SMU Data Science Review

Volume 7 | Number 1 Article 11

Identifying Features and Predicting Consumer Helpfulness of Product Reviews

Triston Hudgins
Southern Methodist University, thudgins@smu.edu

Shijo Joseph
Southern Methodist University, sajoseph@smu.edu

Douglas Yip
Southern Methodist University, douglas.yip@gmail.com

Gaston Besanson
Begin type..., besanson@gmail.com

Identifying Features and Predicting Consumer Helpfulness of Product Reviews

Triston Hudgins[1], Shijo Joseph[1], Douglas Yip[1], Gaston Besanson[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA
[2] 08940 Cornellà de Llobregat,
Barcelona, Spain
thudgins@smu.edu
sajoseph@smu.edu
dyip@smu.edu
besanson@gmail.com

Abstract. Major corporations utilize data from online platforms to make user product or service recommendations. Companies like Netflix, Amazon, Yelp, and Spotify rely on purchasing trends, user reviews, and helpfulness votes to make content recommendations. This strategy can increase user engagement on a company's platform. However, misleading and/or spam reviews significantly hinder the success of these recommendation strategies. The rise of social media has made it increasingly difficult to distinguish between authentic content and advertising, leading to a burst of deceptive reviews across the marketplace. The helpfulness of the review is subjective to a voting system. As such, this study aims to predict product reviews that are helpful and enable strategies to moderate a user review post to improve the helpfulness quality of a review. The prediction of review helpfulness will utilize NLP methods against Amazon product review data. Multiple machine learning principles of different complexities will be implemented in this review to compare the results and ease of implementation (e.g., Naïve Bayes and BERT) to predict a product review's helpfulness. The result of this study concludes that review helpfulness can be effectively predicted through the deployment of model features. The removal of duplicate reviews, the imputing of review helpfulness based on word count, and the inclusion of lexical elements are recommended to be included in review analysis. The results of this research indicate that the deployment of these features results in a high F1-Score of 0.83 for predicting helpful Amazon product reviews.

KEY WORDS AND PHRASES: product reviews, review helpfulness, NLP, Natural Language Processing, BERT Classifiers, Naïve Bayes

1 Introduction

The rapid ascension of computer and mobile technology has thrust the world's population into the Age of Information. In 1990, it was estimated that 0.5% of the world's population was online. In a short, ten-year period, the United States saw exponential growth in internet users. By 2000, 50% of the population in the United States was utilizing the internet (Laughton, R., 2021). Figure 1 demonstrates this growth globally.

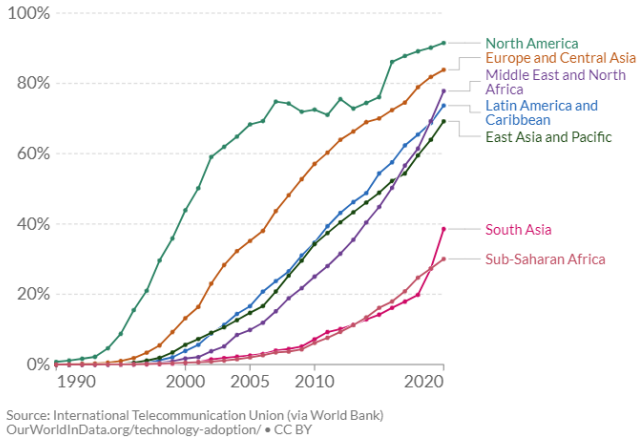

Figure 1. Population utilizing the internet by major global geographical groups from 1990 to 2020.[1]

Along with the increase of internet usage, online sales have increased. The most successful online retailers have taken advantage of this surge. They rely on the information tied to each account to tailor product recommendations and review structures that can be valuable to the potential consumer. Figure 2 demonstrates the growth of Amazon's net sales, of which at least 70% are generated through online retail sales.

[1] https://ourworldindata.org/internet

Figure 2. Amazon Historical net sales from 1996-2022. This includes Online Line Retail and Amazon Web Service sales.[2]

The volume of online reviews has increased more so as the population's internet access has been simplified through smartphones. From mid-2011 to early 2021, the percentage of the United States population that own a smart phone increased from 35% to 85%. Globally, there has been an increase in smartphone users by 49% from 2017 to 2022 (Turner, A., 2022). This has allowed users to make a review from almost anywhere. The result of this convenience is evident, with 72% of consumers reporting that they have written a review for a local business. Popular review sites like Yelp saw an increase in their revenues of 5-9% for every one-star increase (Donaker, Geoff, et al., 2019). These reviews significantly impact purchase decisions, with 94% of consumers claiming that positive reviews will make it more likely for them to purchase a product. In comparison, 92% of consumers claim that a negative review has the inverse effect, where they are less likely to utilize a product (Statista, 2022).

The last two decades have seen a rise in the availability of professional product reviews and individual review structures on common online retail sites. These reviews have become commercially important as customers pay more attention to online information such as reviews and images to reduce their purchase uncertainty (Zhao et al., 2020). A 2016 Pew Research study shows that most consumers consult customer reviews prior to making a purchase (Smith, A., & Anderson, M., 2020). Figure 3 illustrates the impact of reviews on different age groups in the US. The significance of online reviews has influenced business and marketing strategy, with many businesses encouraging consumers to fill out surveys or leave reviews through purchase incentives or reward programs. The largest online retailers understand that effective management of reviews has a positive impact on profit margins and have created an environment where reviews are an integral part of their platforms and business strategy.

[2] https://www.marketplacepulse.com/stats/amazon/amazon-net-sales-94

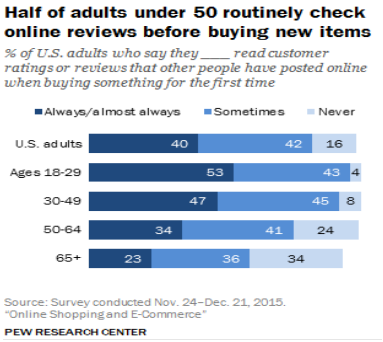

Figure 3. Influence of online review before purchasing products by age group.[3]

Determining the reliability of online reviews is not without risk. False reviews and review spamming are strategies used to either skew a competitor's perception or boost a competing product's rating. These reviews can impact a user's decisions on a platform. Both users and distributors can benefit from methods that eliminate harmful reviews from product recommendations. Accurate product reviews can give a user the power of information to influence a purchase decision and can give online retailers a level playing field among competitors.

