CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Investigation of Significant Features for Reviews Helpfulness

by

Afnan Arshad

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2021

My dissertation work is devoted to My Family, My Teachers and My Friends. I have a special feeling of gratitude for my beloved family. Special thanks to my supervisor whose uncountable confidence enabled me to reach this milestone

CERTIFICATE OF APPROVAL

Investigation of Significant Features for Reviews Helpfulness

by

Afnan Arshad

(MCS183057)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Examiner Name	Organization
(b)	Internal Examiner	Examiner Name	Organization
(c)	Supervisor	Dr. M Shahid Iqbal Malik	CUST, Islamabad

Dr. Muhammad Shahid Iqbal Malik
Thesis Supervisor
June, 2021

Dr. Nayyer Masood
Head
Dept. of Computer Science
June, 2021

Dr. M. Abdul Qadir
Dean
Faculty of Computing
June, 2021

Author's Declaration

I, **Afnan Arshad** hereby state that my MS thesis titled "**Investigation of Significant Features for Reviews Helpfulness**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Afnan Arshad)
Registration No: MCS183057

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Investigation of Significant Features for Reviews Helpfulness**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Afnan Arshad)
Registration No: MCS183057

Acknowledgement

"And your God is one God. There is no deity [worthy of worship] except Him, the Entirely Merciful, the Especially Merciful" [2:163]. First and foremost, I wish to say thanks to Allah (S.W.T) for giving me blessings, power, and knowledge to finish this research. Secondly, I wish to express my gratitude to my supervisor Dr. Muhammad Shahid Iqbal Malik for his help, precious time, and supervision. I pay my thanks to him sincerely for his assistance, motivation, and advice in this field of research. He helped me from the understanding of this subject till the write up of final thesis. I am deeply grateful to my family and my parents for their support and encouragement till the end of my MS thesis. Their prayers and guidance have lead me here.

(Afnan Arshad)

Abstract

In e-commerce business, whether a brand is well-known or is just getting started, product reviews are critical to its success. Manufacturers and merchants may benefit from reviews by analyzing purchasing trends, gaining a better knowledge of consumer demands, and incorporating this knowledge into future company initiatives. E-commerce platforms nowadays incorporate a larger number of customer reviews, and the rate at which they are obtained is also expanding. Most items have online reviews that surpass humans' abilities to examine within reasonable time restrictions, causing inconvenience and difficulty for buyers to read all product evaluations. Customers, in fact, demand a limited number of product reviews that are relevant to them. This situation introduces new policy and decision-making issues for both firms and customers. Many online platforms have long used a review ranking mechanism based on manual quality rating to reduce information overload. Due to fast increase in online reviews, review helpfulness is attracting increasing attention of experts and researchers. It assists users in decreasing the risks and uncertainty associated with online purchase. This study we used six type of features which are word2vec, fast text, GloVe, LDA, Elmo and BERT and three machine learning methods which are MLP, CART and random forest for review helpfulness prediction. Two amazon review data sets (video games, health and personal care) are used for analysis. Our results show that all six type of proposed features deliver the best performance as compared to the state-of-the-art baseline [1] features. At last we also applied wrapper backward elimination method for features selection and its improved results by 14% in video games and 10% improved in health and personal care amazon review data sets in term of MSE evaluation metric. As a result of the findings customers will be able to submit better reviews, merchants will be able to manage their websites more intelligently, and customers will be able to make better purchase decisions.

Contents

List of Figures

List of Tables

Abbreviations

BOW	Bag of Words
CART	Classification and Regression Tree
CBOW	Continues Bag of Words
DV	Dependent Variable
ELMO	Embeddings from Language Models
GloVe	Global Vectors for Word Representation
IVs	Independent Variables
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LSA	Latent Semantic Analysis
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Square Error
NB	Naive Bayes
NLP	Natural Language Processing
PLSA	Probabilistic Latent Semantic Analysis
RandF	Random Forest
RMSE	Root Mean Square Error
SVC	Support Vector Classification
SVR	Support Vector Regression
TFIDF	Term Frequency–Inverse Document Frequency
Word2Vec	Words to Vector

Chapter 1

Introduction

A wide variety of the shopping activities have been changed due to the arrival of the e-commerce. These days, most people are doing online shopping and are enjoying the ease brought by this business pattern. People can even use electronic devices like mobiles and laptops for the online shopping i.e., from booking bus tickets to ordering food from restaurants. With e-commerce, the customers can access detailed product information before purchasing the product. In the present days, the online shopping platforms ask for feedback of purchased product from their customers, mostly in the review forms. The reviews are written by the customers who have already purchased a product and posting the feedback of that purchased product. Since reviews are personal experiences, opinions, and feedbacks by the customers that's why it has great influence on the online shopping? Online product reviews help customers to get better insights of product and to identify whether the product is according to their requirements or not. In short, the reviews help the customers to make more informed purchase decisions. The reviews are important because 90% of the consumers take purchase decisions after reading the online product reviews and 72% of them act after reading positive online product reviews [2].

