Python script mimicked the AJAX request and authentication token generation, and downloaded all reviews for every book. The script got blocked by Goodreads when it reached the business genre. Hence, our initial dataset collection consists of all books in the arts and biography genres, and some books in the business genre. Given that a book can be assigned to multiple shelves, we crawled 1,095,810 books from 685 genres but only for a few genres we got all books available on Goodreads. Hence, we chose to focus on the

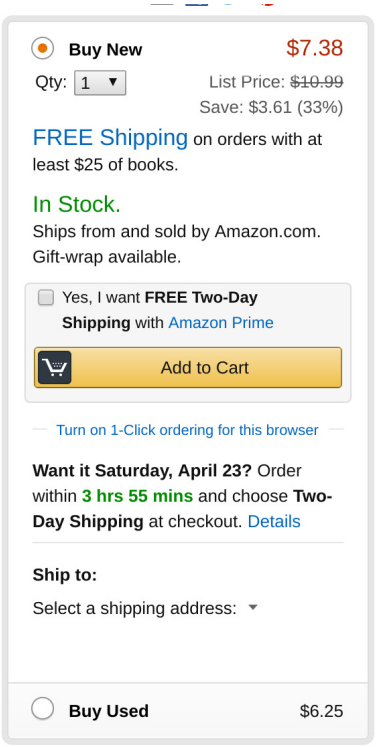


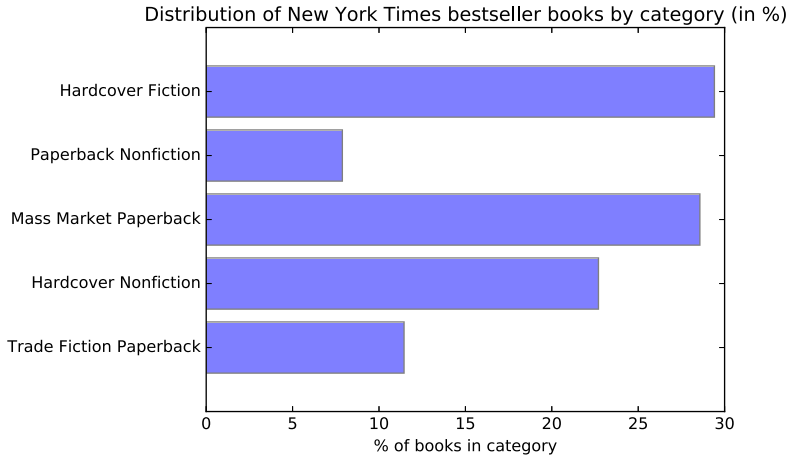
Figure 3–4: Amazon provides a convenient basket management panel to the side of the book description for easy purchase. Multiple purchase options are offered.

biography genre because it is highly coherent (i.e. presents a single individual's life) and very popular. We should note that "genre" is user defined on the Goodreads platform, thus we considered as "biography" any book added to this shelf by at least one Goodreads user and having at least one review. Our dataset contains a total of 10,328 biography book pages from Goodreads, and 1,600,471 reviews. Having access to the full set of reviews for a genre (as opposed to a sample) allowed us to avoid the limitations of non-random sampling resulting from the proprietary review sorting algorithm on Goodreads. For every biography book from Goodreads we obtained the equivalent book page on Amazon by using regular expressions to extract the ISBN number. We then crafted requests to the Amazon.com search engine with the ISBN number as a query parameter. We used wget [25] to send these requests. If the book was found, the response was a redirect to the book page on Amazon. We extracted the unique identifier of every book from the page URL (it is often the ISBN number but for some books, available only in Kindle format on Amazon, it is the ASIN number). Using this identifier we requested every customer review page for every book. Amazon started sending CAPTCHA responses to some of our requests. These responses were easy to identify because they were all of the same size, much smaller than a regular review page. We iteratively filtered the list of review requests that received CAPTCHA and resent them until we got all review pages. We were able to crawl a total of 10,574 book pages, 246 books fewer than our Goodreads dataset. The books we couldn't match between the two platforms were a result of Goodreads not listing the ISBN number or Amazon not offering the specific book for sale. A total of 945,548 Amazon reviews were collected.

3.2.2 New York Times Bestsellers

Between February and April 2015 we crawled the New York Times Bestsellers lists published from January 3, 2010 to January 3, 2015 covering the categories: hardcover-fiction, hardcover-nonfiction, trade-fiction-paperback, mass-market-paperback and paperback-nonfiction[1] . Most of these books are very popular and likely to attract reviews written and read by mainstream readers, which was in line with our focus for the second study. See Figure 3–7 for the distribution of books into the five categories.

Figure 3–7: The distribution of books across New York Times Bestsellers categories between January 3, 2010 and January 3, 2015.

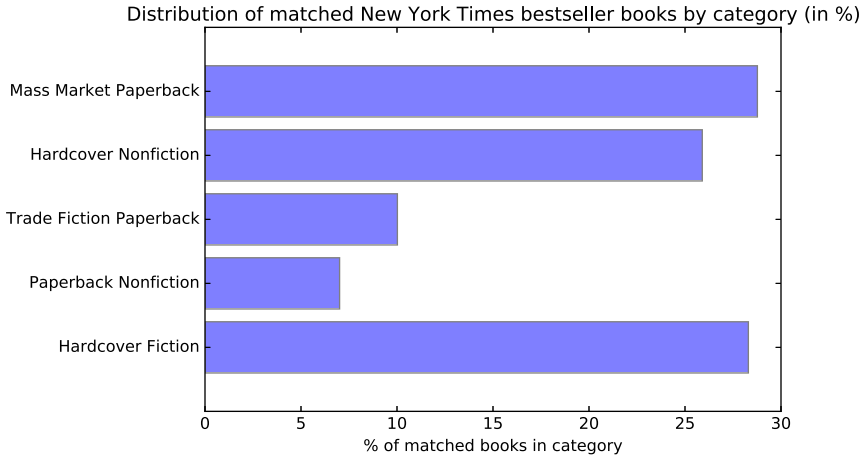


We extracted the ISBN from the lists and crafted requests for Goodreads and Amazon to get the corresponding book page. For instance, a request to Amazon may look like: "http://www.amazon.com/dp/1594480001" while on Goodreads we would

[1] http://www.nytimes.com/best-sellers-books/

request "http://www.goodreads.com/search?q=1594480001" resulting a redirect to the book page if there is a match. Due to the redirection of requests, sometimes we got a book edition with different ISBN number from the one we requested. We were able to match 3,381 books with reviews on both platforms out of 4,176 listed on the New York Times bestsellers. Figure 3–8 shows the distribution of the matched books by categories. It is largely the same as the original New York Times bestsellers distribution (Fig. 3–7) allowing us to draw conclusions about bestsellers without category bias. For every one of these books we crawled up to 60 of the most recent reviews available on each platform.

Figure 3–8: The distribution of the 3,381 New York Times Bestseller books we were able to match between Goodreads and Amazon by category.



After removing duplicate and foreign language reviews (5,195 from Goodreads and 31 from Amazon) identified using the langid language identification tool [19], our final cleaned review dataset included 189,329 Goodreads reviews and 195,195

Table 3–1: Metadata we collected for each review.

Attribute	Amazon	Goodreads
Review author	ID + nickname or anonymous	ID + nickname
Rating	1-5 stars	1-5 stars
Promotion Count	# helpful votes # unhelpful votes	# likes
Comment Count	# comments	N/A
Verified Purchase	Y/N	N/A

Amazon reviews, for the same 3,381 books matched between the platforms. Table 3–1 shows the metadata we extracted for each review.

CHAPTER 4
Results and Discussion

4.1 Biography Results

We start our analysis with the biography book reviews dataset as it is more comprehensive for the specific genre. Next, we demonstrate the validity of our results for a broader range of mainstream books and show the outcome of the additional experiments we performed to gain a deeper understanding of the underlying differences between Goodreads and Amazon.

4.1.1 Statistical Analysis

As part of the review metadata we collected the username of the reviewer on both platforms, giving us a total of 581,409 unique users on Goodreads and 631,922 unique users on Amazon. However, as discussed previously, Amazon allows anonymous reviews on its platform. For the purpose of the statistical analysis we excluded 44,023 reviews written by anonymous users from the Amazon dataset. Our results show that on average Goodreads users tend to write more reviews, indicating higher engagement with the platform (2.75 reviews per user) than Amazon (1.51 reviews per user). On the extremes of both platforms we have few users who contribute disproportionally high number of reviews (a Goodreads reviewer wrote as many as 371 reviews in our dataset, and an Amazon one contributed 699 reviews). Thus, Goodreads has fewer reviewers but they are more engaged with the platform.