Natural Language Processing (NLP), a field that "covers any kind of computer manipulation of natural language" (Bird, 2016), can deploy statistical strategies to interpret and analyze unstructured product review text. The concept is a complex task for a machine to derive insight from the unstructured nature of human language, especially when considering the tone behind the text. With machine learning advances, it is now increasingly possible to detect the nuances of speech, and the use cases of NLP have grown substantially over the last fifteen years. There are two major aspects of NLP: natural language interpretation (NLI) and natural language generation (NLG). Interpretation, or understanding, is the flow of information from human to machine, while generation is from machine to human. This paper will focus on an NLI use case to identify helpful online product reviews through various tools like Naïve Bayes and Bidirectional Encoder Representations from Transformers (BERT). Large language models, like BERT and (Generative Pre-trained Transformer) GPT utilize transformer architecture, which is a type of deep neural network that allows for parallel processing of input sequences, making it highly efficient for processing large amounts of text data.

[3]https://www.pewresearch.org/internet/2016/12/19/online-shopping-and-e-commerce/

It is necessary for a machine to apply transformations on text data if it is to understand and model human language. The text transformation tools in Table 1 are some of the tools available when creating NLP algorithms. Each transformation returns an array or matrix of statistical counts.

Table 1. Word Vector Representation

Word Vector Model	Word Vector Application
Tf-Idf	**T**erm **f**requency-**i**nverse **d**ocument **f**requency weight gives an indication of word importance in a sentence or document in relation to the entire corpus.
BoW	**B**ag **o**f **W**ords analyzes text/document based on word count disregarding word order and grammar.
Count-Vectorizer	Converts text into a matrix of token counts.

These tools can be used to identify key phrases or patterns in poor and helpful reviews and identify spam reviews. The goal of this study is to predict helpful product reviews and enable strategies to moderate a user review post to improve the helpfulness quality of a review**.**

2 Literature Review

2.1 Helpfulness of reviews

Most online retailers have created a system that allows other users to flag reviews as helpful. The ample amount of review data with feedback has allowed for multiple in-depth studies. Heng et al. (2018) noted that many studies have examined the factors influencing review helpfulness. These included review ratings, the length of words, positive and negative sentiments, and emotions that were used to identify the helpfulness of a review. The experience shared by other reviewers, in conjunction with helpfulness vote counts, further informs and supports a consumer's decision to proceed to purchase the product.

Amazon was among the first to recognize the value of multi-user engagement by soliciting feedback for reviews through developing a "Was this review helpful to you?" option that tallies and records additional details of review helpfulness. Zhu et al. (2020) confirms that a reader determines the credibility of helpfulness reviews based on helpful votes it receives from other users and deems reviews with helpfulness votes more credible than those with no votes. However, in their study, they recognized that a considerable number of reviews do not receive helpful votes. As a result, sourcing quality data for helpful review to assess the review quality of the product is challenging (Tang et al., 2013). Indirect methods to predict helpfulness focus on features such as the review content, product features, and reviewer features. Bilal et al. (2020) utilizes the length of review, rating, readability, polarity, and subjectivity to assess a product

review's helpfulness. As such, further examination of how helpfulness can be extracted directly from votes counts or indirectly from details from product reviews and ratings shall be reviewed.

Studies have analyzed product review word length to gain an understanding of a review's impact on consumer decisions. Mundambi et al, (2010) uses the Tobit Regression method to determine review helpfulness through the utilization of word count, total votes, and product ratings. The research found that a lengthier review was more helpful for moderate-rated products. However, the study also suggests that caution is to be exercised on the extreme ratings. The review rating range is bound to a response limit due to certain factors such as: whether the product was essential, the selection was biased based on extreme views, and so forth. These factors will have no impact by the review response. In their study, Gamazu et al. (2021) employed Pearson's R correlation analysis to examine the relationship between the number of annotators and the number of helpful votes received for a given product review. The results indicated a positive correlation, suggesting that longer reviews tend to receive more helpful votes from other users. As such, the lexical analysis of reviews will need to be considered when identifying review helpfulness. Gamazu et al. (2021) also highlighted in their multi-document summarization study the usage of the length, sentiment, and annotation of a helpful sentence, as well as an Amazon Web Services (AWS) sentiment analysis tool as a basis to developing their model to review helpfulness of their reviews.

The helpfulness of a product reviews refers to the extent to which it aids consumers in making purchase decisions and is categorized into three components: "(1) perceived source credibility, (2) perceived content diagnosticity, and (3) perceived vicarious expression." Li et al. (2013). The focus category of product review helpfulness for this study will be based on perceived content diagnostic. Perceived content diagnostic (problem solving) refers to the provision of advice for solving problems. In the context of product reviews, the review content should provide comprehensive information about the pros and cons of a product to aid the consumer in making a purchasing decision.

2.2 Product reviews

With the abundance of information available, product reviews have become a crucial factor in determining consumer engagement and purchase decisions in the market. The availability of data and the ease of access to information has heightened the significance of product reviews in shaping consumer behavior. Gamzu et al. (2021) demonstrate that the desire for a more streamlined shopping experience among consumers has resulted in a migration towards online shopping. Hari (2019) supports this trend with correlations between the shift in consumer engagement and the shift towards online purchasing. To capitalize on this trend, many organizations have strengthened their communication outlets to influence consumer purchasing decisions (e.g., detailed product reviews). Hari (2019) also highlights that in-store shopping behavior is influenced by research conducted online, with 82% of smartphone users checking competing products on their devices and 45% reading reviews prior to making a

purchase. This emerging consumer engagement dynamic makes product reviews critical to both the consumer and the provider.

