UNDERSTANDING BIAS AND HELPFULNESS IN ONLINE REVIEWS

A Thesis

by

SIDDHARTH VERMA

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Dr. James Caverlee
Committee Members,	Dr. Xia Hu
	Dr. Eric Ragan
Head of Department,	Dr. Dilma Da Silva

December 2018

Major Subject: Computer Engineering

ABSTRACT

In today‘s era of easy access to information, online consumers have become more informed in their decision making about the products they would like to buy. Online product reviews have played a key role in the increase in consumer awareness and online research activities about products. Due to the vast number of product reviews (in thousands) for each item, it becomes cumbersome to makes sense of all the information and form a perspective or develop a sentiment about the product. In order to tackle this problem, large websites such as Amazon provide a helpfulness score along with each review, to help uninformed consumers get an idea of the authenticity, quality and perspective of a particular review, which are written by consumers themselves having experience in purchasing or using that product.

We aim to study reviews from the Amazon product review dataset and understand how various review attributes influence the review helpfulness score as well as, how this influence varies across diverse product categories. For this purpose, we will look at key statistical features from the star-ratings as well as context based features extracted from the reviews. As an addition to our existing task, we will also discuss possible origins of biases in the system and look at model building approaches that can reduce the effect of intrinsic biases in a particular product's review helpfulness voting activity. This research will contribute significantly towards understanding the characteristics of helpful product reviews across different categories and lay foundations for future methods for preventing biased helpfulness voting on online product review platforms.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. James Caverlee and Dr. Xia Hu of the Department of Computer Science and Engineering at Texas A&M University and Dr. Eric Ragan of the Department of Visualization at Texas A&M University (now at the Department of Computer & Information Science & Engineering at University of Florida). The dataset used in the thesis work was provided by Dr. Julian McAuley of the Department of Computer Science, University of California at San Diego as a publicly available contribution. All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by funding from part-time student worker positions and Teaching assistant fellowship at Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

The ubiquity of e-commerce retail has brought about a paradigm shift in the consumer shopping experience. Over the last two decades, the boom of the .com revolution has percolated into most, if not all aspects of our lives; day-to-day shopping being no exception [9].

As e-commerce businesses are becoming the new norm for consumer shopping, there is another evolving trend that is not widely discussed – consumer decision making. Consumers today have a truly exceptional medium to communicate ('instant' in terms of speed and 'global' in terms of scale), as well as terabytes of product data to explore, analyze and make informed decisions on their purchases; something which was inconceivable two decades ago. The modern day shopper has become more knowledgeable and selective about buying products than his historical counterpart, and reads product reviews more frequently [12]. This selectivity comes from the ability to leverage data in the form of product experiences/feedback by other consumers, enabling a modern-day shopper to make informed decisions on purchasing products.

Product reviews are key contributors to this skeptic and research-driven behaviour of online shoppers. Reviews are essentially feedback, opinions and/or summaries of other shoppers that have had experiences purchasing or using a particular product. Why are product reviews useful? Simply because they add a layer of transparency to a product's attributes – such as utility, quality/durability, or performance. This feedback is particularly useful to prospective buyers that are deciding on purchasing that product. Moreover, many product reviews also provide feedback on the service platform (such as Amazon delivery experience) or on the sellers (reflecting the company brand). Such form of information albeit highly opinionated, is useful

to online shoppers that may be looking to buy a product in that category for the first time. Finally, product reviews are public which allows the entire marketplace community to view this information.

Figure 1.1: Product review example on Amazon

1.1 Review Helpfulness

Apart from being helpful to online shoppers, product reviews are also abundant in volume and can be quite overwhelming to go through exhaustively. To alleviate this issue of information overload, e-commerce retail companies like Amazon have come up with a scoring methodology which provides a validity metric on the review itself: a review "helpfulness voting" system. This allows users to sort product reviews by their helpfulness scores allowing them to read a few of the most-helpful reviews and quickly make decisions, thereby improving the overall shopping experience.

When users read a review, they have the option of providing their feedback as to whether they found the review to be helpful or not. The previous score is also visible to a reader. Figure 1.1 shows a typical Amazon product review with the feedback question shown at the bottom, and the current helpfulness score shown at the top.

The helpfulness score indicates how many people found a particular review helpful as opposed to not helpful. This voting methodology adds an additional layer of quality control and enables the community to act in a self-regulatory way; reviews by the community are being voted on as "helpful" or "not helpful" by the community itself.

1.2 Bias in Review Helpfulness Voting

Since product reviews are an important factor in consumers' decision making on product purchases, they also impact product sales and the overall seller revenue/profit. Therefore, manipulation of product reviews becomes an easy route for dishonest agents, who try to alter certain review attributes to drive their sales. Such malpractice is prevalent on e-commerce review platforms. A few years ago, it was "fake reviews" by bots/fake accounts [32], which has been alleviated to a large extent on sites like Amazon, and are continually being eradicated [31] [26].

However, manipulation of review helpfulness votes has emerged as a new challenge to e-commerce platforms [15]. Bad actors are able to pose as "interested buyers" and generate a bias in helpfulness voting. To further exacerbate this problem, there are 3^{rd}-party vendors that offer SEO-like services to boost upvotes and suppress downvotes using external agents [1] [27]. They can either selectively up-vote critical/negative reviews of their competitors as "helpful" or over-the-top positive reviews of their own products. Such externally generated biases can cause significant shifts in review rankings, influence consumer decisions and affect overall product sales [5]. Figure 1.2 shows various instances of helpfulness voting manipulation, including Amazon seller forums.

Another source of bias on online review platforms is of a more intrinsic nature - users' preferences [30]. Different consumers that read reviews find different aspects of the review helpful. For example, a review on an electronic product like headphones

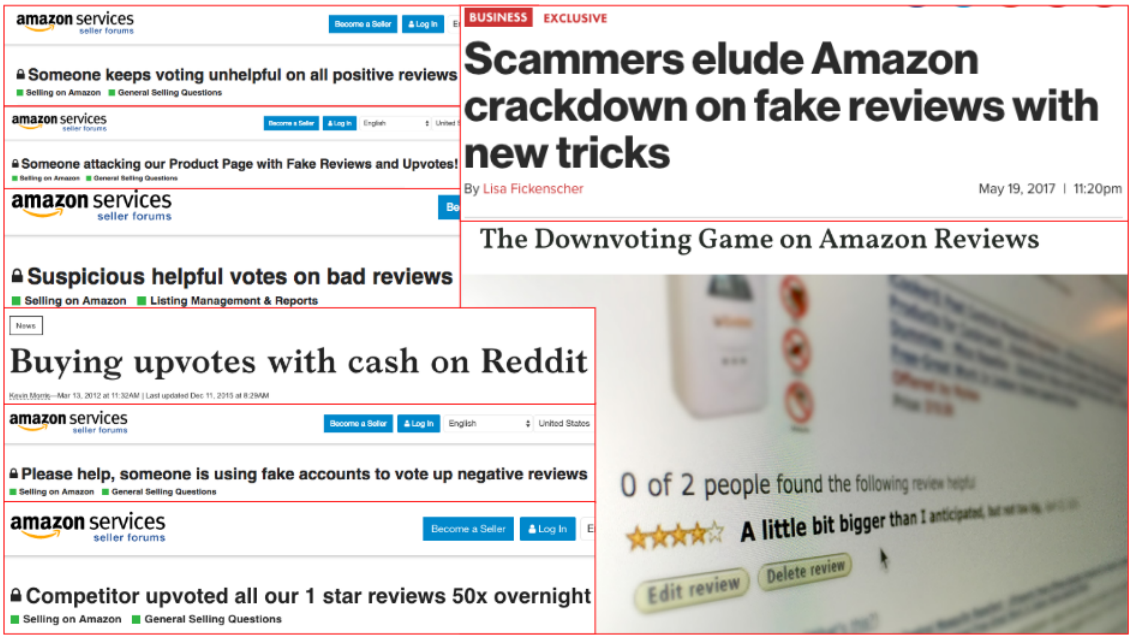


Figure 1.2: Various articles on manipulation of Amazon review helpfulness votes

may say "used it for 3 days. great product. love it!". This review can be perceived as helpful by some consumers since the reviewer expresses his/her opinion on the product and provides sentiment information as well. But it is not surprising if another consumer does not find it helpful, possibly due to the lack of a more detailed analysis (e.g., a clear comparison of pros versus cons). Such forms of biases can affect product sales and create an unfair marketplace for consumers, as well as sellers.

1.3 Research Goals

This thesis research aims to understand the constituents of helpful reviews, analyze sources of bias in online reviews and explore machine learning based models that can compensate for this bias in review helpfulness voting. We evaluate our models using large-scale crowd-based feedback and conduct a comparative analysis across different product categories.

Concretely, we focus on two primary tasks – (i) analyzing factors affecting review

helpfulness; and (ii) building models to prevent bias in helpfulness voting of product review. We discuss these tasks in detail below.

1. Analyzing factors influencing review helpfulness.

We first aim to analyze various review attributes and understand how they influence product review helpfulness. More specifically, we seek to explore review features from different angles – statistical, semantic, and contextual, and understand how each feature's influence varies across different product categories. Helpfulness score prediction has been an integral part of review ranking and recommendation tasks and has been studied extensively [13] [19] [18] [22] [16] [25].

However, factors that determine review helpfulness for a particular product cannot be generalized across all product categories. Therefore it is important to understand how the influence of various review attributes vary by category time. [22] conducted experiments on 6 products within a single category (Electronics) in total. [25] conducted a more comprehensive analysis using 5 different categories with 41,850 product reviews for the task of helpfulness score prediction. In comparison to previous literature, our work is significantly more extensive as we conduct our analysis across 6 product categories and 3 product items within each category.

2. Building models to mitigate bias in helpfulness voting.

In the second part of this thesis, we introduce the notion of bias in helpfulness voting activity, and discuss possible intrinsic sources of bias on review platforms. We further extend our discussion by proposing preliminary machine learning techniques to generate review similarity clusters which are based on various aspects of a product review. The goal of this approach is to prototype methods for mitigating bias in helpfulness voting by promoting under-seen reviews as potentially helpful candidate reviews.

1.4 Challenges

In this section, we discuss the key challenges encountered during our research. We also elaborate on how we addressed each challenge while ensuring that our approach remained most suitable to address our goals.

Lack of Ground Truth.

The bane of any machine learning task is having data but no labels (or class information). Since supervised learning requires annotations, labelled data makes training easier, effective and also allows for varied experimentation with different models. Model evaluation and model comparison becomes a well-defined task and the boundaries for the state-of-the-art approaches can be pushed further.

In contrast to the above, unlabelled data not only makes it difficult to evaluate model performance, but even preliminary analysis for feature selection becomes a challenge as there is no response to compare feature importance/correlation against. Having a similar unlabelled dataset, we had two possible approaches to consider

- **Create labels for data:** Converting an unsupervised task into a supervised learning problem using data annotation methods can overcome the problem of model/feature selection. However, in our case, it is quite difficult to quantify the extent to which a given product review may be biased; review bias can only be detected in comparison to a control variable or across a certain feature (such as time).

- **Unsupervised learning:** Approaches such as clustering, generative modeling can be used to find patterns in the data, which may yield better neighborhoods for recommendation due to high similarity in latent space.

In order to evaluate our models in the absence of labels, we generated crowd-based

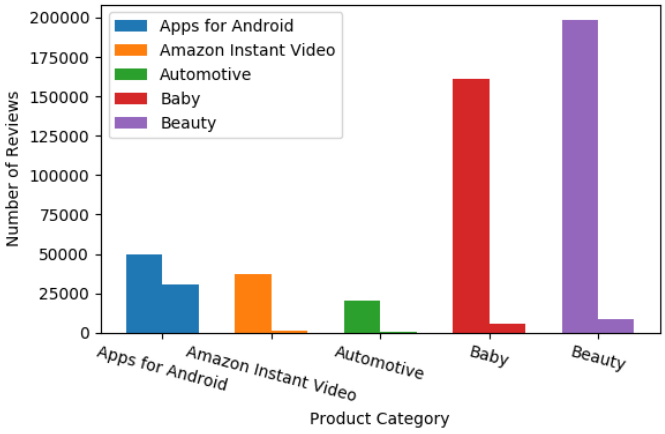
annotations from surveys on Amazon's Mechanical Turk platform. We conducted separate surveys for each model's generated recommendation of similar style reviews for a given product review and compared the results in terms of reduction in crowd bias towards older reviews.

Sparsity in user-item matrix.

In order for us to perform our analysis on all product categories listed in the Amazon product review dataset [20], we required that category have a large review corpus that aligned with our requirements.

Books	5-core (8,898,041 reviews)
Electronics	5-core (1,689,188 reviews)
Movies and TV	5-core (1,697,533 reviews)
CDs and Vinyl	5-core (1,097,592 reviews)
Clothing, Shoes and Jewelry	5-core (278,677 reviews)
Home and Kitchen	5-core (551,682 reviews)
Kindle Store	5-core (982,619 reviews)
Sports and Outdoors	5-core (296,337 reviews)
Cell Phones and Accessories	5-core (194,439 reviews)
Health and Personal Care	5-core (346,355 reviews)
Toys and Games	5-core (167,597 reviews)
Video Games	5-core (231,780 reviews)
Tools and Home Improvement	5-core (134,476 reviews)
Beauty	5-core (198,502 reviews)
Apps for Android	5-core (752,937 reviews)
Office Products	5-core (53,258 reviews)
Pet Supplies	5-core (157,836 reviews)
Automotive	5-core (20,473 reviews)
Grocery and Gourmet Food	5-core (151,254 reviews)
Patio, Lawn and Garden	5-core (13,272 reviews)
Baby	5-core (160,792 reviews)
Digital Music	5-core (64,706 reviews)
Musical Instruments	5-core (10,261 reviews)
Amazon Instant Video	5-core (37,126 reviews)

(a) Corpus Sizes



(b) No. of reviews

Figure 1.3: (a) Number of reviews across product categories (b) Review count after filtering for helpful votes (Left Bar: Total, Right Bar: >= 10 helpfulness votes)

Figure 1.3 (a) shows the skewed count of review data collected for each product category. Even in a 5-core matrix setting, the corpus varies from 10k to 9 million reviews depending on the category selected. Figure 1.3 (b) shows the drastic decrease

in review count when categories were filtered for reviews containing at least >= 10 helpfulness votes. Therefore, it became unsuitable to conduct our experiments on sparsely populated categories. In order to explore accumulative bias, we wanted categories where reviews have large temporal variance i.e. contain a mix of old and recently posted reviews. Additionally, we also wanted categories where a large number of reviews were available, so we could obtain a good set of similar review candidates by applying unsupervised clustering methods.

Finally, we also wanted to maintain diversity in categories in terms of their utility. Categories such as *Books*, *Movies & TV*, *Video Games* are experience goods as they need to be experienced by a user to be consumed and can their usefulness is highly subjective. In contrast, *Electronics*, *Home & Kitchen* and *Health & Personal Care* are categories with more functional dimensions and usually have technical specifications which are not subject to opinion. Having both types of categories would help us investigate the variation in the influence of factors determining review helpfulness.

Absence of temporal history.

In order to find bias in review helpfulness, it is worth exploring to track the deviation of review helpfulness score as a moving average over time. Based on the above time series, we could draw comparison between a 5 year intervals for reviews in two different era's (2005-2010 and 2010-2015). It would then be a matter of comparing slope coefficients for both series and detecting statistically significant differences. However, we do not possess records of the evolution of product reviews over time. Instead, we have a snapshot of the state of product reviews at a given time i.e. when the data was collected. This makes detecting temporal bias complicated as we do not have historical data for any review under consideration.

We can however, get an understanding of the underlying bias by using a block

design in our experiments. We categorize the review helpfulness score distribution into 3 blocks – low, medium and high and look at the proportional distribution of review age within each block, to check for presence of bias.

1.5 Outline of the Thesis

The rest of this thesis is organized into 5 sections. Section 2 talks about the related literature on review helpfulness and factors that influence helpfulness scores, as well as works on analyzing biases in voting scores. Section 3 discusses our approach on data selection, analyzing features and their importance to helpfulness scores across diverse product categories. Section 4 introduces the notion of biases in review helpfulness voting, and explores methods to detect two sources of intrinsic biases - Accumulative and Opinion biases. Section 5 focuses on different modelling approaches to mitigating voting bias and their evaluation using various methods. Finally, our research conclusions, limitation and scope for future work are discussed in Section 6.

2. RELATED WORK

This section focused on understanding related research literature on review helpfulness and bias in review helpfulness voting. There is a good amount of literature on the task of review helpfulness prediction. We will discuss some of their approaches in the forthcoming sections and how our work differs in context to existing approaches.

2.1 Analyzing Review Helpfulness

Although determining whether a review was helpful or not seems to be a subjective opinion of an online shopper, research has shown there there are deeper underlying structures in the review information that govern consumers' evaluation opinions. For example, Danescu et al. [8] analyzed review helpfulness by constructing various hypotheses based on statistical features such as rating deviation from mean, direction of deviation, and deviation by variance of ratings. However, their focus was entirely on the rating aspect of the review and not on the review text. Other works such as [13] include various feature classes (structural, lexical, syntactic etc.) to rank reviews based on helpfulness score. Chen et al. [6] conducted empirical based studies to verify different hypotheses by investigating correlation between review features and helpfulness scores. All of the above endeavors focus on either a small set of products or a single product category. Alternatively, we analyze a set of highly reviewed products across various categories, and aim to explore the change in feature relationships with helpfulness scores across different categories. Additionally we explore possible sources of bias in helpfulness voting, and instead of predicting helpfulness, we propose models that may remove biases.