Table 4–1: Statistical comparison of Goodreads and Amazon biography book reviews. Statistics annotated with a * have statistically significant differences ($p < 0.00001$). Positive difference indicates higher value on Goodreads.

Statistic	Goodreads	Amazon	Difference
# of Reviews*	1,600,471	945,548	+654,923
# of Users*[a]	581,409	631,922	-50,513
Avg # Reviews/User*	2.75 ± 5.35	1.51 ± 2.90	+1.24
Avg # Reviews/Book*	147.92 ± 357.15	95.65 ± 355.50	+52.27
Avg Review Length* (words)	87.09 ± 131.10	117.84 ± 164.36	-30.75
Avg Review Length* (sent.)	5.0 ± 5.78	5.95 ± 6.79	-0.95
Avg Sentence Length*	99.25 ± 124.44	110.44 ± 96.61	-11.19
Avg Stars*	3.86 ± 1.04	4.33 ± 1.09	-0.47
Avg Stars/Book*	3.88 ± 0.31	4.27 ± 0.51	-0.20
Avg Sentiment/Review	0.04 ± 0.08	0.04 ± 0.07	0
Avg Sentiment/Book	0.04 ± 0.02	0.04 ± 0.02	0

[a] Anonymous Amazon reviewers were not included

Our dataset allows us to study the distribution of reviews per book because we collected all reviews for the set of books on each of the platforms. Both platforms may display the same review on multiple editions of a given book, and such duplicates were counted only once for this analysis.

When looking at review length we considered three measures: number of words per review, number of sentences per review and average sentence length. Surprisingly, we found that Amazon reviews for biography books are significantly longer on average based on all of these measures. This result is in agreement with existing literature comparing Amazon and Barnes & Noble [5] and can be explained by the persuasive intent of Amazon reviews: they are marked as"helpful" or "not helpful" depending on whether they can persuade a reader to buy or not to buy the book. Other studies

into participatory culture [17] would lead us to expect that Goodreads, being a platform centered around people with common interests, should attract more expressive and detailed reviews. However, according to the content analysis we discuss below, Goodreads reviewers engage in a more colloquial discussion, which explains the lower average length.

4.1.2 Book Ratings

Figure 4–1: The distribution of stars over all ratings for biography books

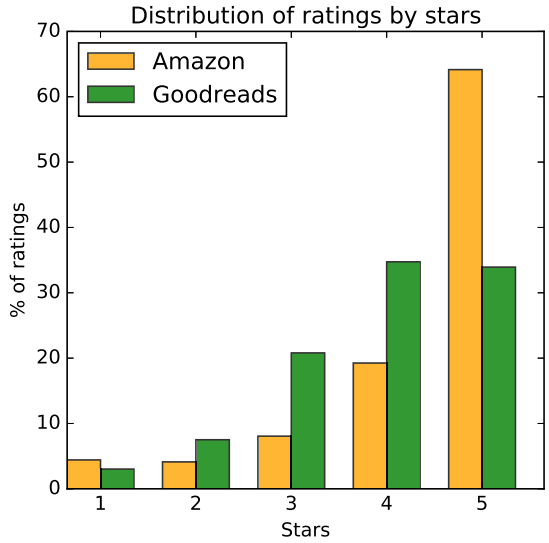
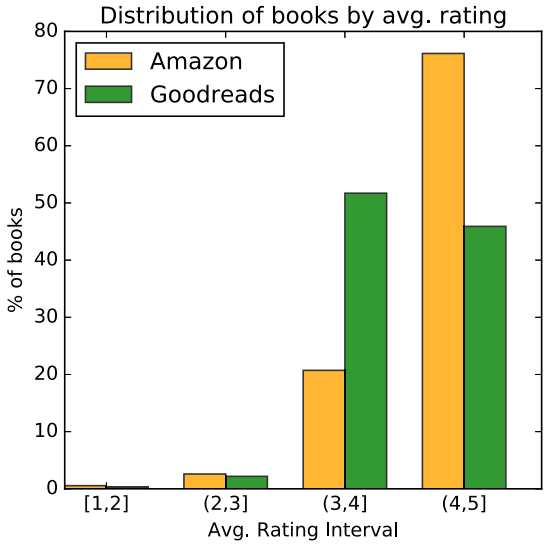


Figure 4–2: The distribution of biography books by average rating



Users may rate books as part of the reviewing process on both platforms. Ratings are on a scale from one to five stars. We studied the differences in the rating distributions for the same biography books between the two platforms and found that Amazon users give higher rating on average than Goodreads users (4.27 vs. 3.88). This discrepancy also exists when we consider other statistics, such as the median (4.4 on Amazon vs. 3.9 on Goodreads). Figure 4–1 shows the distribution

of biography books by their average rating on the two platforms. We see that most books (76.14%) have average rating between 4 and 5 stars on Amazon, while less than half of the biography books on Goodreads (45.90%) have average rating in the same interval. On the other hand, the frequency of books with average rating between 3 and 4 stars is double on Goodreads (51.70%) than it is on Amazon (20.72%). Given that these ratings are for the same books, it appears that books are rated higher on Amazon than they are on Goodreads. Is this result an indication of high disagreement between Goodreads reviewers (large number of very low ratings and equally high number of very high ratings) or does the average Goodreads user simply give lower ratings? To understand the underlying differences in ratings we leveraged the completeness of our dataset at the review level to evaluate the distribution of individual user rating submissions. We found that Amazon users are more likely to rate books with 5 stars (64.15% of all ratings), while Goodreads users more often provide a less-than-perfect rating of 4 stars (34.74% of all ratings on the platform), as shown on Figure 4–1. The distributions of ratings between the two categories (5 stars and 4 stars) are antithetical: only 33.93% of Goodreads ratings are 5 stars (almost half of the Amazon equivalent) and only 19.25% of Amazon users gave 4 stars. If we look at the other extreme (1 star) we find that Amazon users are more likely to give 1 star rating than Goodreads users (4.42% of ratings vs. 3.02% on Goodreads), but the contrast becomes even more striking at at the 2-3 stars range where the frequency of Goodreads ratings is double that on Amazon. Clearly Goodreads users opt for more balanced votes (2, 3 or 4 stars) which may indicate that they are evaluating the book in the context of other books they have read, or that their rating is a comprehensive

summary of multiple aspects of the book. On the other hand, Amazon reviewers are more likely to rate a book as a "buy" by giving it the maximum of 5 stars, or sometimes as "don't buy" by marking it as 1 star. This rating behaviour can be explained by the implicit goal of Amazon reviews to facilitate the buying decision. Extreme ratings are more persuasive when attempting to sway the reader towards buying or not buying a book, as has been demonstrated in previous work [5].

4.1.3 Sentiment Analysis

As a first step in uncovering the content differences between Amazon and Goodreads reviews we consider their sentiment valence. For our experiment we used the SentiWordNet dictionary [3]. This lexical resource assigns a positive, negative and objective score to each word sense. We split our reviews into words and for every word we looked up the posterior polarity of the most popular meaning [14] in SentiWordNet. Based on the positive and negative scores of every word we calculated the following measures for each of the two platforms:

- average positive review sentiment (sum of all positive words' values in a review normalized by all words)
- average negative review sentiment (sum of all negative words' values in a review normalized by all words)
- absolute sentiment (sum of the absolute value of all words in a review normalized by number of words)
- difference between positive and negative sentiment (sum of all positive values minus all negative values for words in a review normalized by number of words)

We found no significant difference in these sentiment measures between Goodreads and Amazon. To account for the difference in review length between the platforms we repeated the experiment for a subset of all reviews: the ones between 10 and 20 words in length. Again, we failed to find significant difference in any of the measures. Finally, as Amazon offers an extra field for reviewers to express their sentiment towards a book (the review title) we repeated the experiment by appending the title to the review text for reviews less than 20 words in length. Still, we found no significant differences in terms of sentiment expression between the platforms. The SentiWordnet results indicate that sentiment is stable between the two platforms, which is unexpected given the discrepancy in star ratings. It is possible that a simple frequency-based SentiWordnet look-up approach does not reflect the emotions expressed in book reviews. For instance, when reviewers discuss the book plot they use words which have sentiment value but are not relevant to the overall sentiment of the review. Finally, we performed an experiment with a hedonometer containing the average happiness value of 10,221 words drawn from a Twitter dataset[7] and evaluated by workers on the Amazon Mechanical Turk crowdsourcing platform. By using sentiment values from the hedonometer dictionary we found a more positive average sentiment for Amazon reviews. Thus, future research may apply a more sophisticated natural language processing (NLP) algorithm in order to account for the meaning of reviews, and this could provide more conclusive sentiment results.