Askalidis et al. (2016) applied the SciPY curve fit function in Python to investigate the relationship between the volume of consumer reviews and their impact on the likelihood of purchase. The results indicate that product reviews with less than 3.5 average stars had a conversion rate of 324%, compared to a conversion rate of 135% for products with an average rating of greater than 3.5 stars. The work provides compelling evidence of the value of online reviews and the impact of product reviews on consumer purchase intent. However, it is crucial to consider the potential for fake reviews and the impact they may have on consumer decision-making. Thota et al. (2022) and Mohwesh et al. (2021) have conducted studies and surveys on the topic of fake reviews, offering insight into how the prevalence of fake reviews can alter consumer perceptions of products.

2.3 Modeling Product Reviews

The utilization of Naïve Bayes in machine learning text classification is well-established. Naïve Bayes is commonly used as a baseline model due to its simplicity and ease of implementation. "Simultaneously, it has earned the dubious distinction of placing last or near last in numerous head-to-head classification papers (Yang & Liu, 1999; Joachims, 1998; Zhang & Oles, 2001) due to the model's severe assumptions" (Rennie et al. (2003)). Naïve Bayes uses a probabilistic classification model that relies on prior probabilities and assumes independence between variables. Unbalanced data can significantly hinder the performance of a standard multinomial Naïve Bayes strategy and can cause the algorithm to favor a class. This study initially utilized a Complement Naïve Bayes (CNB) model as described in Rennie et al. (2003). The CNB adjusts the probability weights of a class by using all training data not in the class. This strategy significantly improves model performance when compared to a standard Multinomial Naïve Bayes model. CNB has been shown to approach the accuracy of a more complex support vector machine model.

The study of NLP is constantly producing modified techniques to aid machines in understanding the nuances and complexities of language. These techniques fall into a sub-category of NLP that is known as NLU (Natural Language Understanding). The key strategies used to predict review helpfulness include the use of persuasive language, emotional language, and specific details about the product or service being reviewed. In addition, factors like review length, reviewer expertise, and overall rating can also be used to predict helpfulness. Among the many tools available is Part-of-Speech tagging which is "the task of assigning words to their respective parts of speech. It can either be basic tags such as noun, verb, adjective, adverb, and preposition, or it can be granular such as proper noun, common noun, phrasal verb, verb, and so on." (Ganegedara, Thushan, 2018). By tagging words accordingly, it may become easier to establish correlations in sentence meaning and behavior. These insights can assist with tasks such as sentiment analysis, text classification, and machine translation.

In 2018, Google introduced a new NLP technique called BERT (Bidirectional Encoder Representations from Transformers). BERT is a significant advancement in the field of NLP due to its superior performance compared to other existing models. Unlike traditional NLP techniques, BERT is designed to be pre-trained, only requiring fine-tuning for specific tasks. The tool has been used for applications such as question-answering and sentiment analysis (Devlin et al., 2019). The pre-training aspect of BERT not only results in accelerated development time, but also in a reduction of required data. Furthermore, BERT is publicly accessible as an open-source model, making it a highly valuable tool for a range of NLP applications.

BERT classifiers present a solution to the overwhelming nature of online reviews by differentiating between reviews that are informative and those that are subjective opinions. In doing so, the consumer is provided with a more accurate evaluation of the product when purchasing. Online customer reviews significantly impact purchasing decisions for prospective customers and help businesses manage customer needs (Cheng et al., 2016). Wu (2020) employed a count representation of helpfulness in a BERT classifier, but Xu (2020) found this method to be insufficient for accurate classification. This study will utilize a refined BERT model that incorporates contextual token embedding, providing a comprehensive approach to the classification of review helpfulness.

BERT employs two NLP techniques, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The MLM task trains the model by hiding words in a sentence and having the model predict the contextual meaning of the masked word. The NSP task trains the model to determine the relationship between two sentences (Devlin et al., 2019). Pre-Training data in BERT comes from multiple data sets: BooksCorpus, English Wikipedia, and Billion Word Benchmark (Devlin et al., 2019). The architecture for both pre-training and fine-tuning is identical, with the pre-training model parameters serving as the starting point for the fine-tuning model. Additionally, classification (CLS) and separator (SEP) tokens are added to the beginning and end of every input, respectively, to serve as separators, as demonstrated in Figure 4.

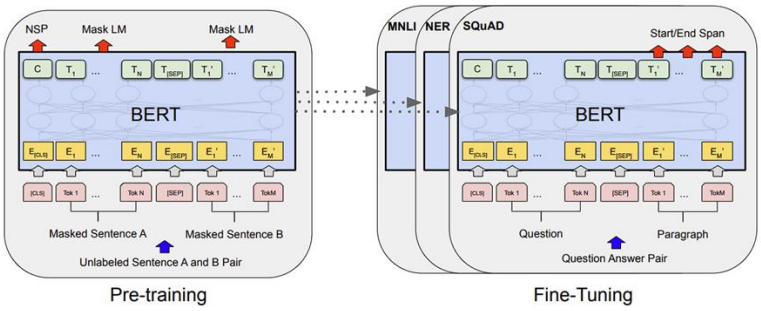

Figure 41. Fine tuning method of BERT Classifiers.

BERT is integrated into the tensor2tensor library in Python. There are two configurations for BERT; $BERT_{base}$ and $BERT_{Large}$. The base model has 12 layers with 768 hidden layers and 12 attention heads, whereas the large model has 24 layers with 1024 hidden layers and 16 attention heads. The following are the steps required to fine-tune BERT:

- The data must be transformed into a specific format by tokenizing it with the BERT tokenizer.
- Special tokens must be appended to the data. The [CLS] token must be added to the beginning of the text, and the [SEP] token to the end of the text.
- A sequence length must be used and plan to look to use a 512-sequence length.
- Adjust the length of the text by padding or truncating it to match the specified sequence length.
- Create attention masks to differentiate between real and padded tokens.