2.1.1 Selecting Relevant Features

A lot of literature has focused on the task of predicting helpfulness scores as a means to another goal – either to provide better review recommendation or identify underlying causes that make a good review/reviewer. Though many of these efforts vary in their approach to model review helpfulness, they follow a similar methodology in performing some analysis or hypothesis testing for selecting features.

Previous studies have explored various aspects of review attributes that can be considered factors influencing review helpfulness. We discuss some of the common attributes in existing literature in the upcoming sections.

Rating statistics. Studies conducted by [8] [14] [23] show that statistical aspects of product rating information are strong contributors in determining review helpfulness. Danescu et al. [8] used rating deviation from mean to show that there exists a directional bias i.e. reviews having a positive rating deviation from mean rating have higher helpfulness than those having a negative deviation. Korfiatis et al. [14] verified a conformity hypothesis to show that reviews having ratings closer to the mean should have higher helpfulness score that those away from the mean. Otterbacher et al. [23] hypothesized that consumers who have "extreme" opinions on products are more likely to exhibit their emotions, and use strong words, which can have a correlation with helpfulness.

Review text-based features. Most studies investigate text-based review attributes and explore the correlation with helpfulness. Previous works like [23] [16] [14] suggested review readability as a useful feature since, consumers perceive them as high quality content and are therefore, likely to up-vote well written reviews as helpful.

Other text based statistics such as number of upper case words, punctuation

marks (!, ?) etc. indicate strong emphasis of opinion and can influence readers compared to reviews with a more monotonic style. Kim et al. [13] explore syntactic features such as Parts-of-speech (POS) distribution of nouns, verbs and adjectives in a review. Lee et al. [18] used proportion of small (single character), medium (2-9 characters) and large (> 10 characters) length words in a review as input to their neural model.

Some studies found that emotion-based attributes (sentiment, positively/negatively structured sentences) also contribute in predicting helpfulness. Connors et al. [7] found that reviews that contain both positive and negative aspects (such as pro/con listings) are generally found to be more helpful that uni-directional oriented reviews. Kim et al. [13] used semantic features from words like "amazing" and 'weak' for describing sentiment in their approach.

The extent to which a review is factual vs opinionated has been deemed useful in some literature. More specifically, the degree of review subjectivity or objectivity has been o of particular interest in some works. Krishnamoorthy et al. [16] utilized subjectivity by calculating the proportion of opinion based words (positive or negative) in a review. Otterbacher et al. [23] explored the effect of objectivity as a feature based on the formulation that highly objective reviews have strong similarity with the product description.

Another feature that has been commonly mentioned is the timeliness of reviews, i.e. temporal information of a review. Krishnamoorthy et al. [16] utilizes review data information based on past research studies that show its relation to helpfulness score. Otterbacher et al. [23] considered the earliest time a review for a product was posted based on the argument that older reviews tend to have less number of votes.

Review metadata. In addition to the aforementioned features, studies have also

investigate some uncommon meta-data such as rating valance [6], product prices [18], reviewer-reader similarity [7] and reviewer innovativeness [24], to name a few.

2.1.2 Modelling Review Helpfulness

Several studies have explored models for a prediction task on review helpfulness. We explore some of the commonly used approaches in these works

Regression/ Hypothesis testing based approaches. Studies that focus on factors determining review helpfulness have explored review features and built hypotheses. In order to validate their propositions, some works have performed correlation tests on their hypothesis [10] [14] [13] [22] [23] [24] using regression models and interpreted their p-value scores as indicators for the features' statistical significance.

Neural network based approaches. Lee et al. [18] selected 20 features from review characteristics as input to their backpropagation multilayer perceptron neural network. They trained a 3-layer shallow network for 100 epochs and validated performance using k-fold cross validation.

Other Predictive Models. Krishnamoorthy et al. [16] compared predictive performance using Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest (RandF) models. [33] used a SVM Regressor with an RBF based kernel while Park et al. [25] used SVM Regressor, M5P (Decision Tree) and RandF for the prediction task.

Crowd-sourced methods. Connors et al. [7] used a survey-based approach with a sample size of 40 business undergraduate students and 20 reviews to understand constituents of helpfulness in a review. Wan et al. [30] used 80 randomly selected graduate students and received a total of 74 valid responses.

2.2 Bias in Review Helpfulness Voting

Since bias in review helpfulness is a relatively new problem, research literature is scarce. Studies such as [2] have explored bias in online reviews focused towards comparing web vs email-prompted reviews. In the online e-commerce setting, Sipos et al. [29] show that consumers vote a review as "helpful" or "not helpful" based on a certain context; for example, voting activity correlates with the amount of ranking deviation of the review from its "true" rank.

3. ANALYSIS OF FACTORS AFFECTING REVIEW HELPFULNESS

In this section, we will focus on the approach used for analyzing bias in helpfulness votes and factors that influence helpfulness voting.

3.1 Dataset selection

We decided to conduct our experiments on the Amazon product review dataset McAuley et al. [20] as it was categorically diverse, contained large review corpus sizes and readily available metadata information (timestamp, helpful voting score, productID) about each review. Additionally there exists a product metadata corpus for each product category which contains more features [20]. The dataset contains more than 140 million product review across 24 different product categories and spans reviews from May 1996 - July 2014. Each review in the dataset is stored in a JSON-dictionary format as shown below in Figure 3.1.


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Figure 3.1: Sample Review JSON data

The JSON key definitions are provided below [20]

- **reviewerID** : Encrypted user ID for the reviewer (uid)
- **asin** : Encrypted product ID (pid)
- **reviewerName** : User name
- **helpful** : A list [x, y] where

  x = number of users who voted review as 'helpful'

  y = total number of users who voted on the review
- **reviewText** : The review in text format (r_{text})
- **overall** : The product rating ($rating_{pid}$) where

  $rating_{pid} \in [0, 5]$ and $rating_{pid} \in \mathbb{Z}_{>0}$
- **summary** : The title of the review (r_{title})
- **unixReviewTime** : The time review was posted (unix time format)
- **reviewTime**: The time the review was posted (raw format)

Due to the high sparsity in the user-item map, we decided to use a 5-core dense subset of the dataset, where at least 5 products have been reviewed by a user, and at least 5 reviews are available for each product. This ensured that we have enough review content for each product.

For our research, we initially performed analysis on all 24 product categories. However, our requirements of selecting older reviews with high helpfulness scores and recent reviews with lower number of votes introduced an additional amount of sparsity in the categories. Thus we decided to reduce our category set to only those

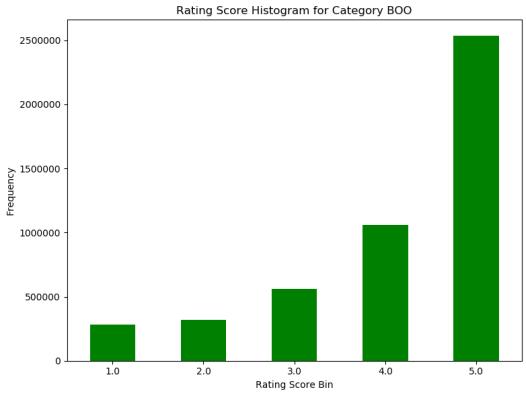
categories having a significant share of review data. We finally performed our analysis on 6 product categories namely - Books, Movies & TV, Video Games, Electronics, Health Personal Care, Home & Kitchen to keep a mix of experiential and functional based goods.

Review Distribution and Analysis In order to better understand how our data is distributed, we explore feature distributions across various product categories in greater detail. We are interested in understanding the global feature distribution and draw comparisons between them across categories. Figure 3.2 shows the star rating distribution for each of the 6 categories. The rating distribution in the data is slightly bipolar, with larger review frequency at extreme scores. This can be ascribed to the fact that majority of the reviewers have a tendency to leave an "extreme" product rating based on whether they were satisfied/unsatisfied with the purchased product.

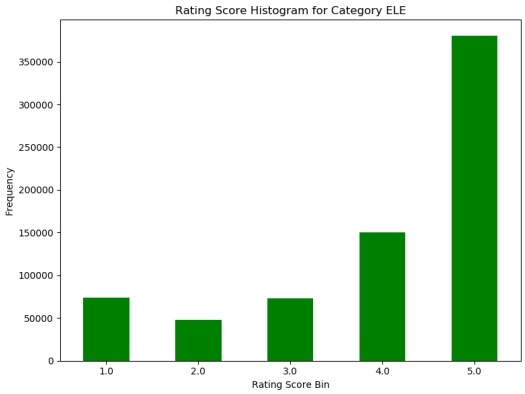
To remove voting scores that can contain significant bias, we filtered out reviews that had less than 10 number of votes in total. Figure 3.3 shows the distribution of review helpfulness score across different categories. The plot contains a similar pattern across all categories - left skewed with larger proportion of reviews containing high helpful scores. However, an interesting observation here is for experiential goods (refer to Figures 3.3(a), 3.3(c) and 3.3(e), the distribution has heavier tail compared to that for functional goods (refer to Figures 3.3(b), 3.3(d) and 3.3(f)).

3.2 Factors Influencing Review Helpfulness

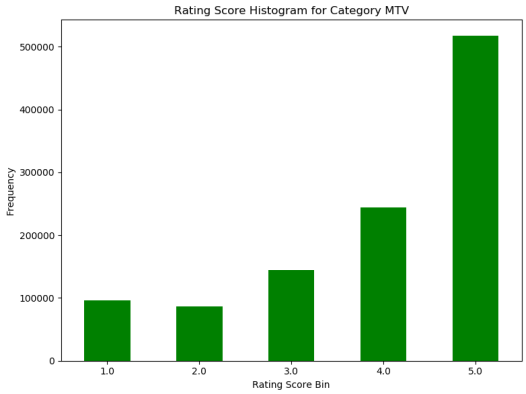
We now explore various review attributes and how they influence review helpfulness scores. Before we being our analysis, it is important to define what review helpfulness means. Since we have the review voting information available, we define



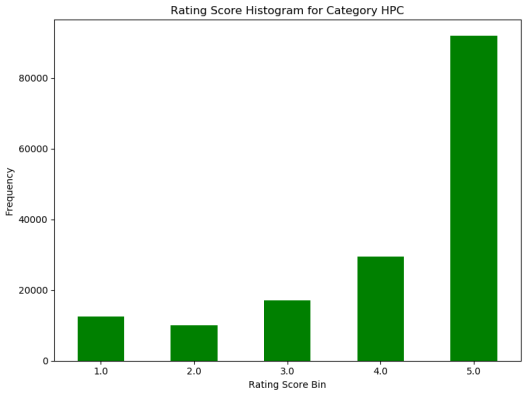
(a) Books



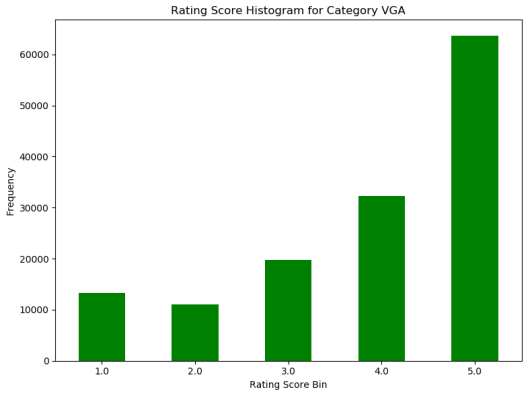
(b) Electronics



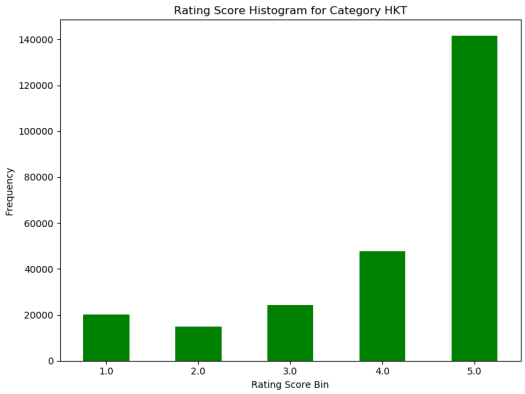
(c) Music & TV



(d) Health & Personal Care

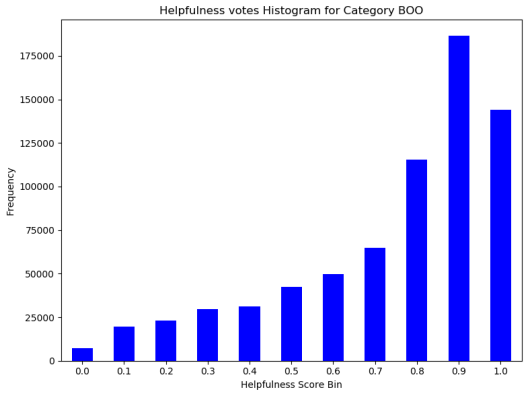


(e) Video Games

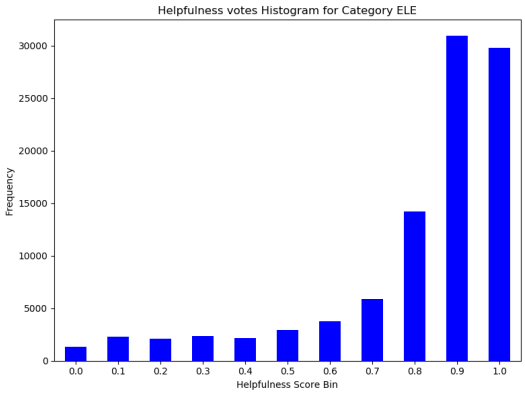


(f) Home & Kitchen

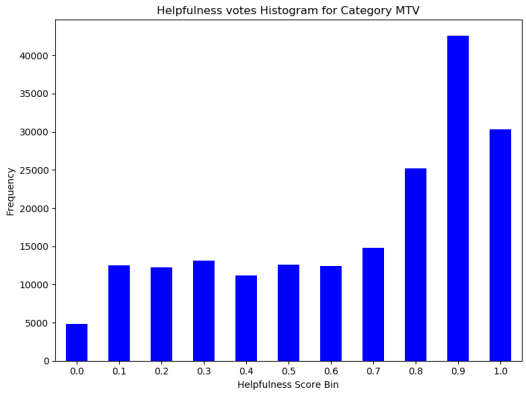
Figure 3.2: Review star rating distribution across categories



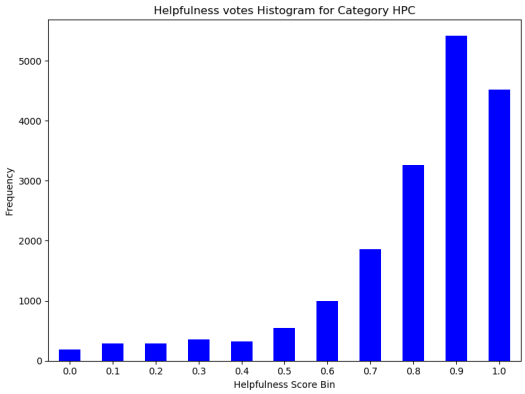
(a) Books



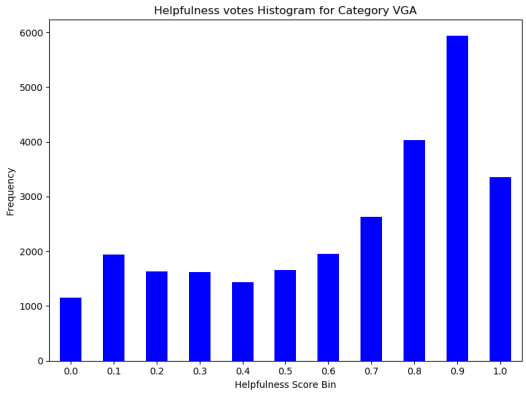
(b) Electronics



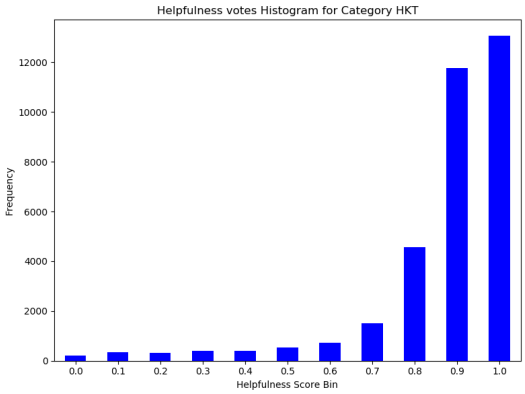
(c) Music & TV



(d) Health & Personal Care



(e) Video Games



(f) Home & Kitchen

Figure 3.3: Review helpfulness score distribution across categories

the review helpfulness ratio r_{HR} as

$$r^p_{hr} = \frac{\text{Number of “helpful” votes}}{\text{Total number of votes}} = \frac{x}{y} \tag{3.1}$$

where x and y are obtained from our the 'helpful' key in our JSON dictionary data.