4.1.4 Content Analysis

To better understand the differences in vocabulary between Amazon and Goodreads reviews we applied a Wilcoxon rank-sum test. First we filtered common English stop-words from our reviews datasets. Then we calculated the frequency of occurrence for each word (unigram) in each of the platforms and excluded rare words which appear less than 1000 times. Finally, we ranked each word by applying the Wilcoxon rank-sum test [34]. This approach has been previously found to identify terms that are more common in one corpus than another without favouring very rare words (the result when comparing frequency ratios) or very common words (which we get when comparing the absolute magnitude of the difference in frequencies). The top ten ranked words for Amazon reviews were: "buy", "bought", "will", "purchased", "reader", "gift", "purchase", "ordered", "highly", "reviewers", "price". A number of these words refer to the purchase, or are used to argue whether the reader should buy the book. As an example, we took a random review which uses the word "highly" and this is the context it was used in: "[...]James' second book "My Friend Leanord" is also one of the most interesting stories that I have ever read. I recommend it highly, as well.Whew! I've been dying to get that off my chest for a long time. Please read the book and see if you don't agree.[...]" In contrast, the top ten words most specific to Goodreads were: "goodreads", "shit", "interesting", "pretty", "memoir", "bit", "listened", "funny", "definitely", "didn", "parts". In this list we see a mix of very colloquial, even vulgar, vocabulary with words that are often used to discount an argument (such as "bit" in "a bit", "a little bit" or "pretty" in "pretty good", "pretty bad", etc.) and literary language ("memoir", "parts"). We queried our Goodreads

reviews dataset for a random review which includes the word "listened" and coincidentally the result was one that uses a number of the top 10 words on Goodreads: "I have not the words. I loved everything about this book, a memoir in mix tapes by a writer for Rolling Stone. In funny, heartbreaking, beautifully written words, the author writes a tribute to his late wife organized by the mix tapes he made and listened to at the time." The book content is a common theme of discussion in Goodreads reviews. Based on our content analysis we can say that Amazon reviews often discuss arguments for and against buying a given book, while Goodreads reviews are more reflective of community conversations and reflection about the book. An interesting direction for future research may be to identify prolific users on each of the two platforms and to determine the extent to which their vocabulary may be influencing the overall review content. This type of study would be most valuable if it correlates user profiles between Goodreads and Amazon in order to answer the question whether the platform influences the way users write or certain users write on one platform and not the other. Anonymity and privacy protection features on both platforms make collecting the user profile data required for such study challenging.

4.2 Bestsellers Results

Our data shows that Goodreads attracts more reviews per book on average for the biography genre (147.92 vs. 95.65 on Amazon) but the trend is reversed for highly popular biography books, with the most reviewed book on Goodreads having 3,000 reviews compared to 15,990 for the most reviewed one on Amazon. This finding indicates that Goodreads may offer a more engaged community of reviewers for non-mainstream books, but Amazon is the platform of choice for bestsellers. Expanding

upon our findings for biography books we performed similar analysis for a sample of the most recent 60 reviews for each of the New York Times bestsellers on the two platforms.

4.2.1 Buying Experience

Our content analysis experiment on biography book reviews showed that some of the words most specific to Amazon are those used when discussing the buying experience. However, the Wilcoxon rank-sum test doesn't help us understand how common these words are on Amazon, it just tells us that they are significantly more common than on Goodreads. To find out how common they are we took 1,000 random Amazon reviews from our dataset and manually annotated each one of them with a binary annotation (true if the review mentions the purchase, delivery, shipping or transaction, false otherwise). We found that only 19 of the 1,000 reviews discuss the purchasing experience in any form. While this is surprising given that Amazon is an on-line shopping site and selling is its main activity, it may be explained by Amazon's reviewing guidelines which stipulate that "Feedback about the seller, your shipment experience, or packaging is not a product review and should be shared at www.amazon.com/feedback or www.amazon.com/packaging". We have no information as to how many Amazon reviewers actively read these guidelines or abide by them, but the platform does state that reviews not complying with the guidelines may be removed or rejected during the moderation process. Interestingly, one of the 19 buying related reviews we identified discussed the purchase experience for an item that is not a book at all. Hence, some reviews do slip through the Amazon moderators.

4.2.2 Linguistic analysis

To find out how the platform influences the writing style of its reviewers we performed a linguistic analysis using the Linguistic Inquiry and Word Count (LIWC) tool [29]. This tool has a predefined dictionary of words and word stems organized in categories such as words describing linguistic processes or psychological processes, etc. It iterates over the input documents (a document in our experiment consisted of all reviews for a given book on a given platform) and records the frequency for each word and each category, as well as the frequency of punctuation marks and sentence / word length statistics. We provide the results from the LIWC analysis in table 4–2.

Our results show that Amazon reviews more frequently use exclamation marks and words from the certainty language category ("unambigu*", "fundamentals", "perfect*", "always", and "guarant*") and the second person singular pronoun "you", while Goodreads reviewers use more diverse punctuation (parentheses, colons, commas, dashes, question marks and other punctuation are all significantly more frequent on Goodreads than on Amazon) and words from the tentative language category, particularly colloquialisms such as "lotsa" (4517.70%), "dunno" (289.68%), "shaky" (180.47%), "kinda" (177.54%), as well as negative and anger vocabulary. These differences indicate that Amazon reviewers express more confidence with their writing style and direct their comments to the reader, potentially in an attempt to facilitate the buying decision. In contrast, Goodreads reviewers are more direct in their expression, discuss more nuanced sentiments and sometimes use profanity (which is explicitly forbidden on Amazon). Their reviews more frequently address

Table 4–2: Frequency of linguistic features (expressed as % of words belonging to each dictionary) in reviews, by platform. All statistics shown are significant with $p < 0.001$. Relative difference is calculated with respect to Amazon, where positive numbers indicate higher frequency on Amazon and negative indicate lower frequency on Amazon.

LIWC Category	Measure	Amazon	Goodreads	Difference
Linguistic processes	Swear words	0.04 ± 0.09	0.07 ± 0.11	-75.00%
	Numerals	0.34 ± 0.30	0.52 ± 1.63	-52.94%
	Exclamation marks	1.12 ± 0.93	0.6 ± 1.98	+46.43%
	Question marks	0.17 ± 0.17	0.22 ± 0.39	-29.41%
	Parentheses	0.22 ± 0.18	0.33 ± 0.26	-50.00%
	Colons	0.11 ± 0.12	0.16 ± 0.19	-45.45%
	Commas	3.37 ± 0.96	4.05 ± 1.25	-20.18%
	Dashes	1.06 ± 0.64	1.35 ± 1.00	-27.36%
	Other punctuation	0.25 ± 0.42	0.35 ± 3.22	-40.00%
	Numbers	1.72 ± 1.00	1.08 ± 0.51	+37.21%
	"you"	0.79 ± 0.41	0.57 ± 0.46	+27.85%
Psychological processes	Positive emotions	6.66 ± 2.02	5.16 ± 3.16	+22.52%
	Tentative language	2.21 ± 0.62	2.54 ± 0.82	-14.93%
	Certainty language	1.76 ± 0.48	1.5 ± 0.53	+14.77%
	Negative emotions	1.71 ± 0.70	1.95 ± 0.88	-14.04%
	Anger	0.53 ± 0.41	0.69 ± 0.51	-30.19%

the community as a whole and not a specific reader trying to decide whether to buy the book. Another result shows that numerals ("firstly", "quarter", "half", "first", "second", "third", etc.) are more common on Goodreads. We manually annotated 100 reviews which use numerals and found that the most common uses for numerals are to place the book within series, to discuss historical events (eg. "first world war", "second world war", etc.) and finally to address specific sections within the book (i.e. "first chapter", "second half", etc.). In contrast, Amazon reviews have higher frequency for numbers. To understand the context they are used in we annotated 100 random reviews that use numbers. We found that the primary use for numbers is to justify the star rating that accompanies a give review (eg. "I gave it 5 stars, but [...]") as well as to refer to time periods and dates (thus, similar to the use of numerals), and lastly a few reviews used numbers to mention specific pages within the book. These differences in frequencies for numerals and numbers suggest that reviews on Goodreads often discuss the book content and its relationship to other books while Amazon reviews are interested in giving an accurate evaluation of the book as a product.