Once the previous steps have been completed, the data is then used to fine-tune the BERT base model for classification. The data is consolidated into a tensor data set and then divided into a training set (90%) and a testing set (10%). BERTForSequenceClassification model, which is a BERT model augmented with a classification layer, is employed in this study. As recommended by Devlin (2019), a batch size of 8 and a sequence length of 512 were used and depicted in the following table.

Table 2. The Hyperparameters for Fine-tuning BERT

Sequence Length	Batch Size	Epochs	Learning Rate	Episolon (eps)
64	32	4	$2e^{-5}$	$1e^{-8}$
128	32			
256	32			
320	32			
384	16			
512	8			

The output of the study will categorize a review as either positive (1) or negative (0) based on whether the review is deemed helpful or unhelpful. Upon completion of the training and testing process, evaluation metrics can be obtained to assess the accuracy and performance of the classifier.

Despite extensive research into classifying product reviews, they remain challenging to discern. A 1-star rating can indicate a consumer's negative opinion as well as factual imperfections of the product. The subjectivity of star ratings makes them prone to ambiguity (Kouvaris et al, 2018). If a product review is given by social influence, it may elicit unrecognized needs or wants. Thota et al, (2022) states the authenticity of online reviews is increasingly being impeded by fake reviews that misrepresent the

product's quality. Therefore, detection of non-relevant reviews is necessary to ensure proper representation of the product and its review.

To understand the quality of the models, the process of creating a training, testing, and validation data set are utilized. A count vectorizer is applied to the text field to create a matrix of token counts. The model is fit and then the testing data is run against the training data to produce scores and a confusion matrix.

2.4 Recommendation Systems (RS)

The prediction of review helpfulness has been the main subject of many studies. Heng et al. (2018) highlighted another purpose of moderating review helpfulness: the optimization of a company's product recommendation system.

Product recommendations have become a vital feature of e-commerce, enabling personalized suggestions based on user preferences (Kouvaris et al., 2018). Recommendation systems originally provided product suggestions through leveraging similarities among a users' purchasing habits. Modern systems can scrape account data and find related topics through various social media platforms. These systems not only aid users by reducing manual searches for related items but also allow retailers to adjust marketing strategy in real-time. Given that consumer trust and risk are inversely correlated for recommendation systems (Kim et al., 2008), it is important to minimize the influence of poor reviews to improve the trust towards helpful product reviews.

Multiple approaches can be used to build recommendation systems. These systems are typically built on the information gathered from consumer profiles, past purchases, demographic data, and if a social network is integrated, the purchases of a consumer's social media connections (Li & Karahanna, 2015). While the significance of recommendation systems is acknowledged in this study, the primary focus will be on the identification and prediction of helpful and informative product reviews.

3 Design and Methods

The following section will provide details on the data set utilized for this study and outline methods to evaluate model results. These steps consist of data preparation and cleaning, Exploratory Data Analysis (EDA), and the development of models.

3.1 Data

This study utilizes the Amazon product review data set from May 1996 to October 2018 (Ni et al., 2019). The data set contains about 233.1 million rows, 29 unique product categories (e.g., Office Products, Pet Supplies, Grocery and Gourmet Food, etc.), and 11 columns described in Table 3.

Table 3. Description catalog of Amazon Metadata

Description of column	Example
Rating of the Product	**5**
Time of the review (raw)	**08 12, 2015**
ID of the reviewer	**A3QHVBQYDV7Z6U**
ID of the product (ASIN)	**0000013714**
A dictionary of the product Metadata	**"Format" is "Hardcover"**
Name of the Reviewer	**The Nana**
Text of the review	**Gift for college student.**
Summary of the review	**This is the tea I remembered!**
Helpful votes of the review	**14**
Time of the review (unix)	**1476316800**
Image that users posted of the received product	**Image of product if applicable**

This research utilizes the category "Grocery and Gourmet Food", a subset containing over 1,140,360 reviews.

3.2 Data Cleansing and EDA

A thorough Exploratory Data Analysis was conducted to uncover relevant patterns and features. The data was cleansed to achieve higher accuracy in these classification tasks. Within the data set, cells with NA were imputed with an average or zero depending on the data format and coding requirements. The following strategies were then deployed to produce the final data set for modeling.

Duplication of Reviews

The size of the data set warranted a check for duplicate reviews. The variables utilized for this check included the reviewer, review time, and review text. Table 4 demonstrates that approximately 15.25% of the data set was duplicate reviews. All the identified duplicate reviews have the identical time stamp, text, and reviewer. This finding suggests that bots may have been utilized to post the same review repeatedly for different products. As such, the duplicate reviews were removed from the data to avoid potentially biased results.

Table 4. Number of Original Reviews

Review	Total Reviews	% of Reviews
Original	969,400	84.75
Duplication	174,460	15.25
Total	1,140,360	100%

Unbalanced data set

Examining helpful vote counts revealed an imbalance in the data. The helpful vote column shows 13.8% of the reviews received a helpful vote count greater than 1. Table 5 shows an unbalanced review sentiment where 70.8% of the reviews have a rating of 5. Unbalanced data sets can create a statistical challenge for accurate modeling. To overcome this, random samples of both helpfulness groupings were pulled to create a balanced data set for modeling purposes.

Table 5. Distribution of reviews by overall rating

Rater Rating	Total Reviews	% of Reviews
5 ☆☆☆☆☆	686,760	70.8%
4 ☆☆☆☆	129,991	13.4%
3 ☆☆☆	71,435	7.4%
2 ☆☆	37,538	3.9%
1 ☆	43,631	4.5%
Total	969,400	100.0%

Helpfulness by rating sentiment

Table 6 shows the number of helpful reviews by rating over-indexes for 1-star reviews. This may suggest that online reviewers are utilizing this information to substantiate their decision not to purchase the product.