1.1 Background

The background knowledge of reviewing the helpful prediction will be discussed in this section. Specifically, about the fast growth of electronic commerce and its influences on the customers and also on the business and the importance of customer reviews and mechanisms will be discussed, which are used to select important and significant reviews will help customers to review in an effective manner.

1.1.1 E-Commerce

After the development of internet technologies and web 2.0, e-commerce has become a global industry [3] worth 2.9 trillion US dollars. Importance of e-commerce can be realized by the Episerver survey [4], according to the survey 26% of customers are shopping online on weekly basis and 62% of customers shop online on monthly basis in 2019, half of the customers' access e-commerce platforms multiple times per week. Figure 1.1 uses national sales data of US and UK, obtained from the US Census Bureau and Office for National Statistics, illustrates the trends of e-commerce. As shown, overall increase in online purchases over the last decade. By the end of 2021, 2.14 billion people [5] will probably use online services to buy goods. By 2040, it is predicted that e-commerce will facilitate 95% of purchases [6]. The prediction for worldwide e-commerce sales in 2023 is to hit 6.5 trillion US dollars [7], which is equal to 22% of total retail sales.

E-commerce has penetrated in the peoples' everyday lives, in different areas ranging from hotel booking and product purchasing to different kinds of virtual assistant services. Offline retailers also can take benefit from the e-commerce. As Walmart achieved double-digit e-commerce growth [9] in the financial years 2017, 2018, and 2019. Presently e-commerce related business and applications have been growing rapidly. In 2019, number of e-commerce companies was 1.3 million [10]. But in 2020, number of live websites using e-commerce technologies to improve business is nearly 12 million. The benefit of e-commerce over traditional in store shopping [11] is that the e-commerce enables the customers to purchase at

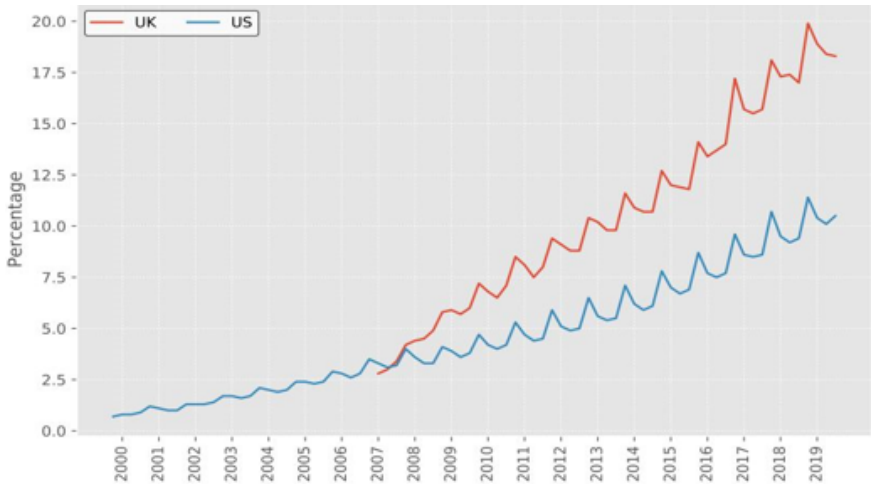

FIGURE 1.1: Projected quarterly e-commerce sales (percentage) [8]

any time without physically visiting the stores. Additionally, in e-commerce platforms, customers can compare prices of a large range of products and select best price options at one place. More importantly, the e-commerce platforms have consumers generated online product reviews that offer rich information to customers for purchase decision making.

1.1.2 Customer Reviews

Currently in the e-commerce environment, online reviews have become essential components and are basic structures of many web communities. Online reviews were supposed to impact 15.44%, Google search result rankings [12] in 2018 are up from 10.8% in 2015. Online reviews have become common practice for the information acquirement.