3.2.1 Rating based features

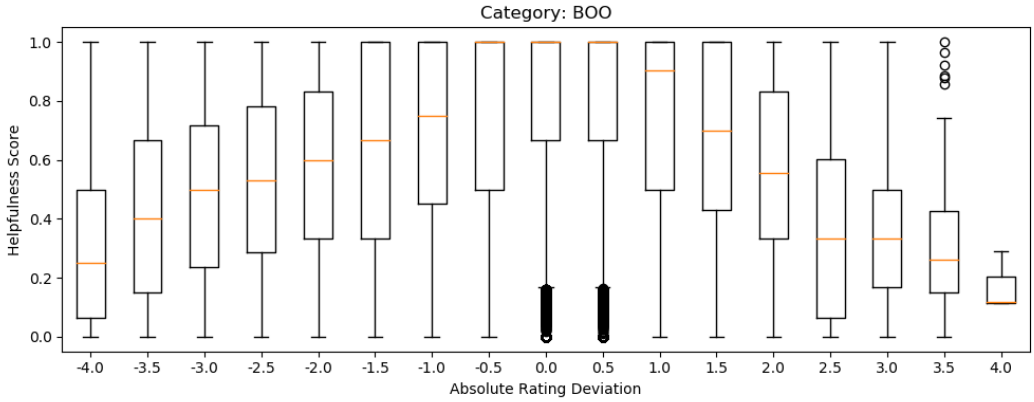
Rating score for a product has been shown to provide useful information about review helpfulness [8]. We compute the mean rating deviation for each category and plot it against the review helpfulness score.

Mean Rating Deviation

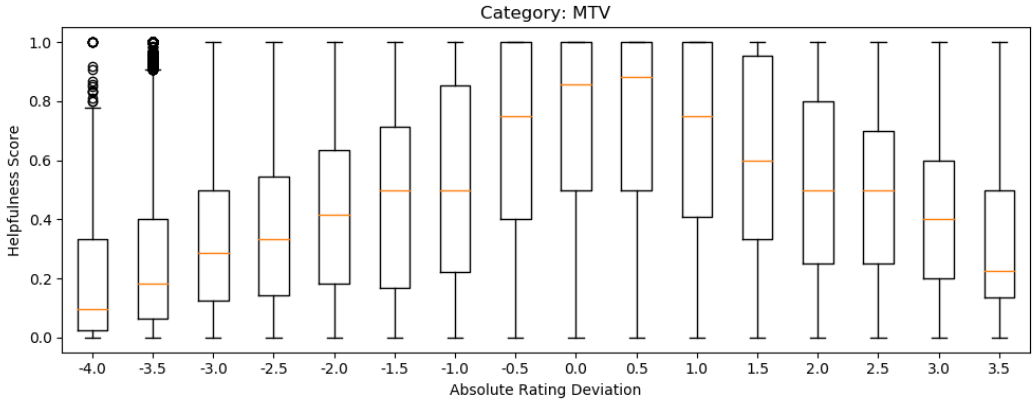
For a given review r on a product p with a rating score r^p_{score} we define the mean rating deviation r^p_{mrd} as

$$r^p_{mrd} = r^p_{score} - \frac{\sum_i^K (r^p_{score})_i}{K} \tag{3.2}$$

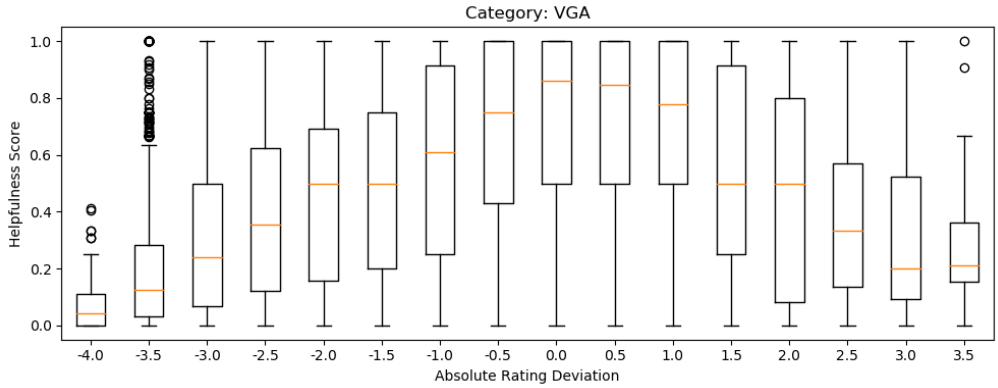
where K is the set of reviews for product p. For the product categories under consideration, as shown in Figures 3.4 and 3.5, we can see that the effect of rating deviation on helpfulness ratio follows a 'bell-shaped' curve for the median values with the maximum helpfulness ration centered around the mean rating i.e (deviation = 0). This is in concordance with the conformity hypothesis in previous literature that a review is more helpful when its rating is close to the mean rating across all reviews for that product [8]. Another interesting observation here is that the distribution of helpfulness score has comparatively heavier tails for experiential goods and lighter tails (with larger number of outliers) for functional goods. This indicates stronger coherence of helpfulness scores in functional goods, when the review rating is close to the mean rating for that category.



(a) Books

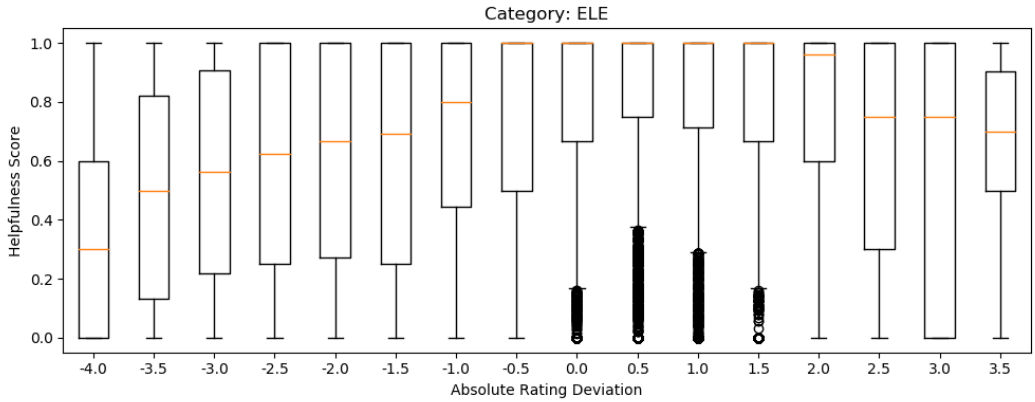


(b) Music & TV

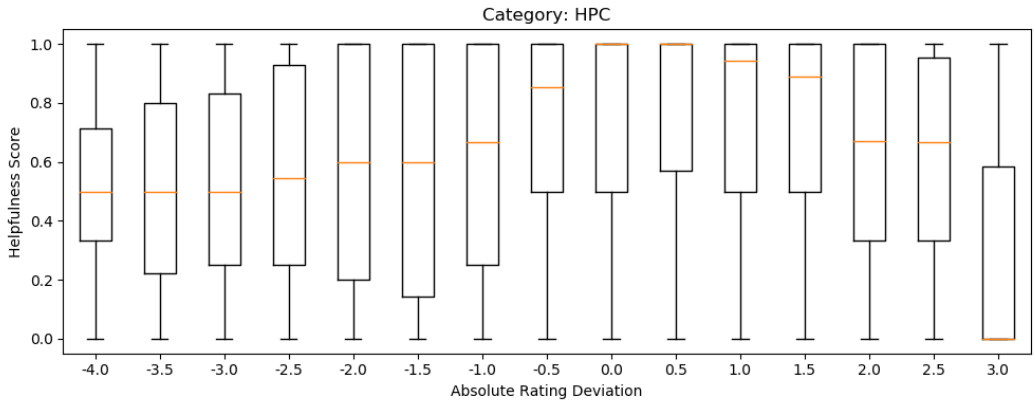


(c) Video Games

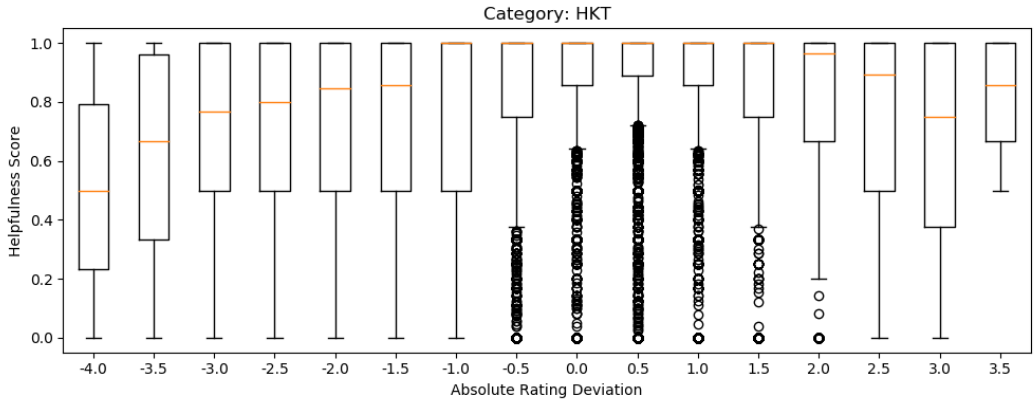
Figure 3.4: Mean rating deviation vs helpfulness score for experiential Goods



(a) Electronics



(b) Health & Personal Care



(c) Home & Kitchen

Figure 3.5: Mean rating deviation vs helpfulness score for functional goods

3.2.2 Text Statistics based features

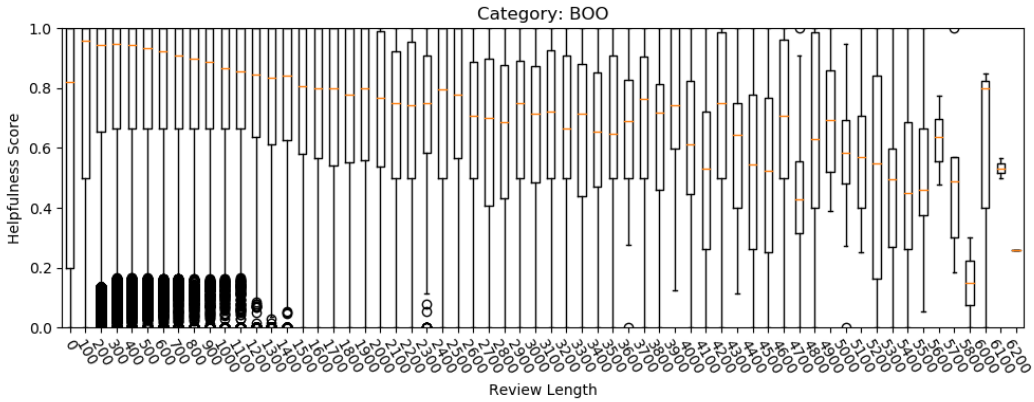
Research has shown that review features such as length, and other statistics (upper case words, punctuation marks etc.) play a role in the consumers' voting behavior on review helpfulness [18] [23] [22] [24]. In our work, we consider the following text based features.

Review length

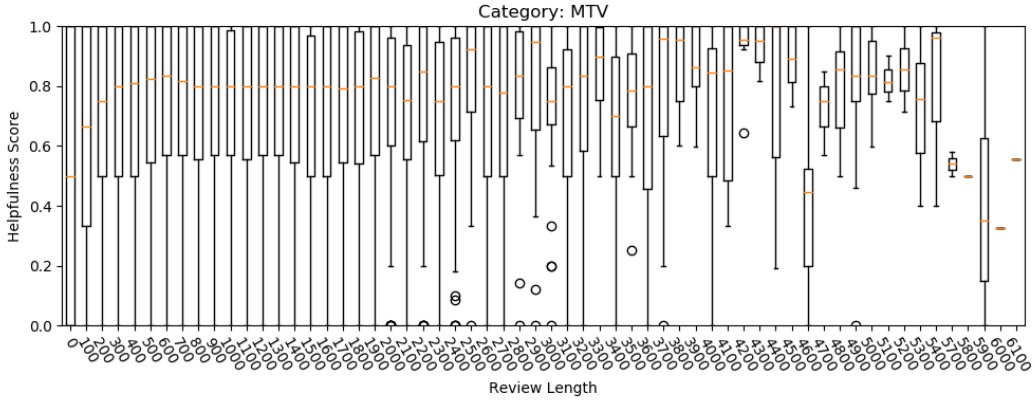
Review length has a non-linear relationship with review helpfulness as shown in Figures 3.6 and 3.7 below. Moreover, the non-linearity is consistent across all product categories, which indicates that longer reviews are in general, considered more helpful than short reviews. From these figures, we see that helpfulness scores for reviews that are less than 100-200 words in length are significantly lower. However, beyond the optimum review length, reviews that are too verbose tend to have lower helpfulness scores as well (although the rate varies based on product category).

Number of Upper Case words

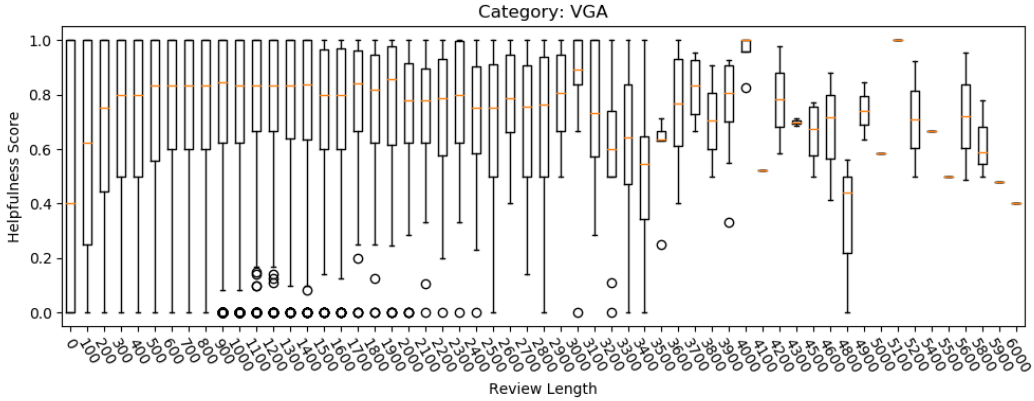
Reviewers use upper case words for emphasis in both positive and negative contexts. Therefore it is of interest to see if how the upper case word count affects review helpfulness scores. Figures 3.8, 3.9 show these distributions across categories. From the plots we can observe that for experiential goods, the helpfulness score is maximum around a count of 20-30 words, and then begins to drop as the upper case word count increases. However, for functional goods, the helpfulness score is comparatively higher for larger counts as well, and slowly decreases with increase in upper case word count.



(a) Books

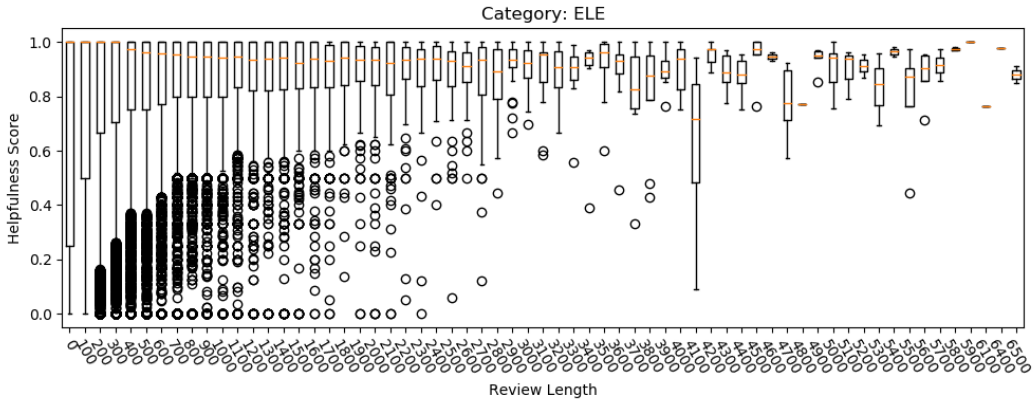


(b) Music & TV

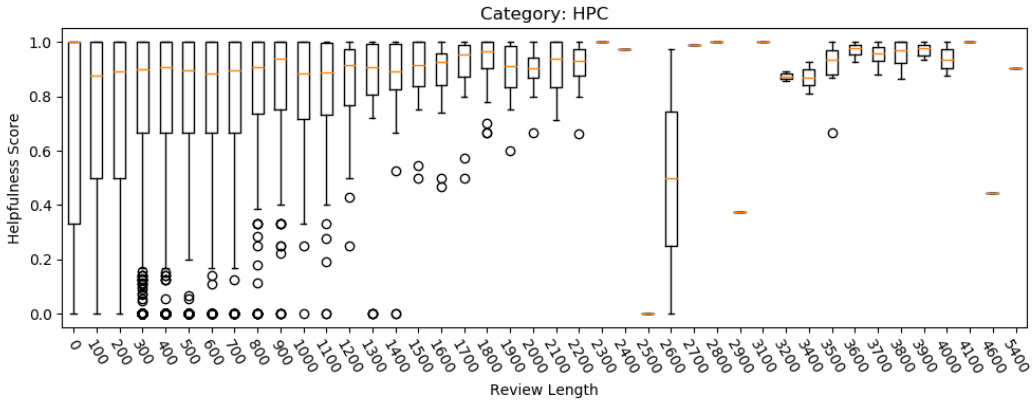


(c) Video Games

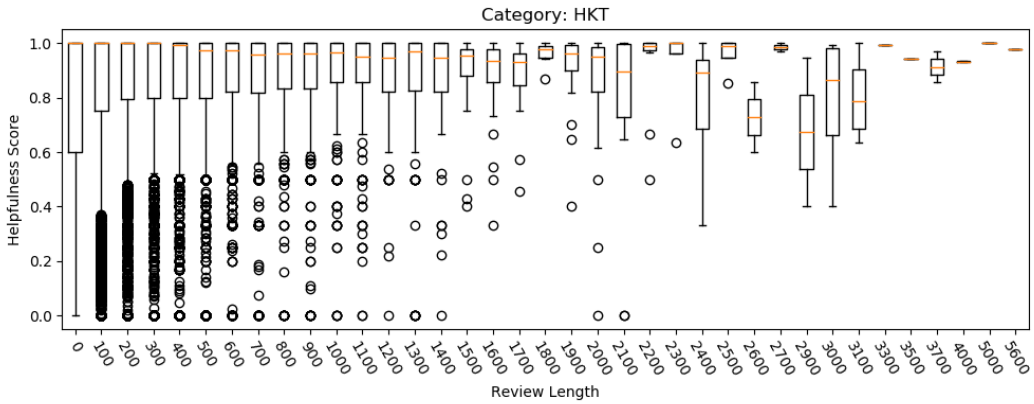
Figure 3.6: Review length vs helpfulness score for experiential goods