Table 6. Distribution helpfulness by rating category

Rater Rating	Total Reviews	% helpful	% not helfpul
5 ☆☆☆☆☆	686,760	13.4%	86.6%
4 ☆☆☆☆	129,991	14.1%	85.9%
3 ☆☆☆	71,435	16.0%	84.0%
2 ☆☆	37,538	20.1%	79.9%
1 ☆	43,631	32.8%	67.2%
Total	969,400	14.8%	85.2%

Statistical Length of reviews

A two-sample t-test was performed to compare the average number of words between helpful and non-helpful reviews. The results, at a 95% confidence interval, yield a P-value ≈ 0, indicating that there is overwhelming evidence of a difference in mean review length between the two conditions. The average word count of the helpful reviews was 89 words, which was three times longer than the average unhelpful review word count. Figure 5 provides additional evidence that helpful reviews are longer than non-helpful reviews as indicated by the steeper and thicker Boxen plot. A Pearson

correlation was performed with the continuous variables. The correlation between helpfulness and review length observed a positive correlation of 0.34.

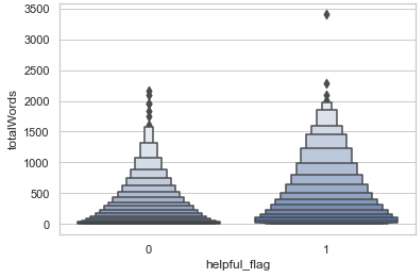

Figure 5.2 Pyramid Graph of Total Words by Reviews that are helpful (1=Helpful)

Drawing on the discoveries discussed above and common knowledge of what constitutes a comprehensive review, total word cut-off lengths were determined to aid in imputation. In Figure 6, section A of the graph contain reviews with fewer than 25 words that were classified as unhelpful, while section B with reviews over 125 words classified as helpful. Mean word counts for each response were used to identify the ranges to impute the response variable. The data set was revised to feature this imputed response variable. The best F1-score was utilized to refine the cut-off points, leading to the optimization of the imputed values. Figure 6 presents the cumulative percentage of each response, providing additional support for the fine-tuning efforts of the response variable.

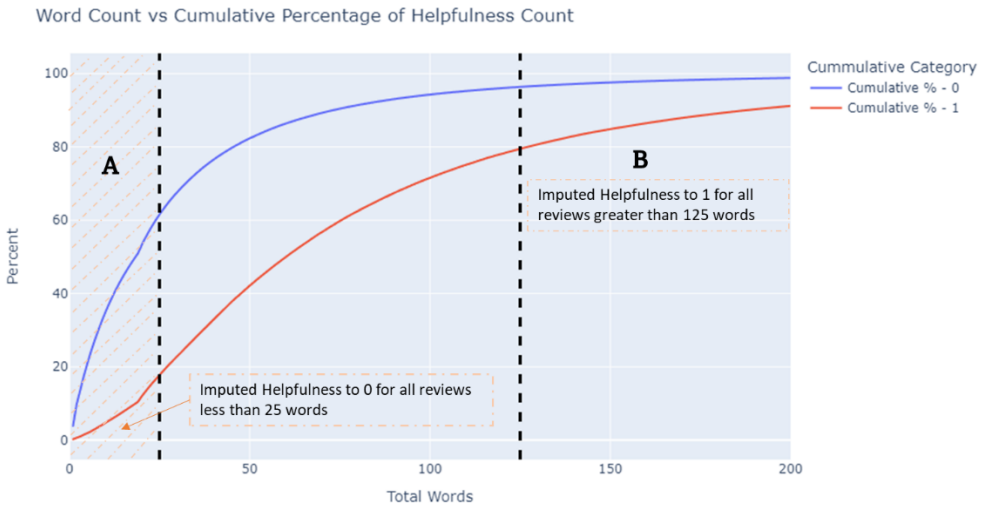

Figure 63. Cumulative Percentage of Review total word count by response (1=Helpful)

Relevant Features

All columns excluding the "review text" and the response variable, "helpfulness", were determined to be out of the scope of this study. All reviewer identification columns were discarded to avoid giving an unfair weight to users who had a higher review count.

Additional Parts of Speech (POS) counts

Rather than using direct lexical and semantic methods to address the characteristics of the review text, Table 7 outlines the additional numeric columns that were added based on selective NLP characteristics and the count for each review. Natural Language Toolkit (NLTK) tags were used to identify nouns, adjectives, verbs, and adverbs.

Table 7.1 Parts of Speech count columns added to data set.

Lexical columns	Definition and detail
Stop Word Count	Counting words that have been defined as common words (e.g., is, I, there)
Numeric Count	Counting numeric text (e.g., 1234)
Upper Case Word Count	Counting words with upper case
Capitalize Word Count	Counting words that are all capitalized
Noun Count	Counting words that are specific or set of objects (e.g., people, places, or things)
Verb Count	Counting words that convey action (e,g. to be, to have, etc.)
Adjective Count	Counting words that describe nouns (e.g., red, big, fury, etc.)
Adverb Count	Counting words that modify verbs (e.g., clearly, often, very, etc.)
Conjunction Count	Counting words that connect words and phrases (e.g, and, because, but, etc.)

3.3 Evaluation Process

A random sample of 50,000 rows was generated into a new dataframe to optimize model prediction testing. This sample was split 80:20 between training and test sets, respectively. Each model demonstrated in this study will use the training set to predict the classification response, which will then be compared to the respective test set.

The performance of the classification models will be assessed through the utilization of a confusion matrix to measure whether the models have accurately predicted the results of the training set. The models will predict and categorize product reviews as either helpful or unhelpful. Performance metrics noted in Table 8 will be utilized to measure and compare the prediction of product review helpfulness. Terms like True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are used to measure classifcation models.

Table 8.2 Metrics used to measure performance of classification models prediction.

Metric	Formula	Definition
Weight F Measure	$2 \text{ x } \frac{Recall\ x\ Precision}{Recall+Precision}$	A single score that harmonizes mean of precision and recall
Accuracy	$\frac{TP+TN}{(TP+TN+FP+FN)}$	Score to the actual value
Precision	$\frac{TP}{(TP+FP)}$	Positive predictive measure deciding the model confidence level
Recall	$\frac{TP}{(TP+FN)}$	Measure of completeness of the results

3.4 Method and Framework of Product Review Helpfulness

Following the literature investigation of potential prediction models, two techniques were selected for modeling Amazon review helpfulness: Naïve Bayes and BERT Classification. Figure 7 pragmatically depicts the testing process that will ensure that each stage evaluates the F1-Score, precision and recall results for each model data set combination.