A survey by Bizrate Insights [13] showed that about 98% of online customers conduct research on a seller through the online reviews. In tourism field [2] statistics can also be found, where before booking 95% of travellers read reviews and then take decision. Hence online reviews play critical roles in decision making. A survey by Fan & Fuel's [14] told that 97% of the participators agree that online reviews influence into their purchasing decisions. A Capterra [15] survey reveals that online reviews influence the buying decisions of almost all software purchasers. 85% internet users in US [16] consider the online reviews as valuable as recommendations

from personal sources like friends and family, the number increases from 85% to 91% [17] if only those users whose age is between 18 and 34 are considered. Those who are providers of goods and services can also get benefits from online reviews. The vendors' can research online review to investigate customer's satisfaction [18], promote product quality and search customer needs.

Online reviews provide a trustworthy source of reference [19] that increases shop keepers' confidence, ease, and experience. 34% of customers [20] trust on content provided by vendors while doing online shopping. During research phase, 66% of buyers do not use vendors' provided materials and use other sources. According to two third of US customers [21] online users reviews are more reliable than the vendor or brand generated content. The manufacturers provided details can be 12 times [22] less trusted than the reviews by mothers who use the Internet. As according to current social influence study [23] not only 68% of customers trusted online reviews more, online peer reviews are also 16% more prominent than traditional media. Online reviews provided by a variety of customers have charm relates to an awareness of the user experiences, prospective and constrains of products [24]. Apart from vendor and manufacturer provided description, buyer can now depend on crowd sources opinions to make more well-versed shopping decisions.

Regardless of the above discussed advantages of online reviews, buyers are facing new challenges in taking advantages of the online reviews. Since 2008, number of online customers' reviews have been increases. As Figure 1.2 shows, the increasing number of customer online reviews posted on the Yelp platform has increased from 4,689 in 2008 to 177,385 in 2018, increased 37 times in a decade and the yearly growth is still increasing, Sourced from US Securities and Exchange Commission. Online reviews volume has exceeded much and require more time and effort to digest all reviews a product can receive. In addition, quality of online reviews may be poor sometime. The content of online reviews depends on reviewer's life experience, educational background and why the reviewer is writing a review. All reviews are not informative, customer require additional manual power to read reviews to get amount of information that help them in purchase making decisions.

Moreover, customers have been shown to have little tolerance in reading reviews. Maximum customers read less than 10 reviews and formed an opinion [25] or take decision about business/product. For example, for travellers to make hotel booking, they used to read average less than seven reviews [2]. In 2019, the average time customers stay in online stores [26] dropped to four minutes and twelve seconds. As volume of reviews increased with unpredictable quality and less customer tolerance for perusing reviews, needs better strategies to select and present only quality and informative reviews to customers.

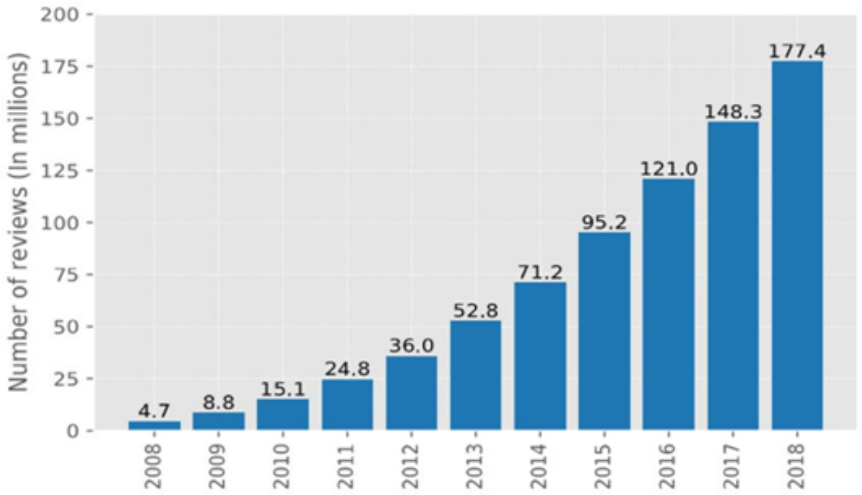

FIGURE 1.2: The number of reviews submitted to Yelp from 2008 to 2018 [27]

1.1.3 Review Ranking

The modern online shopping platforms ease customers to read efficient reviews by taking different measures. One commonly used method is to ask customers for feedback towards other customers' reviews. By asking questions at the end of each review like "Was this review helpful to you?" or "Did you find this review helpful?" online platforms take helpfulness votes from customers and then for each review the voting data was collected and analysed that what people think about the reviews by calculating the helpful reviews. Helpfulness is ratio between helpful and total votes.Figure 1.3 1.4 is the example of Amazon feedback mechanism; other platforms also have similar mechanisms. As customers' feedback accumulates, the votes showed that how helpful these reviews are for customers. From the voting,

the data helpfulness of each review can be calculated easily by using formula mentioned above and helpfulness reflects the quality of review. The platform uses this mechanism and rank reviews by their quality and become self-managed platform. In fact, Amazon uses mechanism described above and earn more than 2.7 billion US dollars every year.