4.2.3 Star Rating Analysis

We showed that the star rating distribution for biography book reviews on Amazon is more extreme than on Goodreads with most ratings in the 5 star category and more ratings in the 1 star category, as compared to Goodreads where most ratings are in the 2, 3 and 4 star categories. We repeat this experiment for New York Times bestsellers reviews to test whether our results can be generalized to other genres.

Figure 4–3: The distribution of stars over a sample of ratings for NYT bestsellers

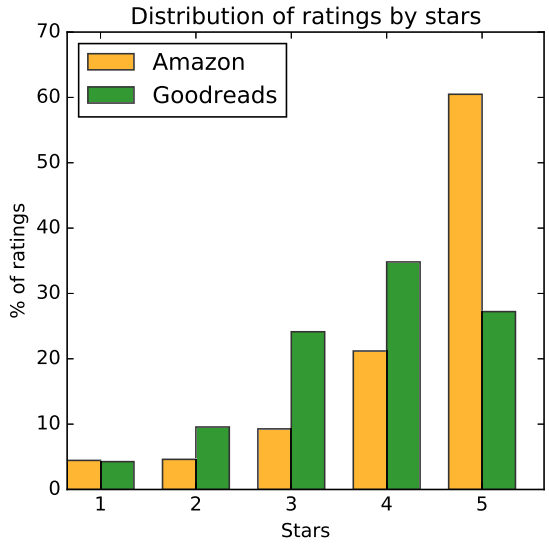
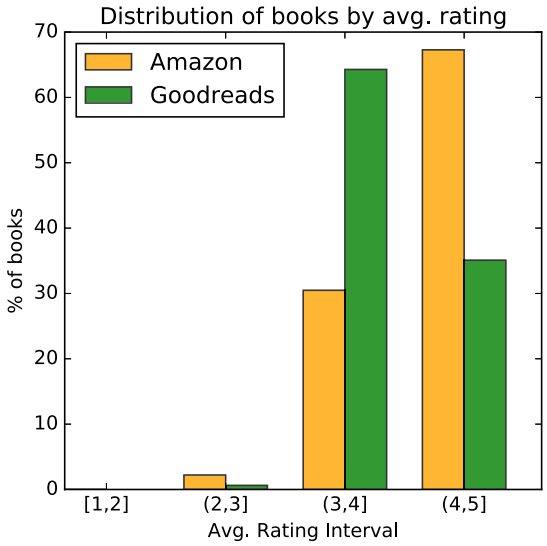


Figure 4–4: The distribution of New York Times bestsellers by average rating



As shown on Figures 4–3 and 4–4 our findings remain valid for the sample of New York Times bestsellers reviews. The average rating per book is higher on Amazon than on Goodreads (4.15 vs. 3.89); similarly the maximum average rating for any book in the dataset is also higher on Amazon (5.0 stars) with no book having a full 5 star rating on Goodreads. Yet, the minimum average rating is lower on Amazon than on Goodreads (1.8 vs. 2.36) suggesting that Amazon ratings are more extreme. If we consider the distribution of books with average ratings between 2 and 3 stars (Figure 4–4) we find that 85 Amazon books fall in this category (representing 2.21% of all books in our dataset), while only 24 Goodreads books have average rating in this range (0.63% of all Goodreads books in our dataset). Most books on Amazon have average rating between 4 and 5 stars (67.28%), while most books on Goodreads have average rating between 3 and 4 stars (64.28%). This finding is very interesting

when considered in the context of biography books because it shows that Goodreads reviewers are even more critical of bestsellers than they are of biography books in general. These statistics are based on the book page as displayed on the platforms, hence they include all ratings and are not calculated from our sample of 60 reviews per book.

If we look at our reviews dataset we find a similar rating distribution (see Fig. 4–3). Only 4.26% and 9.56% of all Goodreads reviews in our sample were accompanied by a rating of 1 or 2 stars respectively. For Amazon we also found that only 4.45% and 4.61% of all reviews were accompanied by a rating of 1 or 2 stars respectively. This finding is similar to the results we presented earlier for biography books, with the understanding that the biography study was based on all available ratings for all biography books, while here we consider only a sample of ratings for books that have achieved bestseller status. if we look at the 3, 4 and 5 star ratings the results are also in agreement with the average rating: most Goodreads reviews (34.82%) are accompanied by a 4 star rating, while for Amazon most reviews carry a 5 star rating (60.48%). On Goodreads 5 star ratings are given by 27.22% of reviewers.

We performed a χ^2 contingency test and found that the observed differences have high statistical significance ($p < 0.001$). If we define 1 and 5 star ratings as extreme, and 2, 3 and 4 star ratings as non-extreme, we can apply a two-proportion Z-test to test whether extreme ratings have the same proportion on both platforms. The results allow us to reject such claim with a high degree of significance ($p < 0.001$), indicating that Amazon reviewers give more extreme ratings than Goodreads reviewers. This finding is in agreement with the linguistic analysis discussed above,

confirming that Amazon reviewers are focused on evaluating a given book as "a good buy" or "not a good buy" while Goodreads reviewers provide a more comprehensive evaluation of the book's content, potentially in comparison to other books.

Two factors may explain the differences in review content and star ratings between the two platforms: population selection and platform influence. It is possible that Amazon and Goodreads attract different types of users (by age, level of education, or other characteristics) who apply star ratings differently. It is also possible that the user base is largely the same but the design, goal and existing content on the platforms influence users to rate books in a certain way. Finally, a combination of these two factors may be at play on Goodreads and Amazon.

4.2.4 Review Classification

Given all the differences we found between the two platforms the next question is are these differences significant enough to allow a classifier to infer the platform of a review. To make a classifier perform well we need to identify features that are not only different between the platforms but also prevalent enough to define all reviews. Initially we designed the experiment as a per-review binary classification problem: each review in our dataset was labelled as belonging to either Goodreads or Amazon, 10% of reviews were set aside as validation set and 10-fold cross validation was applied on the remaining reviews. Our features included: review length, frequency of words from the linguistic analysis discussed above, frequency of 101 terms and expressions identified by a literary expert and a set of binary features based on the mention of the book title, book author, the title of another book from our dataset, and the name of any author from our dataset. We trained an SVM classifier on these features but

the results we obtained were not satisfactory (62% accuracy). To better account for the potential impact of book genre on the language of reviews we repeated the experiment separately for fiction and non-fiction reviews, but the results were similar. Finally, we looked at the review distribution by length on each of the two platforms and identified three major subsets of reviews (1-50 words, 50-150 words and 150-5048 words). We performed the classification experiment on each of these subsets to uncover differences between the platforms that may be more pronounced within reviews of certain length. This separation offered a marginal improvement of the classification accuracy for short reviews but overall the results remained at <64% accuracy, indicating that while the differences between the platforms are significant they may not be manifested at the review level. Thus, we approach the classification problem at the review ensemble level: given all reviews for a given book from a platform, can we determine which platform they were taken from?

Feature selection

We chose features for the classifier based on the results we obtained from the statistical and content analysis (see Table 4–3). Working at the ensemble level gave us the freedom to choose features with varying level of granularity. For instance, we consider review length to mean the number of reviews that fall within a given range within the review ensemble (i.e. this feature would have value 10 for length 1-100 words if 10 reviews within the ensemble have length more than 1 word and less than 100 words). The review length ranges we used were: 0-5 words, 6-20, 21-40, 41-55, 56-220, and greater than 220 words. We chose these range values as they maximize the information gain. Other features are at the ensemble level, such as

Table 4–3: Features used for the classification of review ensembles, by level of granularity.