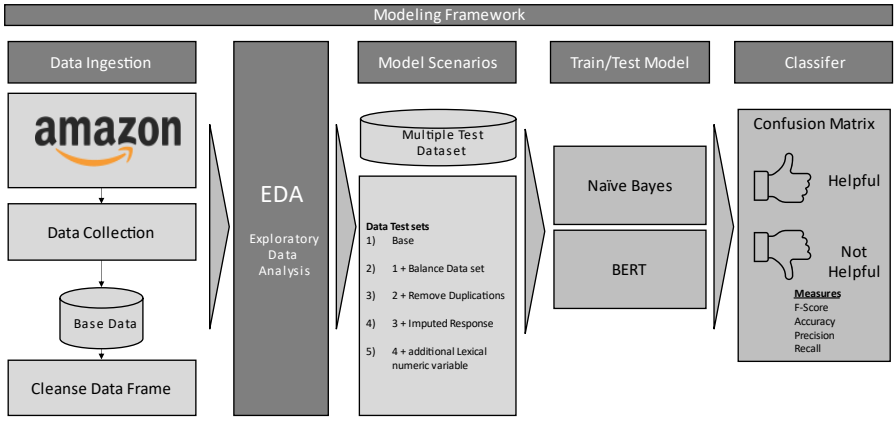

Figure 7.4 Modeling Framework of Product Review Helpfulness.

The modeling approach will begin with a base case and incrementally modify the data to support the understanding of each change as well as determine what features are needed for companies to refine how reviews are written to improve review helpfulness.

To avoid overfitting, a 70:30 train-test split was applied to both the Naïve Bayes and BERT classifier models. The performance of the predictive model was evaluated using cross-validation, a technique that involves repeating the process several times with different subsets and averaging the results to estimate the model's performance on new data.

4 Results

To identify the most effective model for predicting helpful reviews, a practical approach was implemented, comparing Naïve Bayes and BERT Classification. Table 9 presents the outcomes of testing the baseline data for both models. The data set is considerably skewed towards unhelpful reviews, leading to suboptimal performance in predicting helpfulness. The precision and recall values were particularly weak, resulting in an F1-Score of 0.51. Given these findings, it was deemed necessary to rebalance the data set to bolster the precision and recall of the models.

Table 9.3 Base review data set results

Model	Accuracy	Precision	Recall	Macro F1-Score
Naïve Bayes Model	85%	0%	0%	0.46
BERT Model	86%	46%	6%	0.51

The balanced data improved the overall F1-Score (as shown in Table 10) with both model's recall improving from <10% to >75%. Given that the accuracy and F1-Score

are trending on an average of 0.7, this will be the base model to compare other feature modifications.

Table 104. Balanced data set (50% helpful/50% unhelpful review) results

Model	Accuracy	Precision	Recall	Macro F1-Score
Naïve Bayes Model	68%	62%	88%	0.67
BERT Model	72%	70%	75%	0.72

The EDA revealed a substantial presence of duplicate reviews, as depicted in Table 4. In response to this discovery, it was determined that all duplicate reviews would be considered erroneous or fabricated and were thus removed from the study. The subsequent analysis of the revised data, as displayed in Table 11, did not yield improved results. Given that the F1-Score and recall performance decreased, the duplicate reviews can be inferred biased towards helpful. Whether the duplicative reviews are favorable or unfavorable, the integrity of the data must be priority. As such, removal of these reviews is necessary to ensure accurate prediction of review helpfulness.

Table 11.5 Removal of duplicative reviews (same exact fields, in user, timestamp and review) to the balanced data set results

Model	Accuracy	Precision	Recall	Macro F1-Score
Naïve Bayes Model	63%	86%	73%	0.66
BERT Model	70%	69%	71%	0.70

Further analysis revealed a disparity between the word counts of helpful reviews vs non-helpful reviews. The response variable was imputed by setting all reviews with a minimum word count of 25 as unhelpful and setting the max word count for reviews to 125 as helpful reviews as shown in Figure 7. The introduction of the imputed response variable, as shown in Table 12, yielded an improvement to the F1-Score for both models.

Table 12.6 Balanced, no duplicative review data set with imputed response variable results.

Model	Accuracy	Precision	Recall	Macro F1-Score
Naïve Bayes Model	75%	75%	75%	0.74
Bert Model	80%	80%	80%	0.80

The use of Part-of-Speech (POS) tagging facilitated the creation of several lexical property counts, which were documented in Table 6. The conversion of the unstructured review data into numeric lexical variables resulted in improvements to the F1-Score, as indicated by the comparison of results from Table 13 versus Table 12. This would suggest that taking the unstructured data of the review and converting it to a numeric representation is extremely valuable when predicting review helpfulness.

Table 13.7 Adding Parts of Speech variables to all user reviews.

Model	Accuracy	Precision	Recall	Macro F1-Score
Naïve Bayes Model	78%	71%	84%	0.78
Bert Model	83%	77%	89%	0.83

The results of this research indicate that the helpfulness of Amazon product reviews can be predicted with a high degree of accuracy. The improvement from the base model, which resulted in a poor confusion matrix, to the final model, which involved multiple data manipulation techniques to achieve F1-Scores of 0.83, is validation for the research capabilities. Not only was the success of this pragmatic approach evident through the improving F1-Scores, but the research also identifies features that can assist with improving review's helpfulness. These features include the removal of duplicate reviews, imputing review helpfulness based on word count and utilizing POS to add lexical elements to all reviews.

5 Discussion

This section will provide interpretations of the analysis and findings that were drawn from the research for predicting the helpfulness of Amazon product reviews.