FIGURE 1.3: Example of Helpfulness Voting System of Amazon [28]

FIGURE 1.4: Example of How Helpfulness Asks for Helpful Votes [29]

Along with the advantages, present voting system can be problematic. Since only a less number of customers [30] vote for review helpfulness of willing to do so. Figure 1.5 shows that power-law distribution is trailed by voting numbers. Solid (Dotted) lines show relation between number of votes and percentage of remaining product reviews. Data collected from three online platforms: Amazon [31], Yelp [32] and TripAdvisor [33].The scarcity is more simple in reviews of less traffic products [34] and currently posted reviews [35]. Moreover, the voting data may have unpredicted favouritisms. Online platforms often use helpfulness-related voting algorithms to rank reviews dynamically. These kind of ranking methodology simply fall into the

category of winner-take-all bias [36]. This means that the more votes the review have , the higher will be the ranking and as a result the review may be attractive for more helpfulness votes. Both conditions bound review ranking to small range of reviews, while the rest of the valuable reviews are ignored. Similarly, some possibly helpful reviews [37] are unreasonably rated as "unhelpful". Furthermore the voting system is a threat to spam reviews [38] and can be misused for voting manipulation. The above defects will reduce the credibility of the votes obtained.

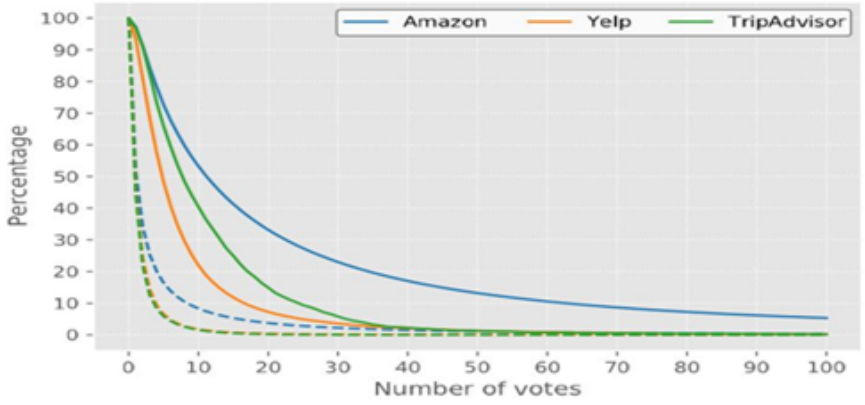

FIGURE 1.5: The Scarceness of voting data [39–41]

An alternative methodology for automatically prediction of evaluative reviews helpfulness can be discussed. The purpose of automatic helpfulness prediction [42] is to recognize and recommend high-quality reviews to users by the use of collected voting data. This branch of research contains information technology, human-computer interaction, behavior analysis and marketing. The aim of automatic helpfulness prediction is to use previously collected data of voted reviews and gained knowledge from it by adopt machine learning techniques to predict helpfulness. Often features also extract from both text and review context, after that applied machine learning techniques on them for helpfulness prediction.

Researcher have also examined other features in terms of review structure, content and social context to investigate what makes an online review helpfulness [43]. In past features such as review length and Unigrams were the most predictive features for product reviews, derived from review text. Syntactic and semantic features are also used with unigrams in previous work. Semantic and pragmatic feature have

also been used for predicting review helpfulness in NLP fields [44]. For example, review readability is correlated with review helpfulness [45].

In term of non-textual features, review star rating is used for product review helpfulness prediction [46]. Moreover, Elapsed time and the reviewers' related information [37, 46] as well as the reviewer and the reader [47] interaction are useful for review helpfulness prediction.

In e-commerce business, whether the brand is popular, or it has just started the online business, the reviews of the product play an important role in its selling. Manufactures and retailers can also take advantages from reviews by analyzing the purchase trends [18], understanding customers' needs and can use this information in future business strategies. In e-commerce, the business can use online product reviews as a threat or an opportunity for future business [48].