(a) Electronics

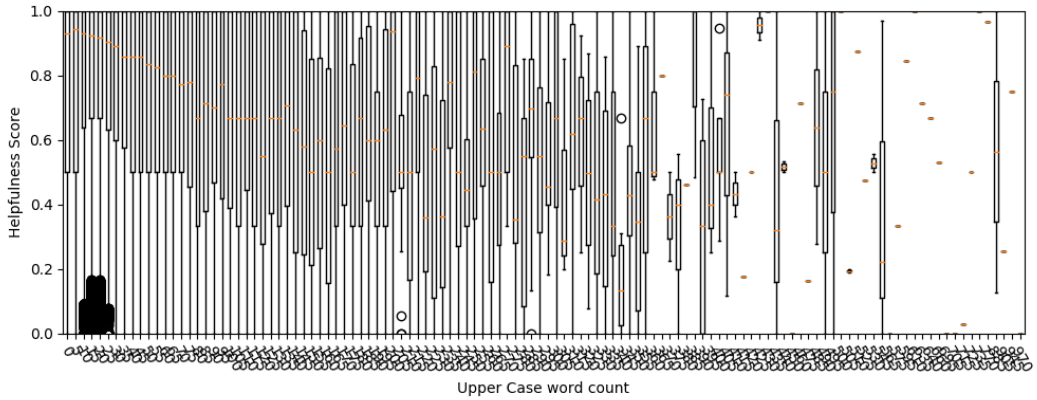


(b) Health & Personal Care

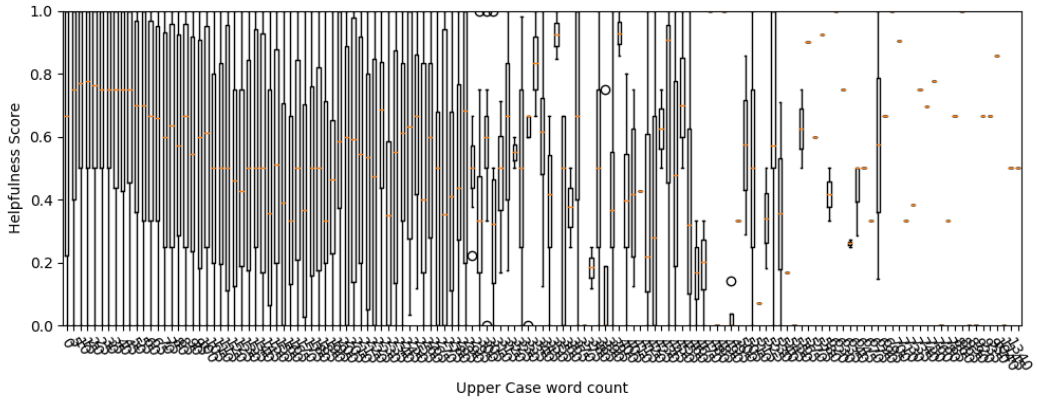


(c) Home & Kitchen

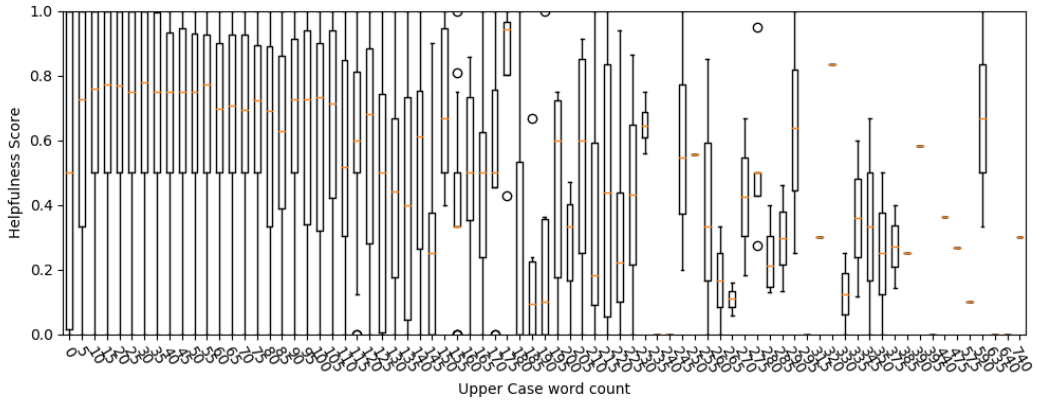
Figure 3.7: Review length vs helpfulness score for functional goods



(a) Books

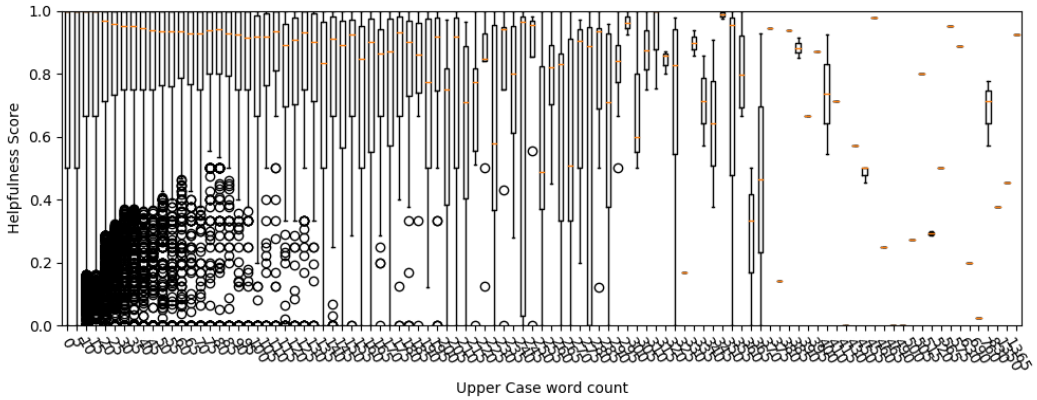


(b) Music & TV

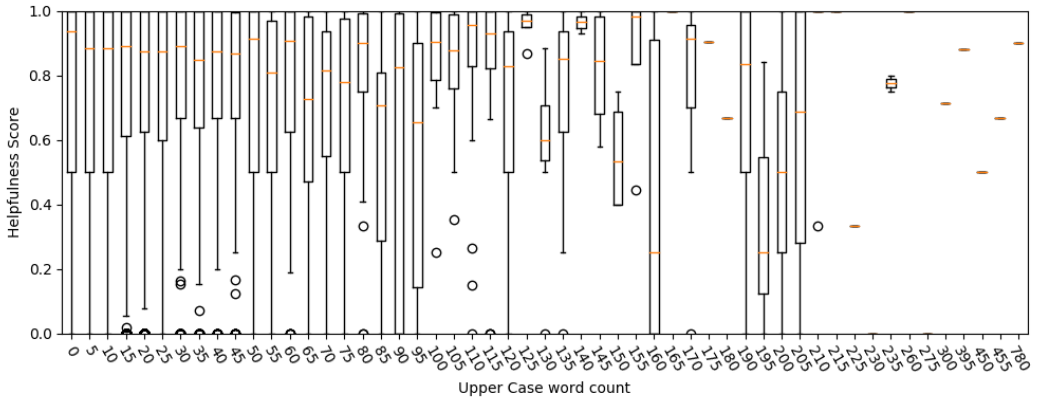


(c) Video Games

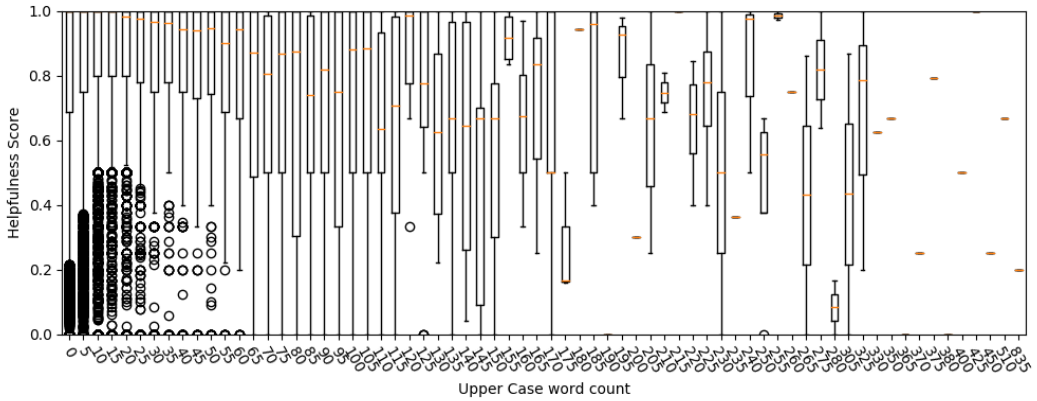
Figure 3.8: Upper case word count vs helpfulness score for experiential goods



(a) Electronics



(b) Health & Personal Care



(c) Home & Kitchen

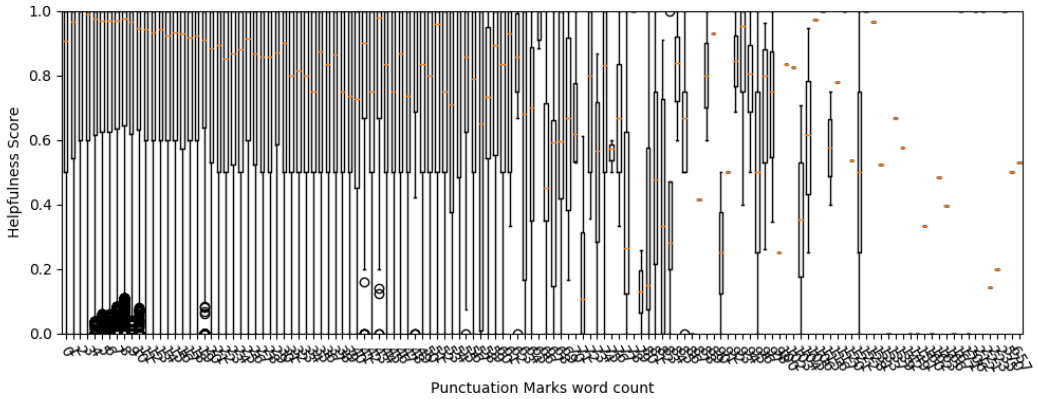
Figure 3.9: Upper case word count vs helpfulness score for functional goods

Number of Punctuation Marks

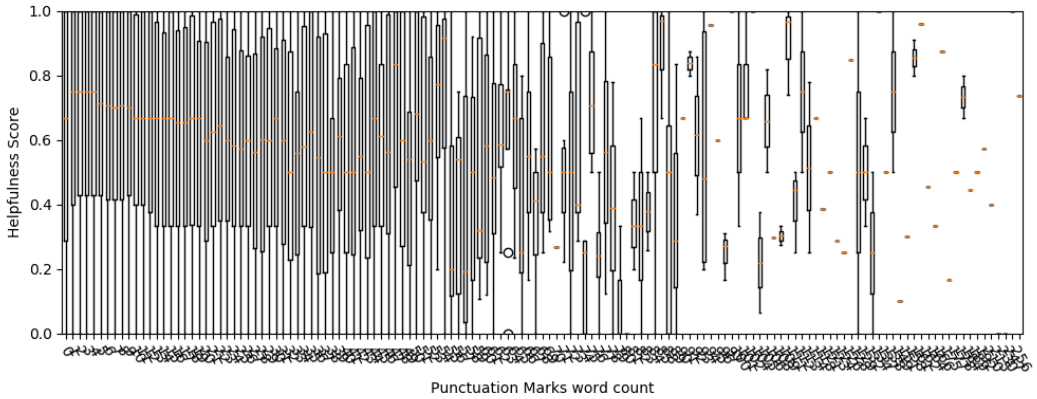
Similar to the above feature, punctuation marks such as ! and ? also are worth considering, since they affect the tone of the message and can increase the readers engagement with the review. Figures 3.10 and 3.11 shows the relationship between punctuation mark counts and helpfulness scores. As discussed in upper case word counts, we see a similar behaviour between experiential and functional goods with respect to how the helpfulness scores decrease as punctuation mark count increases.

Number of Votes

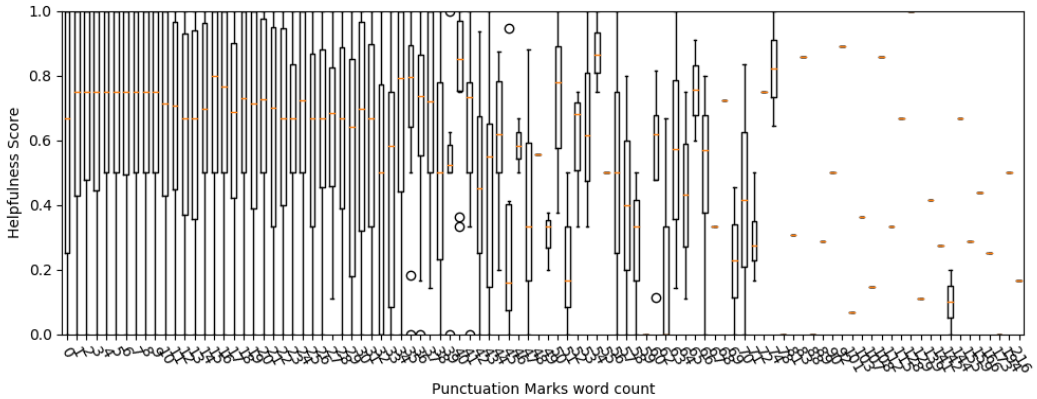
We look at voting activity on reviews since it indicates the readers' collective orientation about whether a review was helpful or not. In order to get a better understanding of how voting affects review helpfulness, we categorized number of votes for a review into various bin sizes. Figure 3.12 shows this relationship across different product categories. A key difference between experiential goods (Figures 3.12(a), 3.12(c), 3.12(e)) and functional goods (3.12(b), 3.12(d), 3.12(f)) is that experiential goods have significantly higher variance in their helpfulness scores for reviews having larger number of votes. This is particularly exemplified in Figure 3.12(e) where the distribution seems to diverge into high and low helpfulness zones.



(a) Books

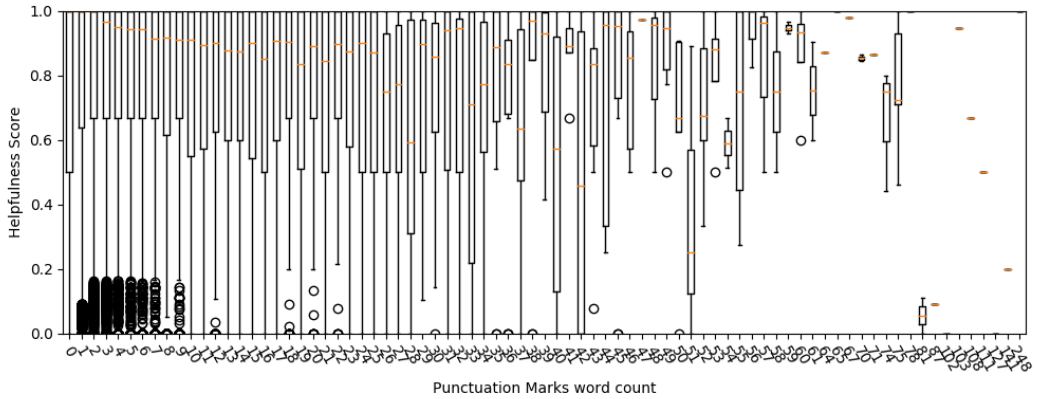


(b) Music & TV

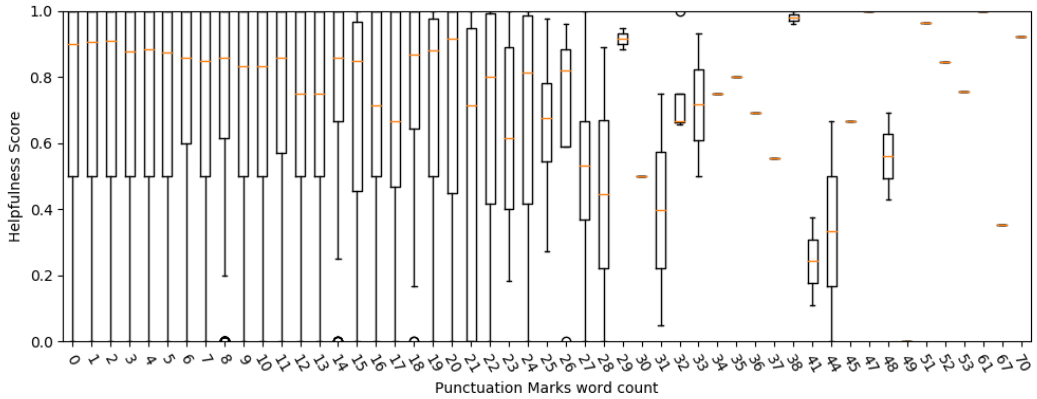


(c) Video Games

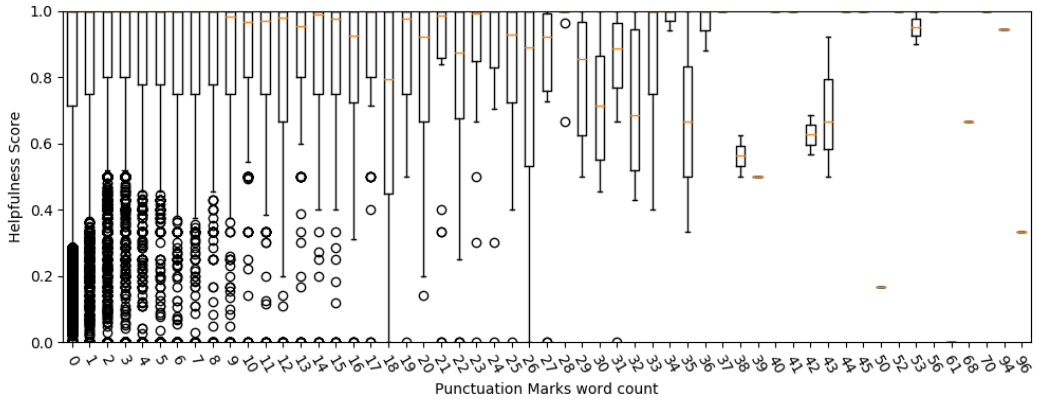
Figure 3.10: Punctuation mark count vs helpfulness score for experiential goods