Granularity	Feature set	# of Features
Review	Review length	6
	Book title / author	4
Sentence	Sentence length	2
Word	LIWC dictionaries	65
	Literary features	101

the number of times the book title is mentioned, normalized by the total number of words in the ensemble. Yet another set of features are at the sentence level, for example we consider the number of sentences of length less than 6 words in the ensemble and the number of sentences of length more than 25 words as two features. For every one of the LIWC dictionaries we summed the word frequencies within the ensemble for all of its words/stems, giving us one feature per LIWC dictionary. As we discovered in our biography book reviews content analysis study, Goodreads reviews often contain terminology specific to the literary domain. We consulted a literature expert to compile a dictionary of 101 words and expressions commonly used in literary analysis ("narrative", "point of view", "protagonist", "ending", etc.). The frequency for each one of these words in an ensemble normalized by the total number of words in that ensemble constituted one feature in our classification dataset. As shown on table 4–3 our classification dataset consists of 178 features per ensemble, all normalized at the respective granularity level. Each of the features is described on table 4–4.

Training and classification

As the New York Times bestsellers lists contain both fiction and non-fiction books, we decided to conduct our classification experiment separately for these two

Table 4–4: Description of each feature used for classification

Feature Set	Description
Review length	Number of reviews having: less than 5 words, 6 to 20 words, 21 to 40 words, 41 to 55 words, 56 to 220 words, more than 220 words
Book title / author	Number of reviews mentioning: the book title, another book's title, the author, author of another book
Sentence length	Number of sentences: shorter than 6 words, longer than 25 words
LIWC dictionaries	The frequency of words from each of 65 dictionaries related to: Linguistic Processes (different kinds of pronouns, articles, adverbs, etc. and swear words), Psychological Processes (words used to describe the social environment, emotions, perception, body parts, space, time, etc.), Personal Concerns (mostly nouns and verbs related to everyday activities, home, money, religion, etc.) and Spoken categories (Assent, Nonfluencies and Fillers)[a]
Literary features	The frequency of each of the following terms: story, end, time, felt, life, character, people, book, characters, plot, world, feel, ending, stories, history, half, place, writing, novel, fact, chapters, written, fiction, chapter, lives, feeling, historical, feels, narrative, voice, perspective, ends, third, details, setting, scenes, writes, read, scene, feelings, genre, telling, first half, reads, places, drama, detail, information, storyline, conflict, protagonist, dialogue, narration, detailed, description, wrote, voices, point of view, structure, describe, fictional, sentence, narrated, describes, sentences, reading, pacing, develop, facts, quarter, developing, write, reader, protagonists, development, portrayed, readers, portrayal, endings, last half, narrates, settings, thirds, narrate, points of view, plots, quarters, fictionalized, plotted, worlds, dramas, descriptive, portrays, depicts, portraying, portray, novels, developed, depiction, depict, develops

[a] http://www.liwc.net/LIWC2007LanguageManual.pdf

genres. We justify our decision with the fact that non-fiction books are utilitarian products, while fiction books are hedonic products [23]. This difference may impact the way they are reviewed, with fiction books potentially triggering more subjective reviews based on reviewer preferences. We split our dataset into training set (90% of the data) and validation set (10%, 444 ensembles for fiction and 220 ensembles for non-fiction). Using the Scikit-learn Python library [28] we performed a grid search on the training set with 10-fold cross-validation and objective accuracy, on decision tree classifier, multinomial Naive Bayes, random forest classifier and support vector machine (SVM), all with default parameters. We identified the SVM as best performing and performed further grid-search with 10-fold cross-validation on the kernel type (linear and RBF) and hyper-parameter values. The highest F_1 scores our classifiers achieved were 95% for fiction book reviews and 92% for non-fiction. These results are very positive in comparison to our initial attempt with platform classification at the review level. The substantially different classification scores can be explained in part by the rich variety of user generated content available on both Goodreads and Amazon. Our statistical and content analysis show that review length and reviewer vocabulary exhibit significant differences at the platform level, however not every review demonstrates this distinctness. Furthermore, with the buying experience experiment we showed that while purchasing related terminology is unique to Amazon, it appears in less than 2 percent of reviews. Thus, an ensemble of reviews is able to capture these infrequent but largely indicative features. Similarly, a user skimming over a book page on Amazon and Goodreads is exposed to at least

15-30 reviews, hence will perceive the reviews as an ensemble with the platform specific characteristics that such grouping entails.

Ablation testing

We performed ablation testing to understand which of the 178 features in our classification dataset contribute the most to the separation of reviews between Goodreads and Amazon. We trained the same classifier with each one of the feature sets listed in table 4–5 and repeated the classification on the holdout validation set for fiction and non-fiction. Review length appears to carry the most weight in our classifier, indicating that the distribution of reviews in the 6 length categories we identified is different between the two platforms, and this difference is independent of book or genre. As our dataset consists of the latest 60 reviews for the same 3,381 books collected at roughly the same time, we can say that the review length difference is a platform specific phenomenon, which can be explained by the platform design and/or intended purpose, or the reviewer base it attracts. The second most important feature set consists of LIWC dictionaries. This is indicative of the differences in vocabulary employed on the two platforms. With their comprehensive nature, the LIWC word categories capture both the colloquial language of some reviews and the sophisticated literary expressions used in others. In addition, the third most indicative feature set (literary features) goes one step further in assigning values to words an expert would use in literary analysis. Both vocabulary and review length are features that are easily perceived by a visitor to the two platforms, hence the importance of our finding for the impact a platform has on its reviews. In contrast, the number of times a review mentions the book title or author, or the title/author

Table 4–5: F1 scores for classification of review ensembles when using each of the feature sets individually.

Feature set	Fiction	Non-Fiction
Review length	0.92	0.86
Book title / author	0.63	0.61
Sentence length	0.62	0.62
LIWC dictionaries	0.84	0.77
Literary features	0.70	0.64

of any other New York Times bestseller, as well as the length of individual sentences, do not appear to carry significant weight for review classification. In general, our classifier achieves better results on reviews for fiction books than for non-fiction. This finding could be an interesting topic for further genre-oriented study. One possible explanation is that the subjective nature of fiction books allows for a richer linguistic expression and an emotional (short) or an in-depth (longer) response to a book.

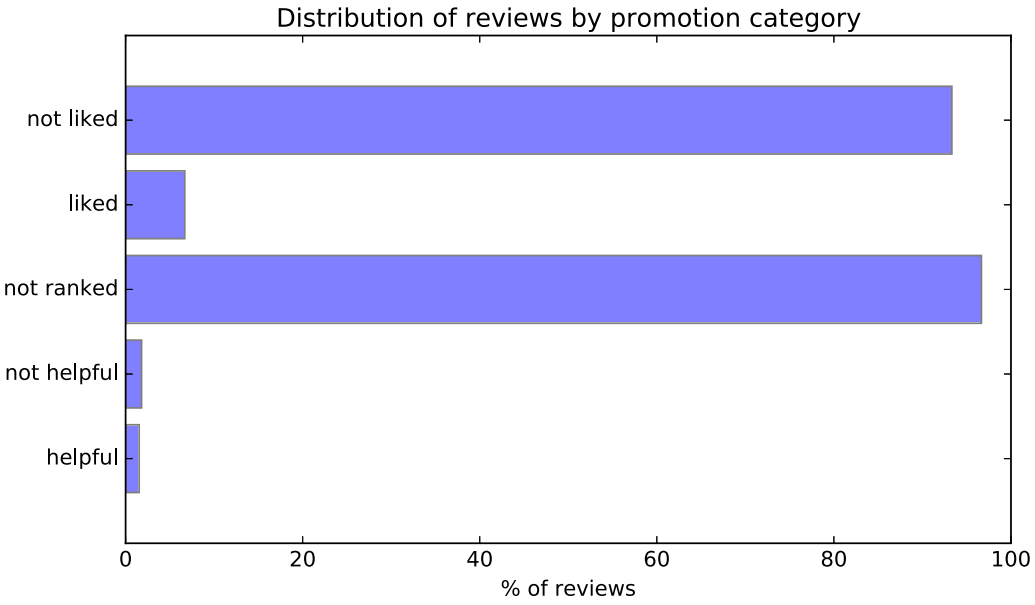
4.2.5 Analysis of Review Promotion

At the time of writing the default review ordering on Amazon is according to the number of helpful votes, while Goodreads uses a proprietary algorithm which considers the number of likes. We've already shown that the two platforms elicit substantially different reviews, but do they also influence users to promote certain type of reviews over others? Here we consider the subtle differences in wording ("helpful" vs. "like") and polarity of the action (positive, negative or neutral vs. positive or neutral) to uncover how the promotion features available on the two platforms are perceived by users. The kind of content users choose to promote may ultimately lead to more reviews of that type being generated on the platform, hence the importance of understanding the promotion mechanisms. At the onset we

expect a prevalence of "helpful" votes to be associated with reviews that facilitate the purchasing decision, for example by providing an objective evaluation of a book. In contrast, we anticipate "like" to be associated with reviews that are enjoyable to read, for instance humorous reviews.