5.1 Result interpretation

Online retailers that rely on review feedback should consider implementing a minimum word count requirement based on one of the key findings of this study. EDA showed a heavily skewed data distribution, and further statistical analysis revealed a significant difference in the mean word count of reviews across different categorical responses. Manual inspection of product reviews also revealed that one-word reviews received helpful votes, while more detailed reviews did not, suggesting that helpfulness is subjective. Thus, implementing a minimum word count requirement can encourage reviewers to provide more informative and detailed reviews, resulting in thoughtful and insightful reviews for other users. Additionally, setting a word count minimum can help filter out low-quality, unhelpful reviews that provide little useful information, prevent spam reviews, and discourage fake reviews that are often short and generic. This feature can foster a sense of community among reviewers.

Furthermore, companies can take review requirements to the next level by setting minimum requirements for the type of word structure used in the review. The application of syntactical and lexical features was found to improve the helpfulness of reviews, suggesting that detailed reviews containing more adjectives and verbs are more influential in consumers' purchasing decisions. However, implementing such a specific feature may also discourage users from posting reviews, and more consumer

research should be conducted to ensure that this feature does not negatively impact web platform user engagement.

5.2 Interesting findings

Another critical finding of this study was the need for Amazon to establish more effective monitoring gatekeepers to eliminate fake reviews generated by bots. Approximately 15% of the data was identified as duplicate reviews, written by the same user, with the same timestamp and content for different reviewed products. While the removal of these reviews did not improve the statistical prediction of helpfulness, the quantity of duplicative reviews underscores the importance of implementing measures to prevent distorted product reviews. Such measures can improve the consumer review experience by removing potential biases.

5.3 Research Limitations

The selected data set for the Amazon Office Products, Pet Supplies, Grocery and Gourmet Food product reviews contained over one million rows. For the purpose of this research, the models were limited to 50,000 rows of data due to high computational requirements for BERT and to avoid model failure. Furthermore, the analysis was limited to these specific categories, with no testing or comparison performed on other categories. As such, it is recommended to exercise caution when generalizing the core findings of this study to other categories.

It should also be noted that the data set was generated prior to the COVID-19 pandemic, making it possible that shifts in behaviors surrounding product reviews, both during and after the pandemic, may not have been captured in this research. The assumption is online consumer purchasing behavior has changed during and post pandemic.

Throughout the research, the BERT model consistently outperformed the Naïve Bayes model F1-Score on average by 6%. However, industry computational cost implications should be considered as the Naïve Bayes models were completed in <1 minute vs the BERT Model which took ~10 minutes to complete for a sample set of 50,000 rows. A cost benefit analysis will need to be completed to understand whether the improved predictive results are worth the cost. Future model can evaluate ensemble models to optimize performance and cost utilizing both Naïve Bayes and BERT Models.

5.4 Future research

It is possible that the selected features within the data set limited the effectiveness of the models. Future studies could benefit from expanding to utilize all features within the data set to include users, timestamp, and rating. Additionally, the findings from this study on the Grocery and Gourmet Food category could be applied to other Amazon product categories, such as Pet Supplies, Automotive Products, and Office Products, to further validate the research results.

Future studies may enhance this analysis by incorporating semantic and discourse analysis techniques for Natural Language Processing, as the current study did not incorporate methods to extract meaning from the reviews. Vertical integration studies, such as recommendation systems, can test new features for review helpfulness to see if improved reviews lead to better user response. It may also be beneficial to incorporate new or existing APIs such as GPT to model deep learning on the reviewed text. If these features prove to be effective, organizations like Amazon and Netflix could potentially increase revenue and profits through improved recommendation systems.

6 Ethics

While it is acknowledged that having more helpful reviews can enhance the consumer experience, caution needs to be exercised due to the potential for review manipulation that targets specific markets which may led to bias recommendation systems. Reviews can be designed to influence user behaviors and preferences, raising questions about whether users are being coerced into making particular choices. Additionally, user autonomy is also a concern, as user reviews can guide recommendation systems towards specific choices, potentially limiting exposure to alternative options. The topic of target marketing should be considered for future research when assessing the consumer helpfulness of user reviews.

Data collection for this study did not require any human subject testing nor did the data expose any personal identification information. The work that is pertained in this research was not impacted by political, social, or economic biases. However, an ethical consideration arose during the imputation of the helpfulness response variable. As the prediction accuracy of the models improved, the question of what constitutes a helpful review was raised. For example, a simple review like "Good Product" may not provide enough detail to be considered helpful, but some users may be seeking binary answers to influence their purchasing decisions. The imputing of a known response variable is subject to debate and requires further analysis of user expectations regarding review helpfulness.

7 Conclusion

The BERT prediction model, trained on a data set with imputed response variables based on word count, removal of duplicate reviews, and added lexical properties, yielded the most promising results. The model achieved an impressive F1-Score of 0.83 in predicting the helpfulness of Amazon reviews, and the fine-tuning of data consistently improved the F1-Score in both Naïve Bayes and BERT models used in the research. This indicates that the applied data modification methods consistently enhanced the accuracy of helpful review predictions.

While the primary objective of the study was to predict the helpfulness of Amazon reviews, certain features were identified that could improve the user review experience. For instance, setting a minimum word count for reviews could foster more detailed product content, and the continuous removal of duplicative or spam reviews could be implemented by companies as a backend process to enhance the consumer experience. Ultimately, improving the user review experience is a crucial consideration in ensuring that product reviews are helpful to consumers.

Acknowledgment A special thanks is extended to Gaston Besanson for his willingness to support the research with predicting Amazon Helpfulness Review utilizing NLP and ML techniques. Thanks, is also extended to Dr. Jacquelyn Cheun and the rest of the SMU MSDS faculty for their guidance and instruction throughout the program.

References

Askalidis, G., & Malthouse, E. (09 2016). The Value of Online Customer Reviews. ACM Conference on Recommender Systems, doi:10.1145/2959100.2959181

Barnes, W. R. (2019). The Good, the Bad, and the Ugly of Online Reviews: The Trouble with Trolls and a Role for Contract Law After the Consumer Review Fairness Act. Georgia Law Review, 53(2), 549–612.