(a) Electronics

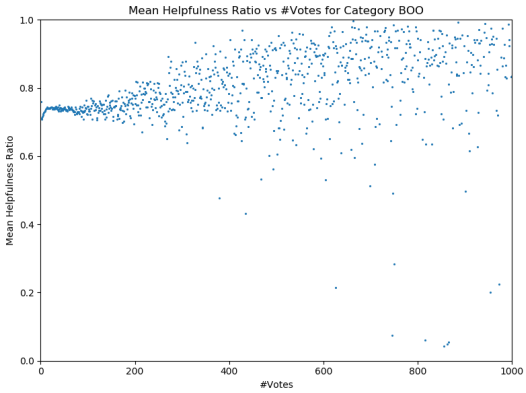


(b) Health & Personal Care

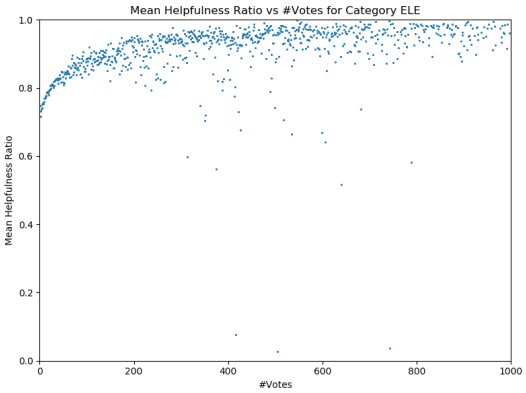


(c) Home & Kitchen

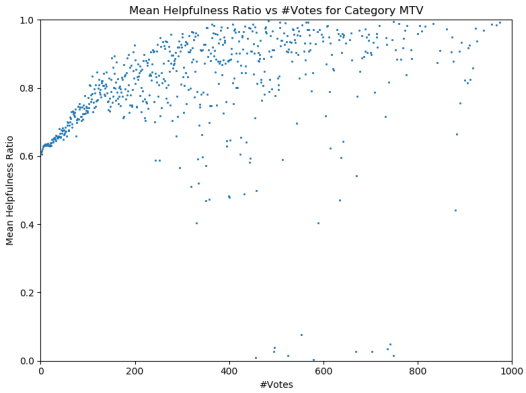
Figure 3.11: Punctuation mark count vs helpfulness score for functional goods



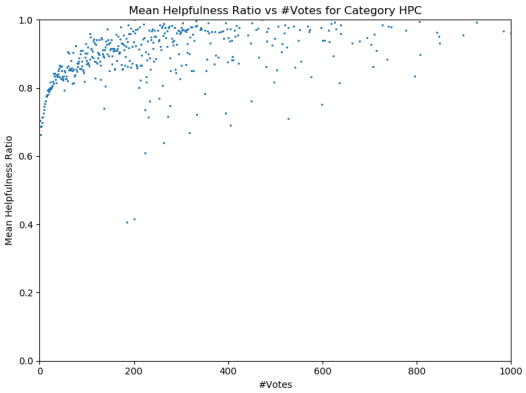
(a) Books



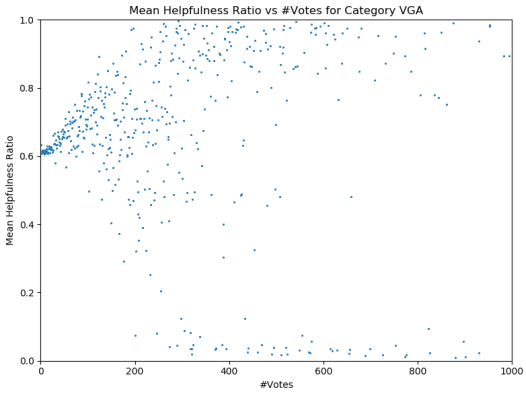
(b) Electronics



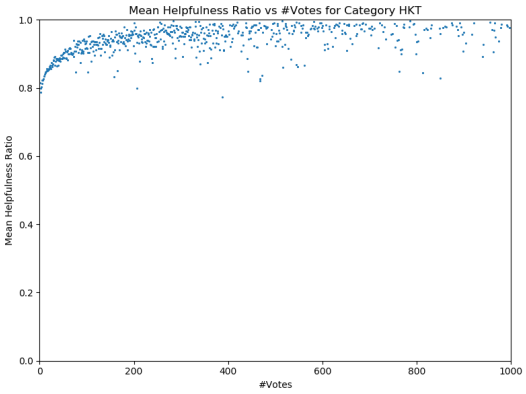
(c) Music & TV



(d) Health & Personal Care



(e) Video Games



(f) Home & Kitchen

Figure 3.12: Number of votes vs helpfulness score across categories

3.2.3 Context based features

Moving away from review and rating statistics, we propose that a review's helpfulness score also depends upon its overall quality. By quality, we mean features that represent the review in a contextual manner, such as writing style, easy of comprehension and sentiment. Therefore we seek to analyze such features, that represent the contextual aspects of the review text.

Review Sentiment

Reviews that contain elements indicating a distinct sentiment towards an entity (product or service) tend to be helpful to online consumers since they indicate a clear opinion of the reviewer on whether they liked/disliked the product. [22] [33]. Therefore we consider sentiment scores of product reviews in our analysis. In order to compute these scores, we use the Stanford nltk sentiment analyzer and select the compound (normalized) component as our sentiment score. [3].

POS Distribution

Linguistic features for predicting review helpfulness had been widely used in literature. We look at the count of Parts-of-Speech of the review text in Noun, Adjective and Verb families, with respect to their effect on review helpfulness. Table 3.1 shows the various tags under each POS family. Our intuition is that more descriptive words should correlated with higher helpfulness scores.

Noun Family Tags		Adjective Family Tags		Verb Family Tags	
POS Tag	**Description**	**POS Tag**	**Description**	**POS Tag**	**Description**
NN	Noun, singular or mass	**JJ**	Adjective	**VB**	Verb, base form
NNS	Noun, plural	**JJR**	Adjective, comparative	**VBD**	Verb, past tense
NNP	Proper noun, singular	**JJS**	Adjective, superlative	**VBG**	Verb, gerund or present participle
NNPS	Proper noun, plural			**VBN**	Verb, past participle
				VBP	Verb, non-3rd person singular present
				VBZ	Verb, 3rd person singular present

Table 3.1: Various POS tags used in our analysis

3.2.3.1 Feature Importance

In order to understand what features were more important predictors for review helpfulness, we performed a correlation test between each review feature considered and our target variable - review helpfulness score. Table 3.2 shows the results of our correlation analysis with Pearson correlation values across different product categories. Adjacent to the correlation scores are the respective two tailed p-values indicating statistical significance of the correlation scores. Our analysis shows that *mean absolute deviation, review length* are comparatively better predictors than the other review features considered for both experiential based and functional based goods. *Review sentiment score* is another feature that has relatively higher correlation for some of the product categories.

Type	Metric	BOO	MTV	VGA	ELE	HPC	HKT
mad	**score**	-0.033	**0.256**	**0.190**	**0.288**	**0.193**	**0.217**
	p-value	0.002**	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **
review length	**score**	**0.147**	**0.218**	**0.166**	**0.101**	**0.102**	**0.103**
	p-value	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **
# votes	**score**	0.024	-0.019	-0.012	0.042	0.073	0.043
	p-value	0.022**	0.075*	0.376	0.000***	0.000***	0.000***
sentiment score	**score**	-0.010	**0.105**	0.085	**0.155**	0.047	**0.146**
	p-value	0.350	**0.000*** **	0.000***	**0.000*** **	0.028**	**0.000*** **
year	**score**	-0.084	-0.066	-0.042	-0.019	-0.030	-0.042
	p-value	0.000***	0.000***	0.002***	0.063*	0.163	0.000***
# upper case	**score**	0.077	0.040	0.085	0.061	0.047	0.027
	p-value	0.000***	0.000***	0.000***	0.000***	0.027**	0.026**
# punc marks	**score**	0.019	-0.026	-0.013	-0.021	0.007	0.007
	p-value	0.074*	0.012**	0.331	0.039**	0.740	0.548
POS Noun	**score**	-0.006	-0.034	-0.002	-0.017	-0.021	-0.001
	p-value	0.588	0.001***	0.852	0.090*	0.318	0.943
POS Adjective	**score**	-0.026	-0.024	0.003	0.010	-0.037	-0.006
	p-value	0.016**	0.024**	0.803	0.309	0.086*	0.618
POS Verb	**score**	-0.001	0.026	0.032	-0.035	-0.003	-0.020
	p-value	0.908	0.012**	0.017	0.000***	0.890	0.102

Table 3.2: Feature correlation scores by product category. BOO: Books, MTV: Movies & TV, VGA:Video Games, ELE:Electronics, HPC:Health & Personal Care, HKT:Home & Kitchen. ***, ** and * represent 0.001, 0.05 and 0.1 significance levels respectively

3.3 Summary of Analysis

In this chapter, we looked at review features from ratings-based, text-statistics based and context based backgrounds. We also analyzed each feature and compared their influence on review helpfulness scores across product categories. Here we re-iterate our key findings.

- Product reviews are more likely to be voted "helpful" when
  - review star ratings are closer to the mean star rating of all reviews in that category.
  - review length is not too short (100 or more words) but not too large.
  - number of upper case words/punctuation marks are at least 10 or more but not too large.
- Experiential and Functional based goods can be contrasted as shown
  - For reviews that have high helpful scores, the distribution is more coherent (with long zipfian tails and more outliers) for functional based goods, while experiential goods have heavy tails with lesser number of outliers.
  - Reviews with large number of votes have significantly greater variance in experiential based goods as opposed to functional based goods.
  - Review features such as *rating deviation from mean, review length* and *sentiment score* are the most correlated features with review helpfulness for both classes of products.

4. ANALYSIS OF BIAS IN HELPFULNESS VOTING

4.1 Analyzing Bias in Helpfulness Voting

More recently, there have been a number of factors that can cause bias in review rankings due to manipulation of helpfulness voting. 3^{rd}-party sellers can act as agents who introduce biased user voting in the marketplace, which can affect a particular products review ranking by showing more positive reviews in the top than critical ones as shown in Figure 4.1. Marketing campaigns as shown in Figure 4.2 that encourage sellers to opt for such services that artificially boost the helpful votes on appreciative reviews and conceal critical reviews by voting them as 'not helpful'.

Such external influence can damage customer trust on the genuineness of the reviewer and eventually the overall review system. Tackling such 3^{rd}-party agents can be tricky, and is currently outside the scope of this thesis. We are particularly interested in more intrinsic sources of bias; that are either created due to the platform structure, or the inherent bias in the community.

4.1.1 Accumulative Bias

This form of bias is self-propagating on platforms where being early has an accumulative reward effect; an early review (that may be average in its true helpfulness) gets viewed first due to its existence, gets voted up which in turn increases its review rank thus forming a self-sustaining cycle referred to as the Matthew Effect. This 'rich-get-richer' form of bias has been studied in the context of online e-commerce reviews [30]. However we aim to provide preliminary models to reduce the effect of this form of bias.

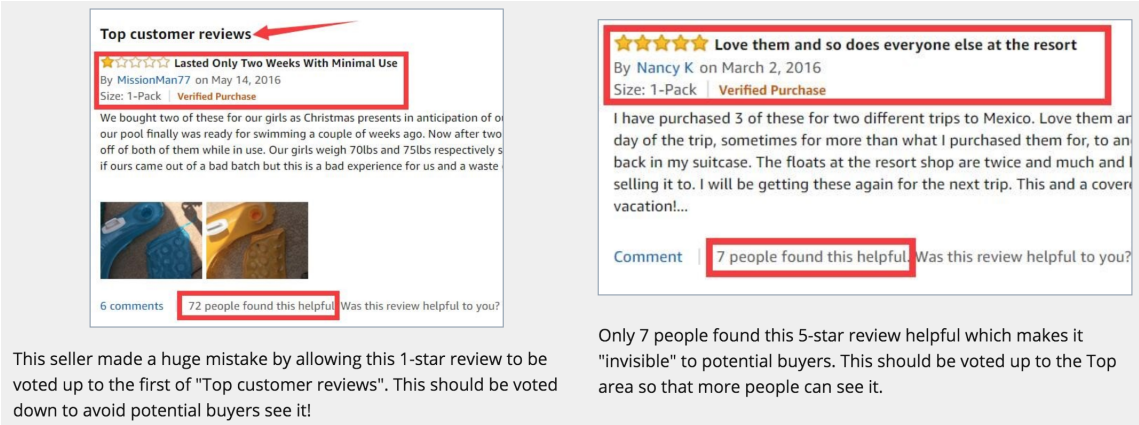


Figure 4.1: Advertisement campaigns to promote manipulation of review voting

4.1.2 Opinion Bias

Previous studies on review helpfulness have tested various hypothesis on the similarity between the reviewer and the reader's mindset/interests [7]. Other works have posited that the review helpfulness scores are not representative of the 'true' scores (i.e. if the entire population of shoppers voted on the reviews) [30]. Therefore, there seems to exist a inherent opinion withing each review reader that drives an opinion-based bias.

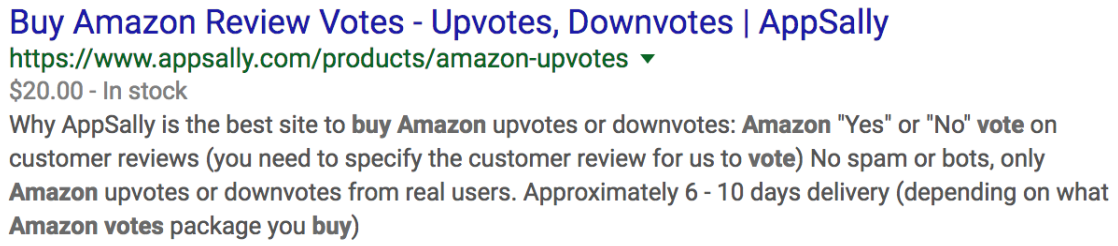


Figure 4.2: 3^{rd} party services selling upvotes to sellers

We are interested in exploring this form of bias in a topical context, where each review can be described using a distribution of key topics. In order to get the context-based topical structure from reviews of a product category, we employ unsupervised machine learning approaches for generative modelling.

4.2 Detecting Bias in review helpfulness

In this section, we discuss methods used for analyzing the presence of intrinsic biases in the review helpfulness setting.

4.2.1 Detecting Accumulative Bias

To detect accumulative bias within a product category, we first define 3 classes for helpfulness scores - low score ($0 < r_{hr} \leq 0.33$), medium score($0.33 < r_{hr} \leq 0.67$) and high score ($0.67 < r_{hr} \leq 1.0$). Then we look at the proportion of reviews in each score class, grouped by year to understand the proportion of year-wise review contribution in each score class. In the presence of no bias, each review group would have an equal probability of being voted as helpful/not-helpful regardless of the time it was posted. Figure 4.3 shows how the proportion of new reviews changes from low helpfulness to high helpfulness class across categories. We see that throughout each category, there is some degree of reduction in the proportion of new reviews, which strongly indicates the presence of accumulative bias.