Distribution

Figure 4–5: Distribution of Amazon reviews into "helpful", "not helpful" and "not ranked" categories, and of Goodreads reviews into "liked" and "not liked" categories. The numbers add up to 200% because they include reviews from both platforms.



We organize our New York Times bestsellers reviews dataset into the following five categories:

- helpful (reviews with more "helpful" than "not helpful" votes on Amazon)
- not helpful (reviews with more "not helpful" than "helpful" votes on Amazon)
- not ranked (reviews with 0 "helpful" and 0 "not helpful" votes on Amazon)

- liked (reviews liked by at least one user on Goodreads)
- not liked (reviews with 0 "likes" on Goodreads)

Notably, the proposed categorization excludes reviews with equal number of "helpful" and "not helpful" votes, which provide an ambiguous signal and confound our analysis. See Figure 4–5 for the distribution of reviews in each of the five categories. The frequency of "liked" reviews on Goodreads is much higher than the frequencies of "helpful" and "not helpful" reviews on Amazon. This finding speaks to the engagement of Goodreads users with the community, or to the meaning of the concepts "helpful" and "not helpful" with respect to the effect of the vote on a review's chance of being seen by other users. Another significant observation is that the frequencies of "helpful" and "not helpful" reviews on Amazon are similar, indicating that both the promotion and demotion mechanisms are applied to improve the visibility of content which the community considers helpful. Most reviews fall in the "not ranked" and "not liked" categories, as can be anticipated given that we took the latest 60 reviews, some of which may not have been on the platform long enough to receive votes.

Content Analysis

To better understand what makes a review helpful we looked at the LIWC dictionaries. First, we calculated the standard statistics (minimum, maximum, mean, median, first quartile, third quartile) and p-value at the dictionary level by counting the number of times any word from a given dictionary appears in a review from a given category normalized by the total number of words in the review. By performing pairwise comparison across the five categories (taking a random sample the size of

the smaller distribution when they were of unequal size) we were able to identify what type of language is associated with "helpful", "not helpful", "not ranked", "liked" and "not liked" reviews respectively. The dictionaries we found to be most distinct were "positive emotions" and "affective processes". We then drilled down to the word level to find which specific words from the two dictionaries are most indicative of each of the five review categories. The positive emotions dictionary contains a total of 405 words and word stems. We found that "great", "good", "love", "enjoy*", "loved", "excel*", "awesome", "fun" and "ok" have significantly higher frequency in "not ranked" than in "helpful" reviews on Amazon. Similarly, "great", "good", "excel*", "ok" and "love" have higher frequency in "not helpful" than in "helpful" reviews. What these words all have in common is that they are generic and express the author's subjective opinion about a book, as opposed to being used for in-depth analysis. Hence, our finding confirms the results of Mudambi and Schuff that superficial reviews are not considered helpful [24]. On the Goodreads side we found the words most frequent in reviews with no likes to be: "good", "great", "enjoy*", "love", "loved", "excel*" and "fun". These words are similar to the ones associated with "not helpful" reviews, i.e. contrary to our expectation users tend not to "like" reviews that use vocabulary also frequent in reviews marked as "not helpful". If we look at the dictionary level, words related to work or employment ("job", "majors", "xerox", etc.), leisure activities ("cook", "chat", "movie", etc.), social processes ("mate", "talk", "they", "child", etc.) and common verbs in present tense ("is", "does", "hear", etc.) are more frequent in reviews marked as "not helpful" than in reviews which received no likes. Such words may appear often when

discussing the book plot (i.e. in "spoiler" reviews) or they can be used to express the opinion of the reviewer, which may not be helpful when deciding to buy or not to buy a book. Articles ("a", "an" and "the") are more frequent in reviews which received "no likes" than in "not helpful" reviews. The use of articles is employed for linking content words together and speaks to the linguistic style of the author[30]. Hence, readers may be driven by their dislike for the quality of writing in choosing not to like reviews with high frequency of articles.

Even more surprising is the comparison between "not ranked" and "no likes" reviews which shows that the same words used to express emotions are more common in "not ranked" than in "no likes" reviews. This finding may indicate that reviewers tend to write more subjective reviews on Amazon than on Goodreads, or it may also be a result of a more sophisticated vocabulary on Goodreads.

Crowd-sourcing Experiment

We took a sample from our dataset such that for a given book we have equal number of reviews in each of the five categories discussed above. This selection resulted 476 unique reviews which we used as the dataset for a controlled experiment on the CrowdFlower[1] crowdsourcing platform. The experiment started with an instruction page indicating that participants should read the review and choose the answer they most agree with. Participants were also advised that non-English reviews and meaningless reviews should be marked as "not helpful" or "disliked". Once they confirm they had read the instructions participants had to read a page of reviews

[1] http://crowdflower.com

and answer a question after each one. Workers were assigned to one of two tasks: helpful or likes. For the helpful task they were presented with a page of 5 reviews and asked to mark each one as either "helpful" or "not helpful", and to choose one of the justifications listed in table 4–6. For the likes task we offered exactly the same justification options, but the question was whether they like or dislike each of the provided reviews. For both experiments we also allowed workers to enter a free-form comment if none of the choices justify their decision. Figure 4–6 demonstrates the design of a task used in this study. Every review was evaluated by six workers, three per experiment. Workers were randomly chosen by the Crowdflower platform and were equally compensated for their effort according to the number of reviews they evaluated. Hence, there was an implicit incentive for workers to evaluate as many reviews as possible in the least amount of time so that they could complete more tasks. Any given participant was allowed to evaluate at most 100 reviews. Workers could not leave reviews unmarked (i.e. our experiment disallowed the neutral choice of not marking a review as either "helpful" or "not helpful"). This limitation was necessary in order to keep the experiment unbiased while countering the tendency of crowdsource workers to spend as little time as possible on a given task (for example by providing the same answer to all questions). To this end we randomly introduced 60 negative test reviews ("not helpful" / "disliked") in our dataset, of the following four types:

- reviews containing only numbers (no words)
- short reviews containing random English words (syntactically incorrect and meaningless)

- longer reviews of random English words (also syntactically incorrect and meaningless)
- the popular "lorem ipsum" filler text (in Latin)

If a worker reads the task instructions he/she will know to mark reviews in the above categories as "not helpful" or "disliked". We disqualified participants who marked a test review as "helpful" or "liked" it, and removed their answers. We believe that this approach to quality control improves the accuracy of our results without introducing bias in the notion of review helpfulness or liking. We did not include the test reviews in any of the results below. Despite of having only negative test reviews we did not observe negative bias (i.e. workers marking all reviews as "not helpful" or "disliked") since the number of reviews each worker had access to was relatively small. For larger crowdsourcing experiments it is worth considering the impact of one-sided test reviews and introducing a more sophisticated quality control mechanism to prevent participants from cheating.

Concepts comparison

Looking at the results from the crowdsourcing experiment we see that workers marked reviews as "helpful" and "like" in similar proportions, 76% and 70% accordingly. The justifications they provided were also very similar (see table 4–6), with the highest percentage of responses indicating that the review facilitated the purchasing decision or helped the reader learn more about the book. Therefore, users are driven by similar motives when marking reviews as "helpful" and "liked".

To test whether these results hold at the review level we calculated Krippendorff's alpha, a measure of inter-coder agreement. In our experiment we consider the

Figure 4–6: The Crowdflower task coding interface as seen by a worker. The justifications are visible because the review was marked as "helpful".