Bilal, M., Marjani, M., Lali, M. I., Malik, N., Gani, A., & Hashem, I. A. T. (2020). Profiling users' behavior, and identifying important features of review "helpfulness''. IEEE Access, 8, 77227–77244.

Bird, S. (2016). Natural language processing with python. O'Reilly Media.

Chen, A., Lu, Y., & Wang, B. (2017). Customers' purchase decision-making process in social commerce: A social learning perspective. International Journal of Information Management, 37(6), 627–638.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pretraining of deep bidirectional Transformers for language understanding. ACL Anthology. Retrieved October 28, 2022, from https://aclanthology.org/N19-1423/

Donaker, Geoff, et al. "Designing Better Online Review Systems." Harvard Business Review, 22 Oct. 2019, https://hbr.org/2019/11/designing-better-online-review-systems.

Laughton, R. (2021, September 8). A history of online reviews. ReviewInc. Retrieved October 2, 2022, from https://reviewinc.com/2021/09/07/a-history-of-online-reviews/

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Applied Sciences, 10(21). doi:10.3390/app10217748

Gamzu, I., Gonen, H., Kutiel, G., Levy, R., & Agichtein, E. (2021). Identifying helpful sentences in product reviews. NAACL 2021. Ανακτήθηκε από https://www.amazon.science/publications/identifying-helpful-sentences-in-product-reviews

Ganegedara, Thushan. Natural Language Processing with TensorFlow : Teach Language to Machines Using Python's Deep Learning Library, Packt Publishing, Limited, 2018. ProQuest Ebook Central http://ebookcentral.proquest.com/lib/southernmethodist/detail.action?docID=5405681.

Heng, Y. et al. (2018) "Exploring hidden factors behind online food shopping from Amazon Reviews: A topic mining approach," Journal of Retailing and Consumer Services, 42, pp. 161–168. Available at: https://doi.org/10.1016/j.jretconser.2018.02.006.

Hari, Harinder. (2019). Customer Engagement Influences On Buying Decision in an Online Context -A Review. Vol-22-. 77-83.

Kim, D. J., Ferrin, D. L. & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decision Support Systems, 44(2), 544-564. Retrieved from https://doiorg.ezproxy.library.wur.nl/10.1016/j.dss.2007.07.001

Kouvaris, Peter; Pirogova, Ekaterina; Sanadhya, Hari; Asuncion, Albert; and Rajagopal, Arun (2018) "Text Enhanced Recommendation System Model Based on Yelp Reviews," SMU Data Science Review: Vol. 1: No. 3, Article 8.
Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/8

Li, M. et al. (2013) "Helpfulness of online product reviews as seen by consumers: Source and content features," International Journal of Electronic Commerce, 17(4), pp. 101–136. Available at: https://doi.org/10.2753/jec1086-4415170404.

Li, S. S. & Karahanna, E. (2015). Online Recommendation Systems in a B2C E-Commerce Context: A Review and Future Directions. Journal of the Association for Information Systems, 16(2), 72- 107. Retrieved from https://www.semanticscholar.org/paper/Online-RecommendationSystems-in-a-B2C-E-Commerce-A-LiKarahanna/a034274d13a2ac9fa3d1e56248a80c9ba2a877a2

Mohawesh, R., Xu, S., Tran, S. N., Ollington, R., Springer, M., Jararweh, Y., & Maqsood, S. (2021). Fake Reviews Detection: A Survey. IEEE Access, 9, 65771–65802. doi:10.1109/ACCESS.2021.3075573

Mundambi, Susan, Schuff, David (2010) What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. MIS Quarterly , March 2010, Vol. 34, No. 1 (March 2010), pp. 185-200

Ni, J., Li, J., & McAuley, J. (2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 188–197. doi:10.18653/v1/D19-1018

Ostheimer, M., & Atleson, M. (2022, June 9). FTC puts hundreds of businesses on notice about fake reviews and other misleading endorsements. Federal Trade Commission. Retrieved

December 2, 2022, from https://www.ftc.gov/news-events/news/press-releases/2021/10/ftc-puts-hundreds-businesses-notice-about-fake-reviews-other-misleading-endorsements

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 616-623).

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S.-G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771. doi:10.1016/j.jretconser.2021.102771

Smith, A., & Anderson, M. (2020, May 30). 2. online reviews. Pew Research Center: Internet, Science & Tech. Retrieved October 2, 2022, from https://www.pewresearch.org/internet/2016/12/19/online-reviews/

Statista. (2022, August 11). US smartphone ownership 2021. Statista. Retrieved October 2, 2022, from https://www.statista.com/statistics/219865/percentage-of-us-adults-who-own-a-smartphone/

Tang, J., Gao, H., Hu, X., & Liu, H. (2013). Context-aware review helpfulness rating prediction. In Proceedings of the 7th ACM conference on recommender systems, ACM (pp. 1–8).

Thota, Aswini; Tilak, Priyanka; Ahluwalia, Simrat; and Lohia, Nibrat (2018) "Fake News Detection: A Deep Learning Approach," SMU Data Science Review: Vol. 1: No. 3, Article 10. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A.: The Stanford Digital Library Metadata Architecture. Int. J. Digit. Libr. 1 (1997) 108–121

Turner, A. (2022, October 1). How many people have smartphones worldwide (Oct 2022). BankMyCell. Retrieved October 2, 2022, from https://www.bankmycell.com/blog/how-many-phones-are-in-the-world

Zhao, Z., Wang, J., Sun, H., Liu, Y., Fan, Z., & Xuan, F. (2020). What factors influence online product sales? online reviews, Review System Curation, online promotional marketing and seller guarantees analysis. IEEE Access, 8, 3920–3931. https://doi.org/10.1109/access.2019.2963047

Zhu, Y., Liu, M., Zeng, X., & Huang, P. (2020). The effects of prior reviews on perceived review helpfulness: A configuration perspective. Journal of Business Research, 110, 484–494.