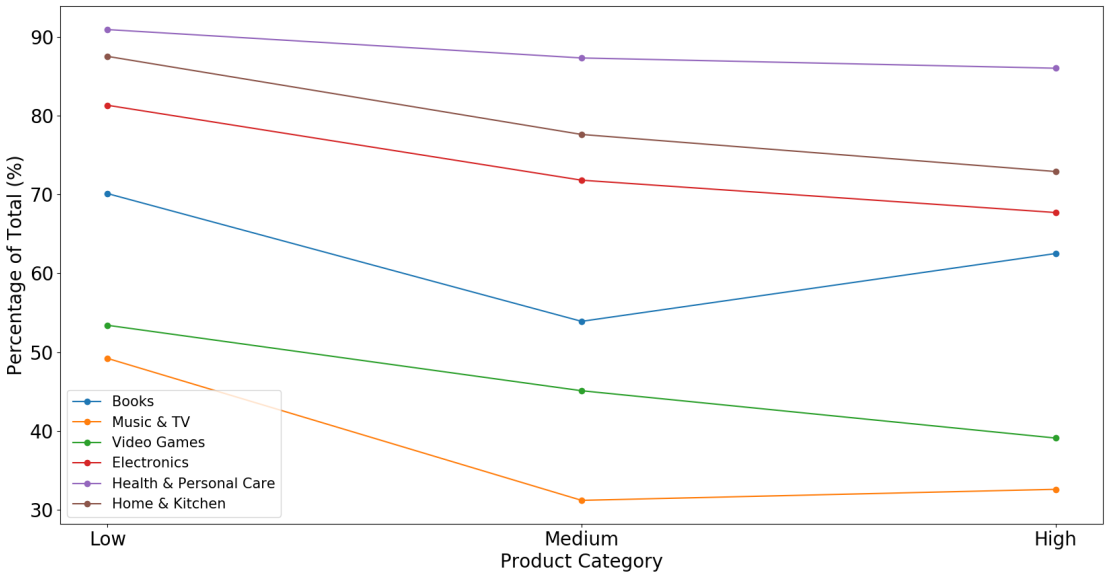


Figure 4.3: Helpfulness score classes vs proportion of new reviews

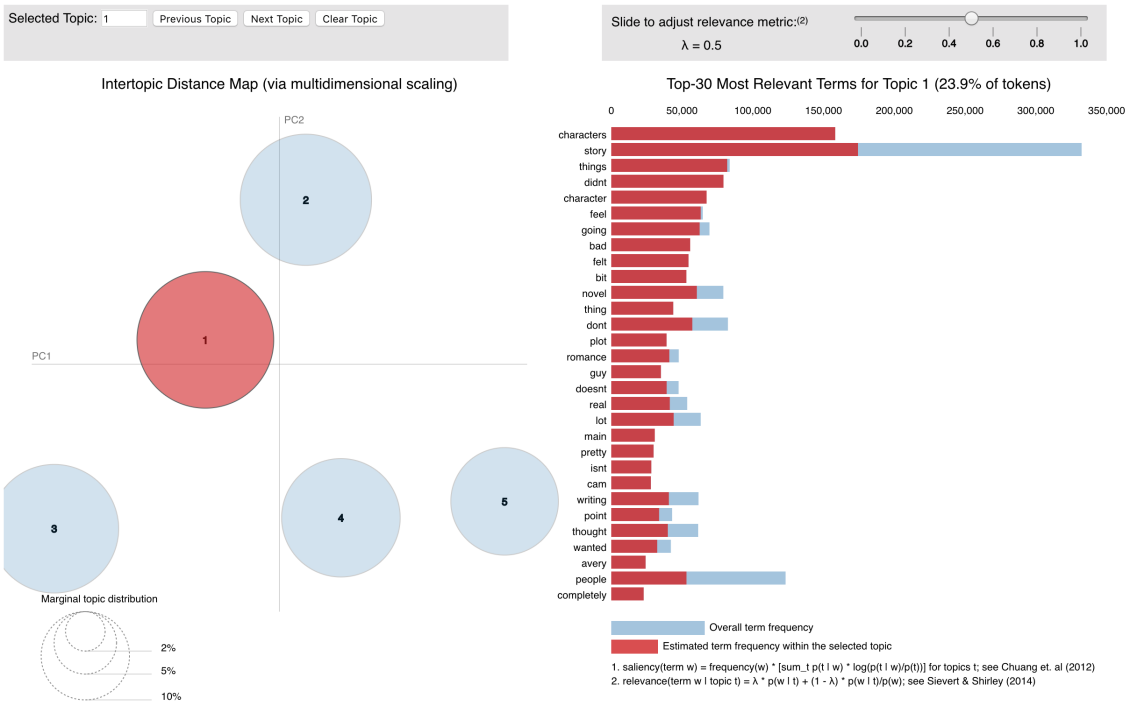
4.2.2 Detecting Opinion bias using Topic Models

As shown in previous studies, consumers tend to have self-developed biased based on their personal interests and preferences, which can determine their opinion on a particular review. For example, if a diverse group of consumers are asked what makes a helpful review for an electronics product (laptop, mobile phone), we might hear varying answers like 'talks about the durability of the laptop over time' or 'lists pros/cons of the product' or 'compares to other products in similar price range'. Such contexts if extracted from review text, can be useful in determining review helpfulness.

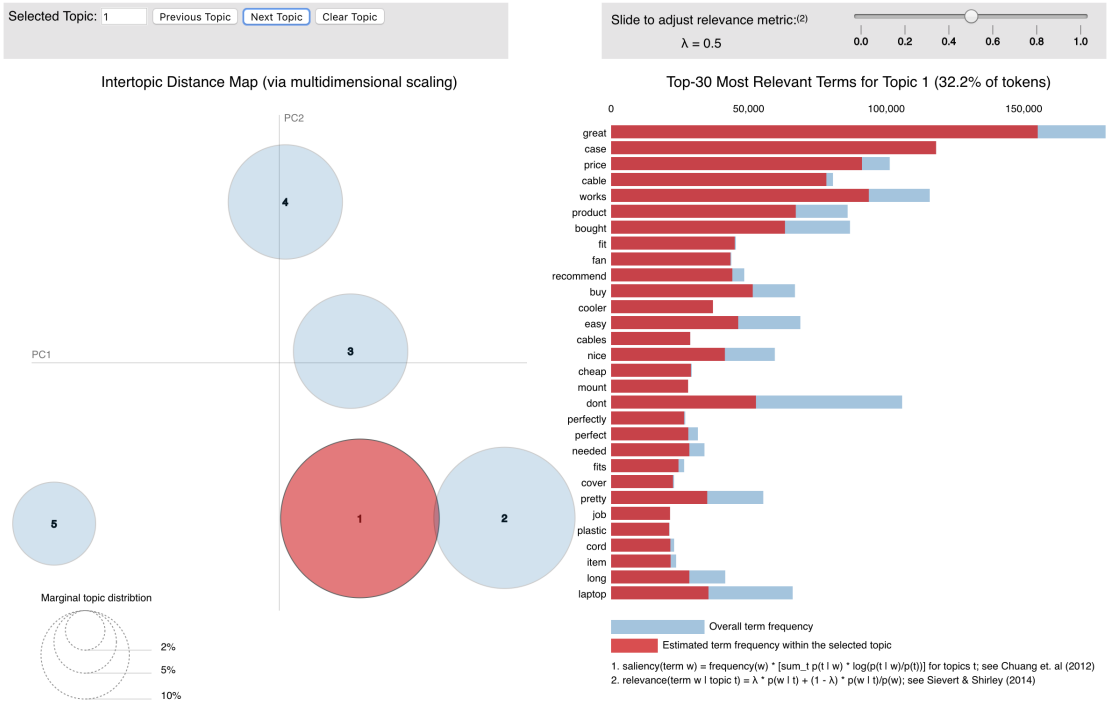
We explore topic modelling as a generative modelling approach to construct topic distributions for product reviews across various categories. We employ a popular topic modelling approach known as Latent Dirichlet Allocation (LDA) [4] which represent documents as a probability distribution of various topics (which themselves

are represented as distributions of words). In order to visualize these topic distributions, we employ a popular library - pyLDAvis [28] which shows the topic cluster orientations in a low-dimensional space as well as the most relevant term distribution for each topic. Figures 4.4, 4.5 and 4.6 show the LDA topic distributions across different product categories.

From these plots we observe that topic distributions vary across each category. However, certain similarities exist in how topics themselves are distributed in terms of context words. For experiential goods, the most dominant topics (highlighted in red) have word frequency distribution with light tails, compared to functional based goods. This indicates that contextually, a category such as *Books* may contain similar descriptive words but *Electronics* will have words that are more uniformly distributed across reviews.

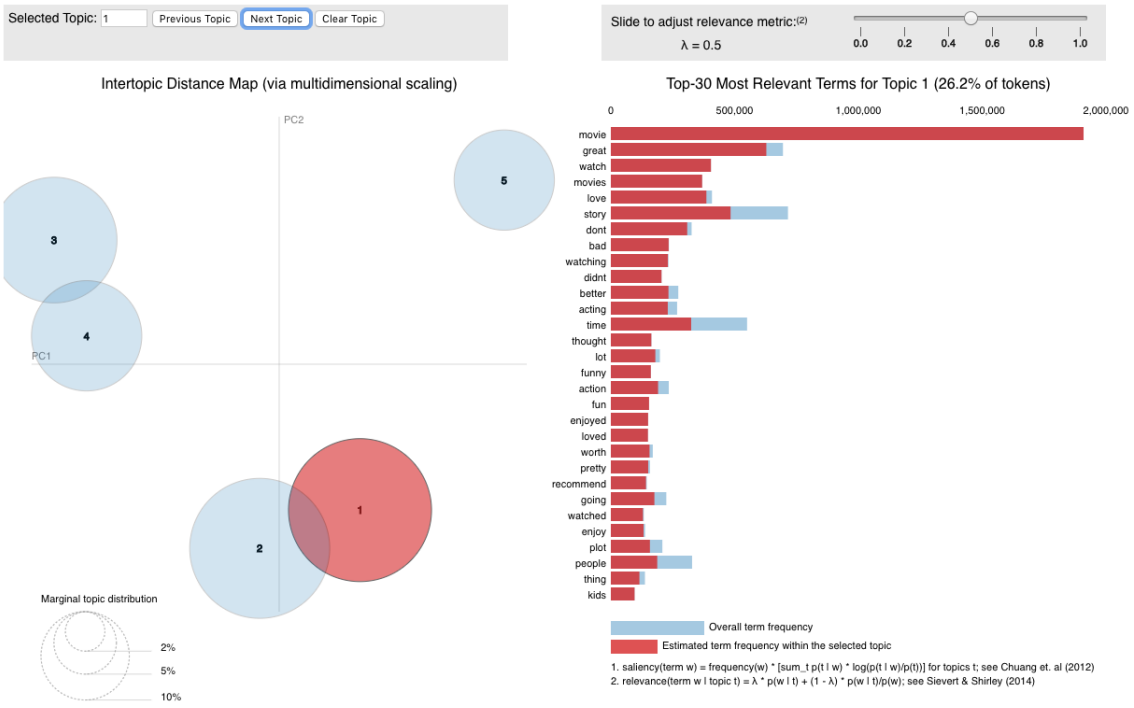


(a) Books

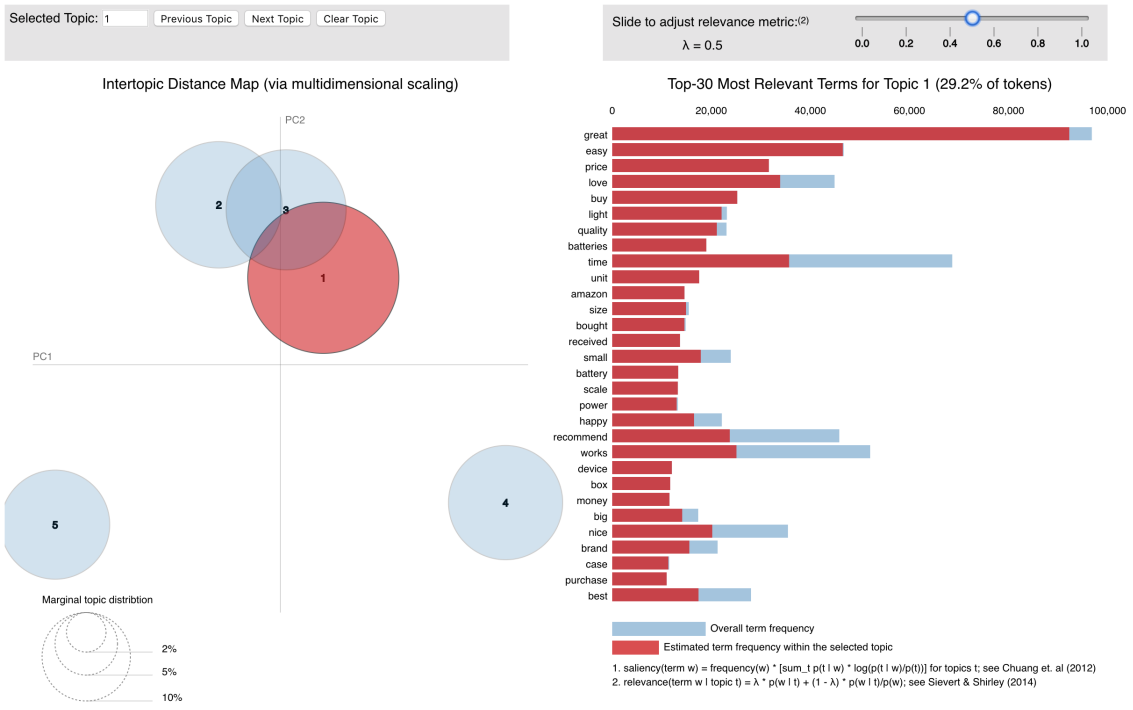


(b) Electronics

Figure 4.4: LDA based topic models for Books and Electronics

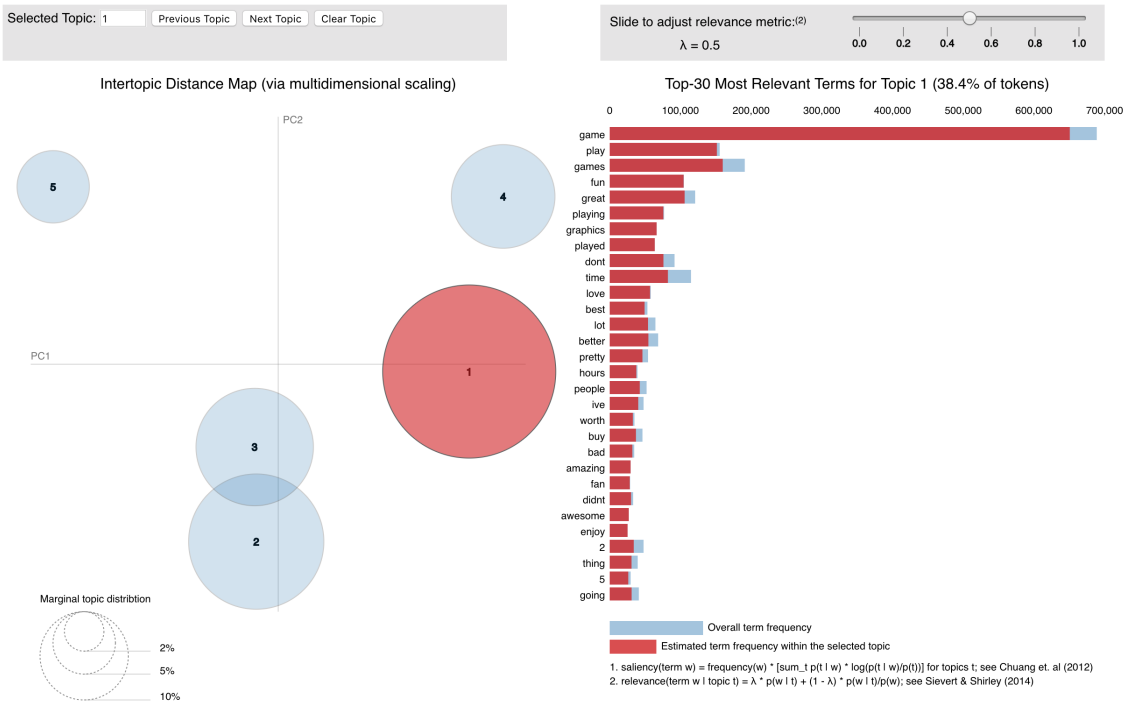


(a) Music & TV

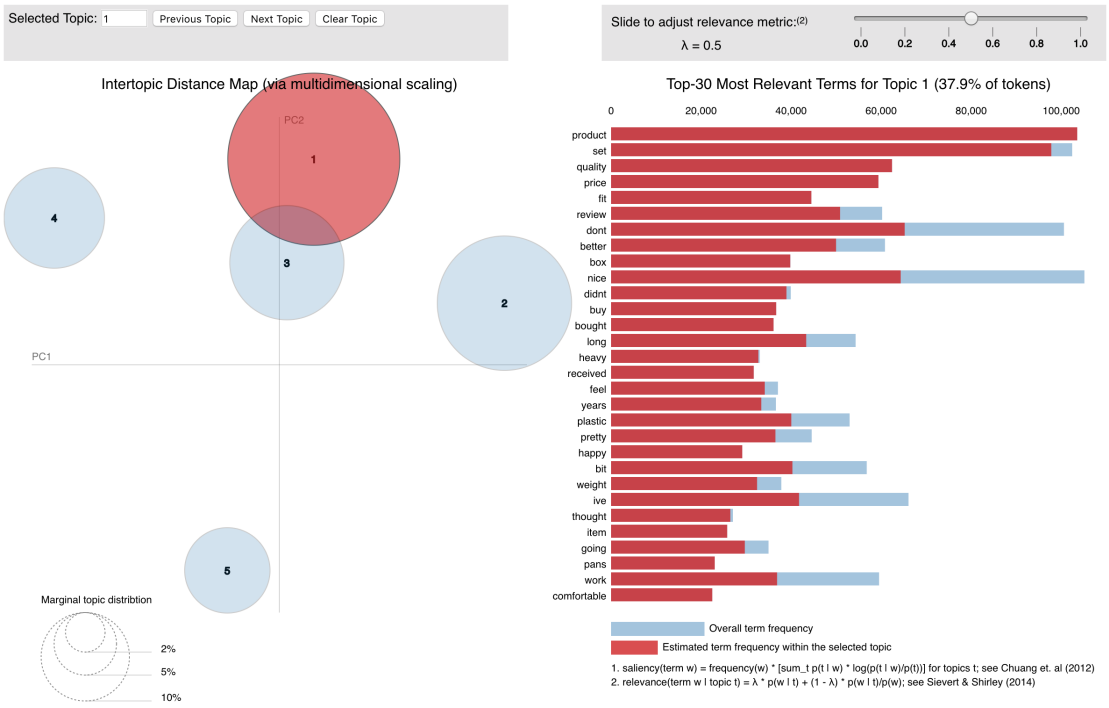


(b) Health & Personal Care

Figure 4.5: LDA based topic models for Movies & TV and Health & Personal care



(a) Video Games



(b) Home & Kitchen

Figure 4.6: LDA based topic models for Video Games and Home & Kitchen

4.3 Summary of Analysis

In this chapter, we explored possible sources of intrinsic biases in review helpfulness voting. We summarize our key findings below

- **Accumulative bias**
  - Also known as "Matthew Effect", this refers to intrinsic bias due to reward policy that benefits entities existing earlier in time, causing them to get higher scores and pushing their rankings up, in a self-perpetuating fashion.
  - We observed how the proportion of newly posted review decreases as helpfulness score bins increase from low to high, indicating that likelihood of older reviews receiving higher helpful votes is more.
- **Opinion bias**
  - We also looked at user preferential bias and explored topic modeling techniques to understand contextual distribution of reviews across different product categories.
  - For experiential goods, word distribution on topics had lighter tails and consisted of few key terms, while functional goods had more uniformity in their term distribution over topics.
  - We explore crowd-annotated results on opinion bias in Section 5.