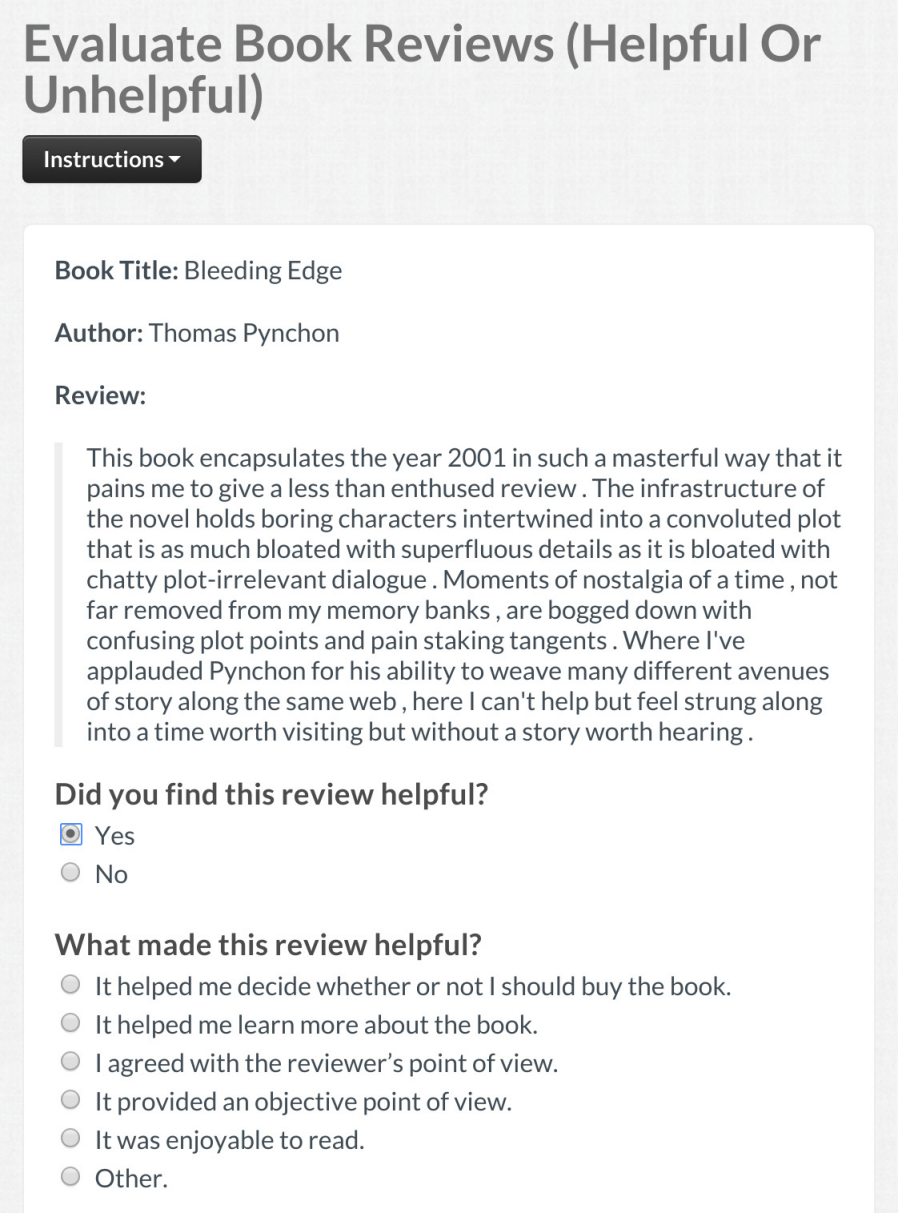


Table 4–6: Percentage of reasons given for applying each label to reviews. Each column sums to 100%.

Reasons	Helpful (Amazon)	Not helpful (Amazon)	Liked (Goodreads)	Not liked (Goodreads)
It provided an objective point of view./ It was too subjective.	14.0	23.6	12.5	25.3
It [helped me/didn't help me] decide whether or not I should buy the book.	42.3	11.0	38.2	11.5
[It helped me learn more/ I didn't learn anything significant] about the book.	38.0	51.1	38.7	43.5
It was enjoyable to read./ It was badly written.	3.6	11.4	6.6	12.9
I [agreed/didn't agree] with the reviewer's point of view.	2.0	2.0	3.7	5.2
Other.	0.1	0.9	0.3	1.6

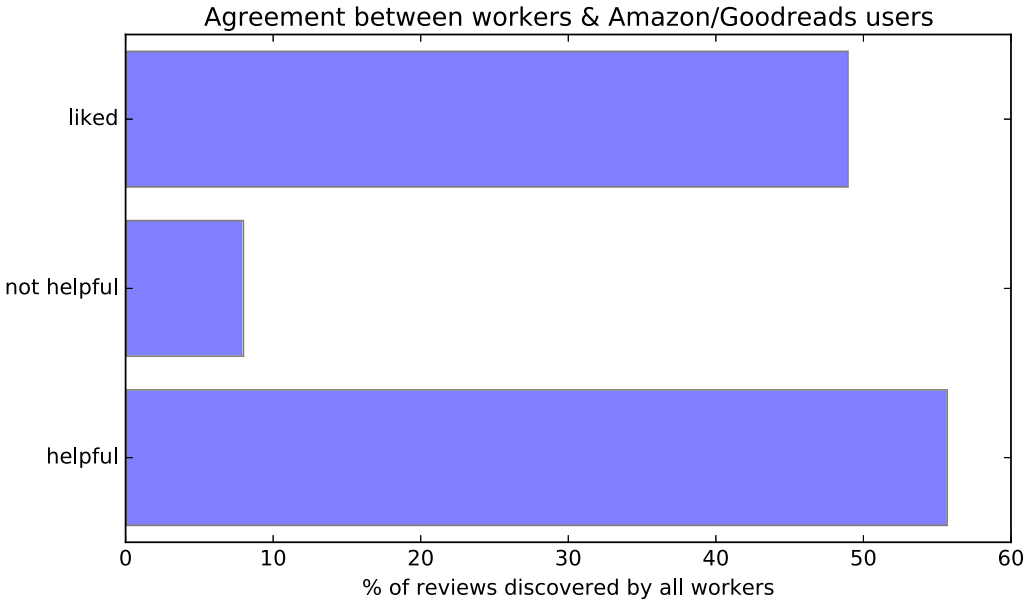
Table 4–7: Overlap between categories expressed as percentage of reviews from each category as marked by all workers.

% of reviews marked as ***Helpful*** also marked as ***Liked***	**51.5**
% of reviews marked as ***Helpful*** also marked as ***Not Liked***	**1.4**
% of reviews marked as ***Not Helpful*** also marked as ***Liked***	**12.3**
% of reviews marked as ***Not Helpful*** also marked as ***Not Liked***	**28.3**
% of reviews marked as ***Liked*** also marked as ***Helpful***	**61.6**
% of reviews marked as ***Liked*** also marked as ***Not Helpful***	**1.4**
% of reviews marked as ***Not Liked*** also marked as ***Helpful***	**11.0**
% of reviews marked as ***Not Liked*** also marked as ***Not Helpful***	**20.7**

agreement between workers who marked a review as "helpful" and those who "liked" the same review (see table 4–7). We assigned a score from 0 to 3 to every review depending on the number of workers that liked it and a score on the same scale for the number of workers who marked it as helpful. Using these scores we calculated Krippendorff's alpha of 0.22, which indicates that there is some correlation between the two concepts but it is not as strong as one may expect by looking at the overall frequencies of "likes" and "helpful" votes. This result poses the question whether workers randomly labeled reviews as "helpful"/"liked" (which we countered through test reviews as discussed earlier) or there is an intrinsic difference between the two concepts (which we did not find with the content analysis).

To test whether the difference in "helpful" and "like" votes for individual reviews is statistically significant or random we applied the Stuart-Maxwell Marginal Homogeneity Test which is an extension to McNemar's test for a $k \times k$ *matrix* of objects to categories. The test produces a chi-square statistic with k-1 degrees of freedom. For our experiment we considered the two scores ("helpful" and "like") as the two raters of the Stuart-Maxwell test. The result ($\chi^2 = 55.7$, d.o.f. $= 3$, $p < 0.001$) shows that the two concepts are highly distinct and the difference is statistically significant. Hence, the decision to mark a review as "helpful" or to "like" it is driven by different criteria even if workers provide similar justifications. Therefore, although seemingly similar, the two concepts ("helpful" and "like") can result the promotion of different types of reviews and ultimately impact the kind of content that gets generated on a given platform.

Figure 4–7: Agreement between Crowdflower workers and Amazon/Goodreads visitors expressed as % of reviews from each category unanimously labeled by all workers.