5. MODELS TO MITIGATE VOTING BIAS

Based on the observation of bias and helpfulness in the previous chapter, we turn here to exploring methods to mitigate this bias. Our goal is to explore preliminary approaches based on unsupervised learning methods that could be used to mitigate intrinsic bias to some extent. Figure 5.1 shows a more generalized view of our approach to model building. We select reviews across different product categories from our curated data. We extract the relevant features based on our analysis of feature importance to review helpfulness.

To this end, we experimented with different techniques such as tf-idf, Doc2Vec, and combining Doc2Vec with engineered feature embeddings derived from review meta-data (Doc2Vec*).

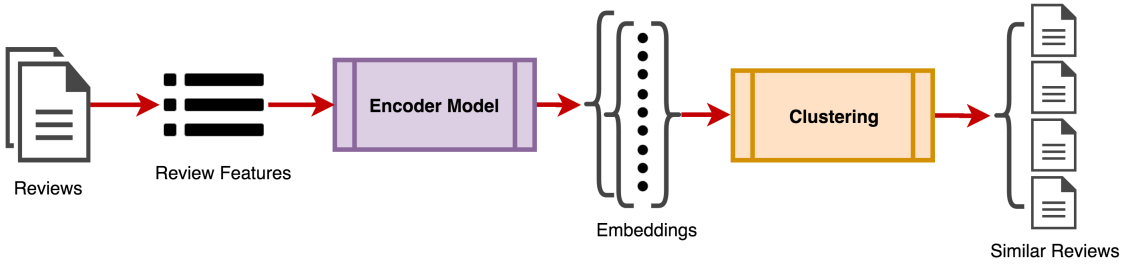


Figure 5.1: General Model building approach

In the rest of this chapter, we introduce our approach to building models that aim to reduce the biases present in review helpfulness voting. Our goals is to recommend a set of top-k most similar reviews $r \in S_{r_t}$ within a category, given a target review r_t. Therefore, we consider various methods to compute review similarity, which differ in

their review attribute selection.

5.1 Baseline Model: tf-idf

Term frequency-inverse document frequency is a widely used numerical statistic in the field information retrieval and document mining. It is used to represent the importance of a particular word in a document or corpus, and therefore, is the basis of modern search engine algorithms.

The tf-idf score is a product of two different numerical scores, the term frequency tf - which indicates how frequent a particular term t is in a given review document d, and the inverse document frequency idf which indicates the degree of informativeness of a particular term t. By degree of informativeness, we mean that words that are more commonly used across most documents (common nouns) contain comparatively lesser information as opposed to more-specific words (technical jargon) that may be present in a select few documents.

To compute the tf component, we compute the raw count of the term t in review document d. The idf component is calculated by first computing the document frequency $df_{t,d}$ of term t as the number of review documents that contain term t. The idf score is then calculated as the logarithm of the ratio of total documents in the corpus to the document frequency of term t.

$$\begin{aligned} tf_{t,d} &= \text{Frequency count of term } t \text{ in review document } d \\ df_{t,D} &= \text{Frequency count of review document } |d \in D : t \in d| \end{aligned} \tag{5.1}$$

$$idf_{t,D} = \log(1 + \frac{N}{df_{t,D}}) \tag{5.2}$$

where N = $|D|$ is the total number of documents in corpus D. Finally, the tf-idf score

for term t and review document d is given as

$$\text{tf-idf}_{t,d} \;=\; tf_{t,d} \times idf_{t,D} \;=\; tf_{t,d} \times \log(1 + \frac{N}{df_{t,D}}) \tag{5.3}$$

Using the tf-idf score information, we create an embedding vector $U_d \in \mathbb{R}^{\mathbb{T}}$ for each review document d where T is the term vocabulary size in corpus D. It is a trivial observation that U_d is highly sparse since the the term count in a document d is significantly smaller than the term vocabulary size $|V|$.

$$\forall t \in T \qquad U_d^t = \begin{cases} > 0 & \text{if } t \in d \\ 0 & \text{if } t \notin d \end{cases} \tag{5.4}$$

It is important to point here that the tf-idf score for each term t is computed using a 'bag-of-words' (BOW) model approach. In other words, the relative ordering of terms in the review document is not considered important; only the presence of a particular term is counted. This approach has two clear disadvantages. Firstly, the embedding vectors are highly sparse since the vocabulary size T is much larger than the set of terms in any given review document d. This sparsity often generates dissimilarities between two reviews that may be similar in context, but different in the usage of terms. Secondly, the bag-of-words model approach doesn't take into account, the semantic aspect of the review document. Therefore, words sequences that have complementary meanings can have a high similarity score, if the terms used are the same.

As discussed in the previous chapter, the general approach to our model building remains relatively unchanged. Figure 5.2 shows our model for tf-idf based embeddings. We pre-process the review corpus, to select ASINs within each category that contain at least 200 reviews. This is done to ensure the non-sparsity of the embed-

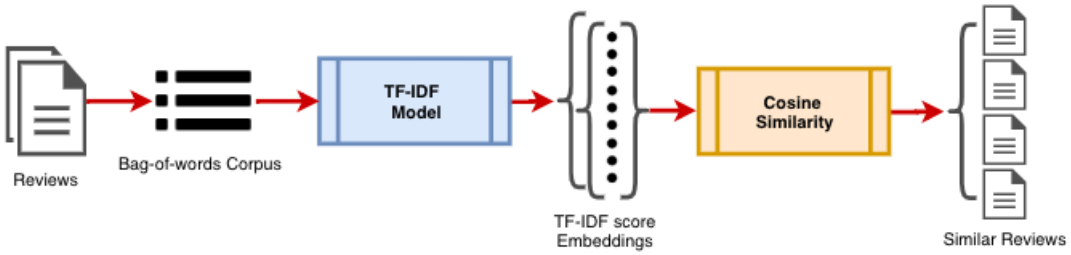


Figure 5.2: Model architecture for TF-IDF based embeddings

ding space. We then create a bag-of-words (BOW) corpus from a particular product's review text for each category.

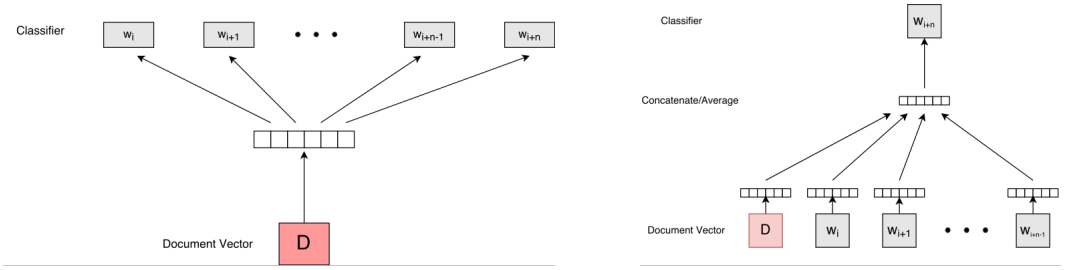
Using the BOW corpus, we create a tf-idf model which gives us an embedding vector with a dimension equal to the term vocabulary size. Using these embeddings, we generate a tf-idf similarity matrix M where any element $m_{ij} \in M$ represents the cosine similarity score between review i and review j for a single ASIN in each category. Our final 'most-similar' reviews are the top-k candidate reviews for each target review based on the cosine similarity score.

5.2 Including Semantics: Doc2Vec Embedding Approach

We can think of improving our baseline by incorporating semantic information from our review document embeddings instead of simply using numerical statistics. Word embedding techniques are an excellent approach to language modeling, which generate dense word vectors (compared to the large and sparse tf-idf vectors) that contain semantic information. A popular method for generating word embeddings is Word2Vec, which uses neural networks get a vector representation of words in context. [21]

Although word embeddings are helpful in determining contextual similarity between words, we are particularly interested in similarity between entire reviews.

We use the Doc2Vec approach to create numerical representation of review documents/paragraphs [17]. Doc2Vec is based upon Word2Vec with the difference that we use an additional feature - the paragraph ID along with the individual words as input to the neural network.



(a) PV-DBOW (b) PV-DM

Figure 5.3: Doc2Vec methods to create paragraph vectors

Figure 5.3 shows two widely used implementations of Doc2Vec for generating Paragraph Vectors (PV) using Distributed-Bag-of-Words (PV-DBOW) which is based on the skip-gram Word2Vec model, and Distributed Memory (PV-DM) based on the Continuous-Bag-of-Words (CBOW) Word2Vec model [11]. The D vector i.e. the reviewID, is unique to each review and therefore can be considered as useful information about the review itself. It is trained along with the word vectors and becomes the numeric representation of the review. In our approach we chose to proceed with the PV-DM model as it has been shown to superior to the PV-DBOW model in terms of achieving state-of-the-art results.

Figure 5.4 shows our approach to generating Doc2Vec embeddings. Similar to the previous work flow, we convert review text into the desired input format for the Doc2Vec model using a tagged document dictionary. The output from the model are the latent embeddings for the review document vector. Each review was represented

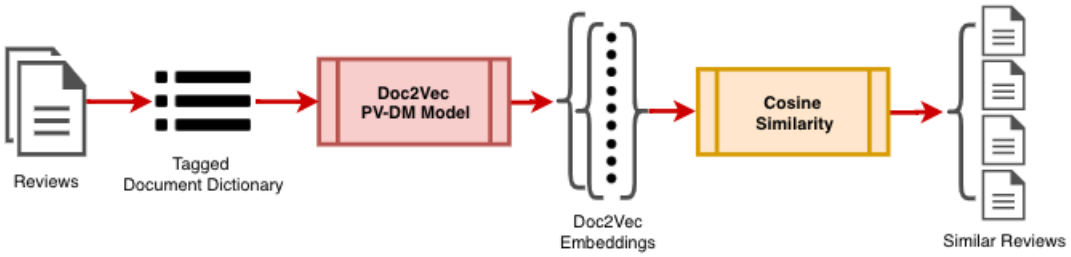


Figure 5.4: Model architecture for Doc2Vec based embeddings

as $r \in \mathbb{R}^d$ where $d = 300$ was a 300-dimensional embedding. For each target review, we then compute the cosine similarity metric and select the top-k most similar reviews.

5.3 Combining Semantics and Metadata: Doc2Vec* Model

The Doc2Vec model yields good representation of review text as vectors and gives similar reviews that are more coherent compared to our baseline models. However, the above methods do not account for any biases that may be present in the review, possible due to review attributes such as length, number of votes etc. Therefore, we create additional embedding dimensions using various review features discussed before. Our new review embedding vector can be represented as $r \in \mathbb{R}^{d+k}$ where $k = 10$.

The 10 additional features selected for the Doc2Vec* model were -

- Rating deviation from mean
- Review length
- Number of votes
- Upper-case word count
- Punctuation marks count
- Sentiment Score
- Year of review post
- POS - Noun (NN) tags count
- POS - Adjective (JJ) tag count
- POS - Verb (VB) tag count

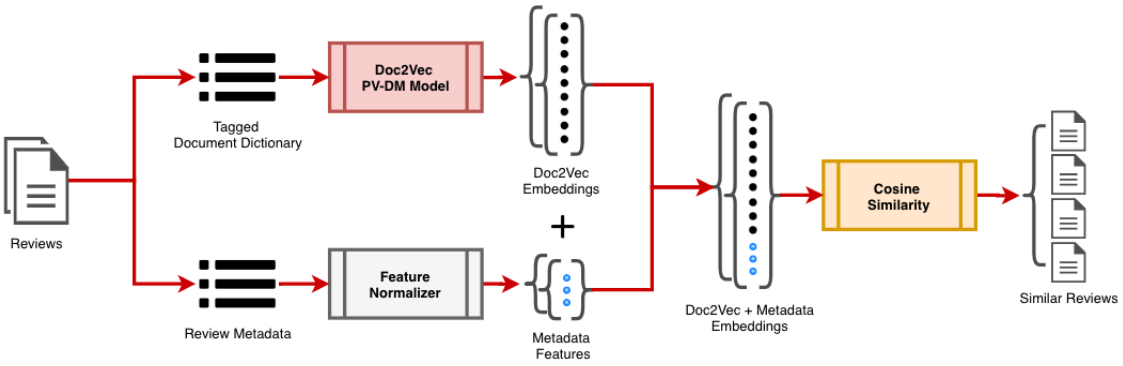


Figure 5.5: Model architecture for Doc2Vec* based embeddings

To ensure that all features are equally weighted, we normalized each feature score (length, num votes, etc.) f_i from its feature distribution F as shown in equation 5.5

$$\hat{f}_i = \frac{f_i - \mu_F}{|\underset{i}{\text{argmax}}(F) - \underset{i}{\text{argmin}}(F)|} \quad \forall f_i \in F \tag{5.5}$$

The normalized review features are concatenated to the existing Doc2Vec review embeddings as shown in Figure 5.5. For our experiments, we had 300 Doc2Vec embeddings and 10 normalized review features. We conjecture that using additional attributes will normalize the review representations therefore reducing intrinsic biases, and yield higher-quality neighbors in terms of contextual similarity.

5.4 Evaluation

We employ crowd-sourced evaluation strategies for analyzing our model performance, due to the unsupervised nature of our task . Since we require evaluation over different model approaches, we found that large-scaled crowd-based annotation platforms such as Amazon's Mechanical Turk would give us results that are most representative of a large online community.

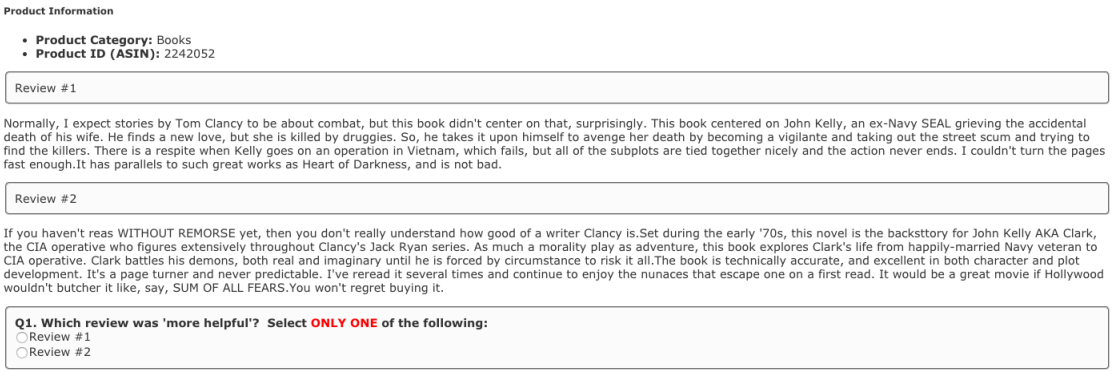
5.4.1 Creating Amazon MTurk Questionnaire

We formulated our survey as a questionnaire designed to show reviewers different product reviews and ask them which one do they find to be more helpful. In order to draw comparisons between the reviews, we decided to keep two reviews side-by-side for each question. The target review had a relatively older date (before 2010) and a high helpfulness score ($hr > 0.7$), while the candidate review was one with the top-k (k=5) highest cosine similarity score to the target review embedding. These two categories of reviews were intentionally chosen since we wanted to contrast whether the effect of accumulative bias can be mitigated to some extent by showing two product reviews that are highly similar.