While our results are significant they may not be applicable to the two platforms (Amazon and Goodreads). Firstly, users on the platforms have the option of not marking a review as either "helpful" or "not helpful" and of not "liking" a review, while in our experiment the choice is binary. In fact, only 40% of the Amazon reviews in our dataset received at least one "helpful" or "not helpful" vote. Secondly, particularly on Amazon, we expect that users are at least partially driven by a financial incentive to identify the most helpful reviews which maximize their utility per dollar, a condition not present in our experiment. To better understand the applicability of our results to existing platforms we looked at the agreement of workers and platform users. We can't compare the reviews that received no likes on Goodreads to the ones that were disliked in our experiment because our result would

be confounded by the reviews that were simply ignored on Goodreads. Similarly, we can't compare reviews which received no helpful or unhelpful votes, or equal number of helpful and unhelpful votes, to any of our two categories ("helpful" and "not helpful"). Thus, we measure agreement between all Crowdflower participants and Amazon/Goodreads users based on "liked", "helpful" and "not helpful" reviews (see Fig. 4–7). We find that Crowdflower workers successfully discovered 48.95% of reviews with at least one "like" on Goodreads. This result by itself is not encouraging, but it should be considered in the context of the subjective nature of "liking" a review and the requirement to have all participants agree that a given review deserves a "like". On Amazon the agreement is similar for "helfpul" reviews (55.67% of reviews with more helpful than "not helpful" votes on Amazon were marked as "helpful" by all participants) but much lower for "not helpful" reviews (7.98%). Hence, we can't say that our experiment is representative of user behaviour on Amazon but we can claim that in a controlled experiment there are significant differences between the two promotion mechanisms ("helpful" and "like"). Clearly there are other forces at play on the bookseller platform which may be explained by the incentive to find the most optimal purchase or by the design of the platform and existing reviews. Understanding the impact of context on review promotion is a topic worthy of further study.

Our crowdsourcing experiment highlights how the semantic difference in the two concepts "helpful" and "like" can affect the way they are applied by users. This difference may explain the differences in promoted reviews between the two platforms.

Our content analysis experiment aimed to identify such differences between "helpful" and "liked", as well as, between "not helpful" and "not liked" reviews but our results based on the LIWC dictionaries don't demonstrate significant discrepancy in vocabulary usage. A different approach, such as the Wilcoxon rank-sum test may be more fruitful.

CHAPTER 5
Conclusion

In this study we applied computational and statistical methods to understand the differences in user generated content on two platforms: Amazon and Goodreads. By collecting and analysing two distinct datasets: all reviews for biography books and a sample of the latest 60 reviews for New York Times bestsellers, we gave our work depth and breadth which allows us to draw broader conclusions about user behaviour on the platforms.

We found that platforms influence user generated content by means of their design, the presence of existing content and their approach to content promotion and moderation. By using reviews as a proxy for the user base we can say that both platforms attract a diverse set of reviewers who generate reviews of varying length, vocabulary richness and content. Hence, given a single review it is very difficult to tell whether it was written on Amazon or Goodreads. However, as our classification experiment demonstrated, when taken as an ensemble, reviews are very indicative of the platform they were written on. We believe that this result is a close approximation of an Internet user engaging with these websites and reading a set of reviews for the same books.

Some of the conclusive results we presented show that reviewers on Amazon often give higher ratings to the same book than Goodreads reviewers. These ratings

are supported by the language used on the two platforms, with Amazon reviewers employing a more positive and commercially acceptable vocabulary, while some Goodreads users resort to profanity. Yet, the presence of literary terminology in Goodreads hints that professional reviews coexist with colloquial content on the platform. Review length is very revealing of the platform if we take an ensemble of reviews. Generally, Goodreads has some very short reviews and a few very long ones, while Amazon reviews tend to be longer on average.

Through a controlled crowdsourcing experiment we showed that the promotion mechanisms on the two platforms: "helpful"/"not helpful" and "like", are used differently by people, even if justified by similar reasoning. By analysing the way Amazon and Goodreads users apply these features we found that on Amazon reviews are both promoted and demoted, while Goodreads users are more receptive to diverse expression forms. General and subjective reviews are often demoted on Amazon and don't get likes on Goodreads, while objective content tends to be perceived positively. Through promotion mechanisms, moderation and reviewing guidelines, Goodreads and Amazon communicate to their users the kind of content that the platforms aim to attract.

The contributions of this work are then two-fold. Firstly, we showcase the tools and methodologies that can be used to perform a comprehensive comparison of user generated content on on-line platforms. Secondly, we demonstrate that two specific platforms, Amazon and Goodreads, are generally successful in influencing user behaviour and attracting substantially different reviews for the same product.

Hence, potential book buyers can benefit from reading reviews both on Amazon and on Goodreads.

References

[1] Customer review creation guidelines.

[2] AAP. *Bookstats*, volume 3. Association of American Publishers, 2013.

[3] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[4] Saeideh Bakhshi, Partha Kanuparthy, and Eric Gilbert. Demographics, weather and online reviews: A study of restaurant recommendations. In *Proceedings of the 23rd international conference on World wide web*, pages 443–454. ACM, 2014.

[5] Judith A. Chevalier and Dina Mayzlin. The effect of word of mouth on sales: online book reviews. *JMR*, 43(3), 2006. They observed that Amazon reviews are longer than Barnes and Nobles, reviewers seem to depend on content, not summaries; lower ratings carry more weight than higher ratings.

[6] Nicholas Diakopoulos and Mor Naaman. Towards quality discourse in online news comments. In *CSCW*, pages 133–142, 2011.

[7] Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752, 2011.

[8] Wenjing Duan, Bin Gu, and Andrew B Whinston. The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of retailing*, 84(2):233–242, 2008.

[9] Marisa Vasconcelos et al. Popularity dynamics of foursquare micro-reviews. In *Proceedings of COSN*, 2014.

[10] A Flood. Amazon purchase of goodreads stuns book industry. *The Guardian*, 2, 2013.

[11] Chris Forman, Anindya Ghose, and Batia Wiesenfeld. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, 19(3):291–313, 2008.

[12] Eric Gilbert and Karrie Karahalios. Understanding deja reviewers. In *CSCW*, 2010.

[13] Goodreads. Goodreads *about us* webpage, September 2015.

[14] Marco Guerini, Lorenzo Gatti, and Marco Turchi. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*, 2013.

[15] Nan Hu, Noi Sian Jok, and Srinivas K. Reddy. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, 57, 2014.

[16] IFPI. *Recording Industry in Numbers*. IFPI, 2014.

[17] Henry Jenkins. *Fans, Bloggers, Gamers: Exploring Participatory Culture*. NYU Press, 2006.

[18] Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, 2007.

[19] Marco Lui and Timothy Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, 2012.

[20] Lionel Martin and Pearl Pu. Prediction of helpful reviews using emotions extraction. In *Proceedings of AAAI*, 2014. They find that influential users tend to use more affective words.

[21] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of RecSys*, 2013.

[22] Loizos Michael and Jahna Otterbacher. Write like i write: Herding in the language of online reviews. In *ICWSM*, 2014.

[23] Sarah G Moore. Attitude predictability and helpfulness in online reviews: the role of explained actions and reactions. *Journal of Consumer Research*, 42(1):30–44, 2015.

[24] Susan M Mudambi and David Schuff. What makes a helpful review? a study of customer reviews on amazon. com. *MIS quarterly*, 34(1):185–200, 2010.

[25] Hrvoje Niksic. Gnu wget, 1998.

[26] Jahna Otterbacher. 'helpfulness' in online communities: a measure of message quality. In *CHI*, 2009.

[27] Do-Hyung Park, Jumin Lee, and Ingoo Han. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4):125–148, 2007.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[29] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[30] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577, 2003.

[31] United States. Securities and Exchange Commission. *Form S-1: Registration Statement Under the Securities Act of 1933 AMAZON.COM, INC.* Securities and Exchange Commission, 1997.

[32] Beverley A. Sparks and Victoria Browning. The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, 32(6), 2011.

[33] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. Normative influences on thoughtful online participation. In *CHI*, 2011.

[34] Frank Wilcoxon, SK Katti, and Roberta A Wilcox. Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1:171–259, 1970.

[35] Qiang Ye, Rob Law, Bin Gu, and Wei Chen. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Computers in Human Behavior*, 27(2), 2011.