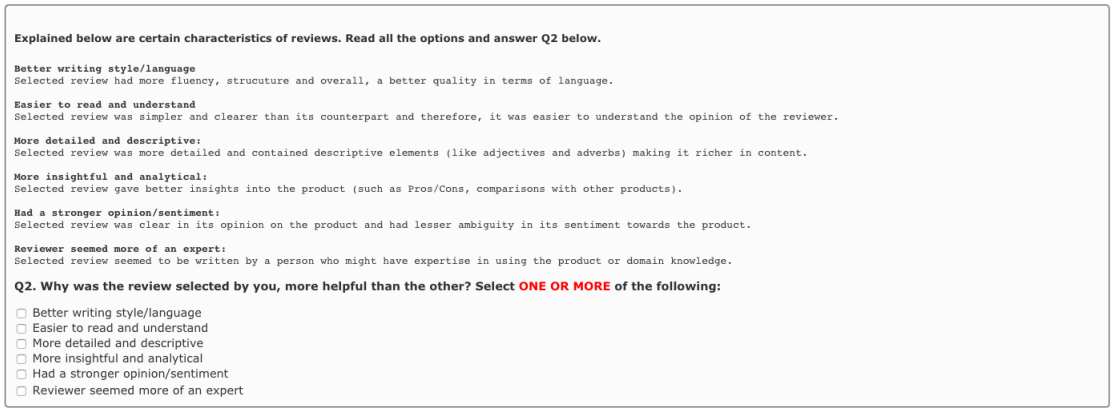
We were interested in two aspects of evaluation for each of the models proposed. Firstly, we wanted to know whether the crowd is more likely to vote for the target review (which had a high likelihood of accumulative bias) or the candidate review which was similar to the target review, but did not have high voting activity and hence could, not be shown to have such a bias.

Secondly, we wanted to get feedback from the crowd on their reasoning for voting a particular review as more helpful over the other. This would provide insight into the self-selection bias of user opinion subjectivity, that is intrinsic to such online platforms. Understanding user-specific preferences with respect to what constitutes helpfulness in a review would be paramount to review recommendation engines that aim to improve review relevancy.

Moreover, we wanted to observe this behaviour across all categories under this study and understand the variation in crowd-behaviour between the experiential and functional class categories. We also wanted to understand how each of our proposed models performed in these scenarios.



(a) Helpful Review Question



(b) Reviewer Feedback Question

Figure 5.6: Amazon MTurk questionnaire sample

We curated a total of 90 review questions (6 product categories $\times$ 3 products (ASIN) per category $\times$ 5 target-candidate pairs per ASIN). To ensure that the target and candidate pairs were shown in an unstructured way we randomly sampled their order of occurrence. For each question, we collected a total of 10 responses to get an unbiased consensus on that question. We conducted independent surveys for each model to and performed a comparative analysis of the crowd's response across each model. Figure 5.6 shows an example of a review question in our survey. We show two reviews and ask the reader to select the 'more helpful' option. The second questions seeks to understand what constitutes the 'helpfulness' aspect in the selected

review, according to the reader. We provided the readers with 6 different options that consisted of different aspects that may be indicators of a helpful review, which are described as shown below

- **Writing Style:** The selected review had smooth flow, well-worded sentences and definite structure.
- **Review Readability:** The selected review was simpler and clearer than its counterpart making it easier to understand the opinion of the reviewer.
- **Degree of Description:** The selected review was more detailed and contained descriptive elements (like adjectives and adverbs) making it richer in content.
- **Degree of Insight:** The selected review gave better insights into the product (such as Pros/Cons, comparisons with other products).
- **Polarity of Opinion/Sentiment:** The selected review was clear in its opinion on the product and had lesser ambiguity in its sentiment towards the product.
- **Reviewer Expertise:** The selected review seemed to be written by a person who might have expertise in using the product or domain knowledge.

5.4.2 Analyzing Top-k similar reviews

The first part of our analysis is to understand how the top-k most similar reviews for a given target review vary across categories and between different embedding approaches. Table 5.1 shows various model results on the distribution of average proportion of Old reviews (before 2010) and New reviews (after 2010) for different products and categories. The proportion was calculated for each target review for a product ASIN using top-100 most similar reviews. We see that the Doc2Vec* yields the higher % of new reviews and therefore, gives the overall best performance.

Class	Type	ASIN	Product	tf-idf		Doc2Vec		Doc2Vec*	
				OR (%)	NR (%)	OR (%)	NR (%)	OR (%)	NR (%)
Experiential	**BOO**	0002242052	Tom Clancy - Without Remorse	54.0	46.0	47.8	**52.2**	58.4	41.6
		0006514006	Phillipa Gregory - The Other Boleyn Girl	81.7	18.3	75.7	**24.3**	83.7	16.3
	VGA	B00005Q8M0	Super Smash Bros Melee	91.7	8.3	87.1	12.8	86.5	**13.5**
		B000FQ9QVI	Super Mario Galaxy	81.9	18.1	75.3	**24.7**	76.2	23.8
Functional	**ELE**	B00004THCZ	Canon Telephoto Zoom Lens	45.6	54.3	35.1	64.9	21.8	**78.2**
		B00004T8R2	Panasonic Stereo Headphones	17.7	82.3	14.7	85.3	11.9	**88.1**
	HKT	B000DLB2FI	Keurig Coffee Filter	27.3	72.7	25.0	75.0	23.1	**76.9**
		B0009ONZ8G	Hoover Vacuum Cleaner	70.9	29.1	62.3	37.7	54.1	**45.9**

Table 5.1: Evaluation results for top-100 similar reviews. OR:Old Review, NR:New Review, BOO:Books, VGA:Video Games, ELE:Electronics, HKT:Home & Kitchen

5.4.3 Analyzing MTurk Responses

Our evaluation from MTurk responses is shown in Table 5.2. For our survey, we provided the (target, candidate) review pairs using the top-5 similar reviews for each target review under consideration. Since the products were not common across each model, we show the category-wide performance of each model in terms of proportions of old vs new reviews voted as 'most helpful' by the crowd. In this scenario as well, we observe that Doc2Vec* model performs comparatively better than the tf-idf and Doc2Vec models. This provided strong evidence that models incorporate semantic information with review metadata are best equipped with the ability to mitigate accumulative bias.

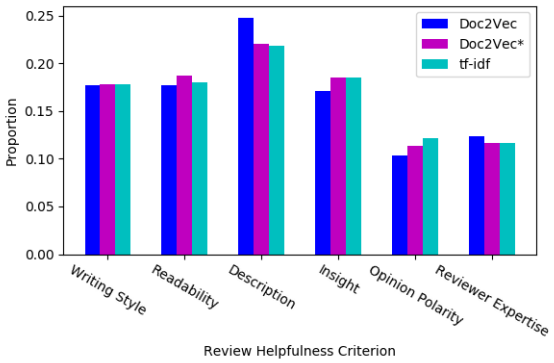
Class	Type	tf-idf			Doc2Vec			Doc2Vec*		
		#Votes	OR (%)	NR (%)	#Votes	OR (%)	NR (%)	#Votes	OR (%)	NR (%)
Experiential	BOO	150	69.3	30.7	150	58.0	42.0	149	**51.0**	**49.0**
	MTV	150	56.0	44.0	150	50.0	50.0	156	**43.4**	**56.7**
	VGA	149	48.3	51.7	150	50.0	50.0	150	**47.3**	**52.7**
Functional	ELE	150	**49.3**	**50.7**	152	52.0	48.0	156	54.0	46.0
	HKT	149	64.4	35.6	150	**55.3**	**44.7**	150	60.7	39.3
	HPC	150	43.3	56.7	150	56.0	44.0	150	**32.0**	**68.0**

Table 5.2: Evaluation results on Amazon MTurk responses. OR:Old Review, NR:New Review, BOO:Books, MTV:Movies & TV, VGA:Video Games, ELE:Electronics, HKT:Home & Kitchen, HPC:Health & Personal Care

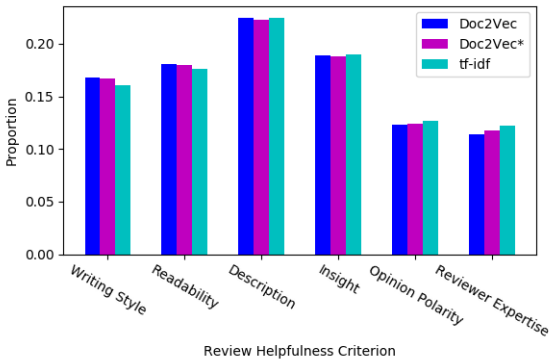
Figure 5.7 shows the distribution of various criterion that user's selected as their primary reason for voting a particular review as helpful, based on review pairs generated by 3 different models. From the plots we can make two clear observations.

Firstly, the frequency distribution of criterion is non-uniform - the most voted reason for helpfulness was *Degree of Description* while *Polarity of Opinion/Sentiment* and *Reviewer Expertise* were voted with the lowest frequency. This confirms our hypothesis on the 'homophily effect' that different users have different opinions when it comes to what aspect of a review do they find the most helpful.

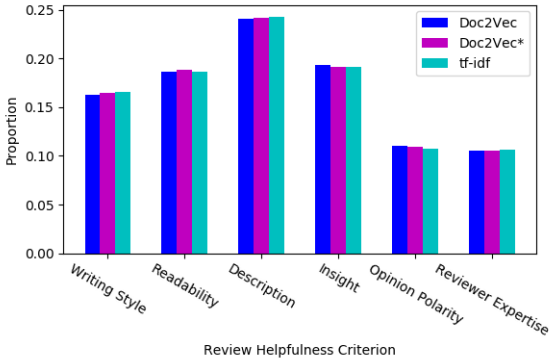
Secondly, the distribution of helpfulness criterion is consistent across all categories, which indicates that the bias is consistent across all product categories and the nature of a helpful review independent of product/category type.



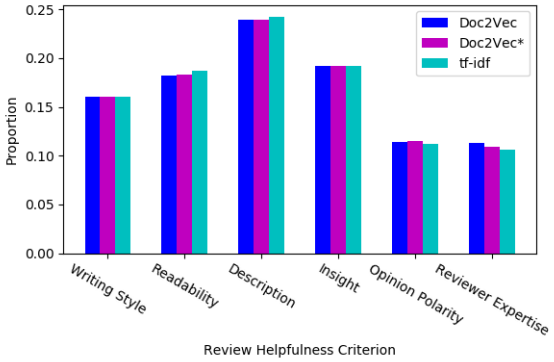
(a) Books



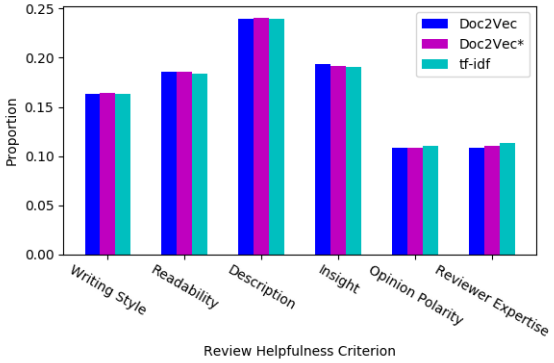
(b) Electronics



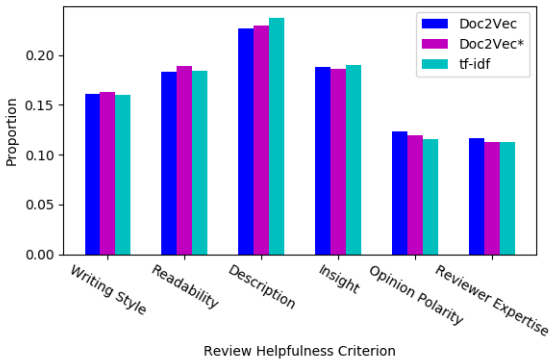
(c) Music & TV



(d) Health & Personal Care



(e) Video Games



(f) Home & Kitchen

Figure 5.7: Most helpful criterion responses from Amazon MTurk survey

5.5 Summary of Analysis

In this section, we discussed different modeling approaches to generate review embeddings and evaluated their performance. We summarize our key findings below

- **Evaluation using top-100 most similar review**
  - We looked at proportion of old reviews (target) and new review (candidates) from the top-100 most similar reviews for each model.
  - Our Doc2Vec* gave the best overall performance (greater for functional based goods than experiential based) while the Doc2Vec model gave the next best results (primarily for experiential based goods).
- **Evaluation using Amazon MTurk responses**
  - We performed a crowd-based annotation task on Amazon MTurk using 6 product categories, 3 products per categories and 5 most similar *(target, candidate)* review pairs, asking users to vote which review was more helpful to them.
  - Our Doc2Vec* model gave significantly better results than all other models, in terms of proportion of new reviews that were votes more helpful.
- **User preferential bias**
  - We also asked users to select reason why a review seemed helpful to them, based on different subjective aspects, and analyzed the voting distribution. Due to the non-uniformity of selected criteria, we showed the presence of user-preferential bias across different product categories.

6. CONCLUSION

In this thesis, we explored different Amazon product review attributes and how they influence review helpfulness scores. We also showed how the degree of influence of review attributes varies based on the nature of product categories. We extended our research by investigating two sources of intrinsic biases in review helpfulness scores (1) accumulative bias and (2) opinion bias and compared various machine learning methods to mitigate the effect of biases. We showed that our Doc2Vec* model gives the best performance by incorporating review metadata and text semantics to create review embeddings and can be used to improve helpful review recommendation.

6.1 Limitations

Some limitations of our study were (1) not considering how helpfulness scores change over a review life-cycle due to lack of data availability (2) experimenting with our modelling approaches in different review settings in order to test for generalization and (3) not exploring all possible review attributes (review readability, reviewer expertise etc.) that were discussed in previous literature, since our goal was to discuss preliminary embedding based methods for mitigating biases.

6.2 Future Work & Scope

Our work contributes to the existing research by analyzing review helpfulness in a wider context by performing cross-categorical experiments. This work also improves upon existing studies on helpfulness voting biases by proposing various modeling approaches to mitigate helpfulness voting bias in product reviews.

Based on our results, we aim to further extend this study by studying the in-

herent characteristics of the review community itself, and to understand underlying attributes of product review writers and readers which determine the likelihood of a review being up-voted or down-voted, from a cross-categorical perspective.

This work can also be extended by exploring the helpfulness 'life-cycle' of a review by mining review helpfulness scores across product categories over a larger time period and analyzing temporal deviations in helpfulness across various review features. Additionally, analyzing review voting activity over time and looking for anomalous 'spikes' in voting activity is an interesting avenue to explore. These approaches can provide stronger evidence to indicate helpfulness voting bias.

A more interesting problem that is an outcome from our work is whether we can predict the 'extent of bias' in a given product review by a framework which can predict the degree of bias in a product review and help us recommend reviews that have low bias scores (that are likely to be unbiased reviews).

REFERENCES

[1] AppSally. Buy amazon review votes, 2018.

[2] Georgios Askalidis and Edward C Malthouse. Understanding and overcoming biases in customer reviews. *arXiv preprint arXiv:1604.00417*, 2016.

[3] Steven Bird and Edward Loper. Nltk: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 31. Association for Computational Linguistics, 2004.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Pei-Yu Chen, Samita Dhanasobhon, and Michael D Smith. All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. 2008.

[6] Yuanlin Chen, Yueting Chai, Yi Liu, and Yang Xu. Analysis of review helpfulness based on consumer perspective. *Tsinghua Science and Technology*, 20(3):293–305, 2015.

[7] Laura Connors, Susan M Mudambi, and David Schuff. Is it the review or the reviewer? a multi-method approach to determine the antecedents of online review helpfulness. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–10. IEEE, 2011.

[8] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, pages 141–150. ACM, 2009.

[9] Madeline Farber. Consumers are now doing most of their shopping online. *Retail: online shopping (June8, 2016) viewed athttp://fortune.com/2016/06/08/online-shopping-increases*, 2016.

[10] 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.

[11] Tim Gollub, Erdan Genc, Nedim Lipka, and Benno Stein. Pseudo descriptions for meta-data retrieval. *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR 18*, 2018.

[12] Nan Hu, Ling Liu, and Jie Jennifer Zhang. Do online reviews affect product sales? the role of reviewer characteristics and temporal effects. *Information Technology and management*, 9(3):201–214, 2008.

[13] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics, 2006.

[14] Nikolaos Korfiatis, Elena GarcíA-Bariocanal, and Salvador SáNchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217, 2012.

[15] Laura J Kornish. Are user reviews systematically manipulated? evidence from the helpfulness ratings. *Leeds School of Business Working Paper*, 2009.

[16] Srikumar Krishnamoorthy. Linguistic features for review helpfulness prediction. *Expert Systems with Applications*, 42(7):3751–3759, 2015.

[17] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[18] Sangjae Lee and Joon Yeon Choeh. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046, 2014.

[19] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Modeling and predicting the helpfulness of online reviews. In *Data mining, 2008. ICDM'08. Eighth IEEE international conference on*, pages 443–452. IEEE, 2008.

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[22] Susan M Mudambi and David Schuff. Research note: What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, pages 185–200, 2010.

[23] Jahna Otterbacher. 'helpfulness' in online communities: a measure of message quality. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 955–964. ACM, 2009.

[24] Yue Pan and Jason Q Zhang. Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing*, 87(4):598–612, 2011.

[25] Yoon-Joo Park. Predicting the helpfulness of online customer reviews across different product types. *Sustainability*, 10(6):1735, 2018.

[26] Sarah Perez. Amazon cracks down on fake reviews with another lawsuit, 2016.

[27] SEOLIX. Amazon up review votes yes, 2018.

[28] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[29] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. Was this review helpful to you?: it depends! context and voting patterns in online content. In *Proceedings of the 23rd international conference on World wide web*, pages 337–348. ACM, 2014.

[30] Yun Wan and Makoto Nakayama. Are amazon. com online review helpfulness ratings biased or not? In *Workshop on E-Business*, pages 46–54. Springer, 2011.

[31] E Weise. Amazon cracks down on fake reviews. *USA Today (October 19), http://www. usatoday. com/story/tech/2015/10/19/amazon-cracks-down-fake-reviews/74213892*, 2015.

[32] Emma Woolacott. Amazons fake review problem is now worse than ever, study suggests. *accessed March*, 3:2018, 2017.

[33] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 38–44, 2015.