Presently, e-commerce platforms have gathered lager number of user generated reviews and speed of gathering is also increasing. In fact, most of products have online reviews that exceeded the ability of humans' examination within tolerable time limits reviews and it cause inconvenience and difficulty for customers to read all reviews of a product. In fact, customers require selective and small set of product reviews which are helpful for them. This situation creates new challenges to both companies and customers in policy and decision making. To eliminate information overload, review ranking mechanism through manual quality evaluation has long been used in many online platforms. Due to rapid growth of online reviews, effective solutions for the ranking of reviews and to filter low-quality content and automatically trace the useful information are needed.

1.2 Problem Statement

The aim of this work is to investigate influential set of significant features to improve the prediction accuracy for review helpfulness and apply a more robust machine learning model for predictive model construction.

1.3 Scope

This work will help the e-commerce firms in finding helpful reviews and in showing those reviews to their customers selectively. E-commerce companies will identify the helpfulness of the review as soon as the review is posted and have no votes, supports a competitive approach for e-business to enhance the ongoing use of online reviews. This research will also ease the consumers as they only get helpful reviews.

1.4 Research Question

In this research, the following research questions have been focused.

- Can latest review contextual features improve the helpfulness prediction along state-of-the-art base line by using random forest machine learning methods?
- Which type of features (Word2Vec, GloVe, Fast text, LDA, BERT and ELMo) is the most contributing features for helpfulness prediction of product review?

1.5 Research Objectives

A large percentage of online reviews contain little or no votes at all. As a result, their helpfulness is difficult to assess. Furthermore, newly published reviews and lesser-known products have less opportunities for other consumers to read them, and hence will receive less votes. As a result, rather of depending entirely on the manual voting mechanism for helpfulness, it is essential to approximate the helpfulness of online reviews using an automated technique in order to fully utilize the entire review dataset. Customers, as previously described, are unable to read all reviews. Thus, the aim of this work is to investigate review features to predict the

helpfulness of product reviews using various machine learning techniques. These approaches used the text of the review for the extraction of features, which ensures that the helpfulness of the review can be estimated as soon as the review is posted on the e-commerce website and the review can be ranked accordingly.

Chapter 2

Review of Literature

This chapter surveys literature on the prediction of the helpfulness of the review. The survey first overviews the popular strategies of the review text representation, followed by the three fundamental perspectives of helpfulness prediction; the helpfulness voting process on contemporary web platforms, the perception of the review helpfulness, and the reasons that online users consider reviews as helpful/unhelpful. Subsequently, the analysis addresses word embedding strategies, presents approaches to label review helpfulness and offers the sources of review for the task.

Existing web platforms both gather the user-generated reviews and offer the relevance of the reviews to the crowd. Helpfulness voting can be represented using theories of message and data processing [49]. In fact, the process of voting goes through a number of processes, from presentation which is the appearance of reviews, receipt which is the interpretation of and attention to of reviews), to giving birth to the belief change, evaluation, and attitude change towards review helpfulness. A research describe the helpfulness voting as a three-step process, the first one is the reviewer writes a product review; second is depending on some parameters, a ratter reads and assigns an internal score to the review; and third one is that if the score reaches any threshold, otherwise, the ratter votes the review as helpful or unhelpful [50] .A growing number of online sites currently only allow votes when users deem a review useful. This voting process eliminates the misuse

and manipulation of votes [51] and helps to cultivate a positive atmosphere [52] for the consumers who search shopping products and reviews. A list of aliases explain review helpfulness, such as review usefulness [53], review utility [54], review quality [36], review in-formativeness [55], review persuasiveness [56], and review trustworthiness [57], is preserved through the review helpfulness.

Researchers have been carefully curated over the past decade to reflect tests for prediction of helpfulness [58]. The proposed features can be classified into content-based and context-based ones, although they follow different naming conventions. Language numbers resulting from the textual content of reviews are contained in the former, while the latter includes contextual detail on reviews [59].

2.1 Content Based Approaches

Following approaches are laying under the umbrella of content-based approaches.

2.1.1 Linguistics and Syntactic Analysis

To derive information from the reviews, content-based characteristics remove different linguistic and syntactic features from analysis documents. The five logical subcategories, ordered by implementation complexity, are discussed as follows.

The structure analyses into the framework of the analysis. The layout of a review illustrates that during the writing of review, critics present their remarks. Most studies examine the review structure from review texts through length (depth) statistics. To model helpfulness, distinct granularity levels of language units have been investigated. Structural details includes the number of paragraphs [60], sentences [61], phrases [62], terms [63], characters [64], and syllables [65] in the study, from rough to fine-grained units. The standard deviation of word and sentence-level counting statistics in and the ratio of analysis length before to after preprocessing [60]. It is even necessary to derive structural characteristics entirely

inside the summary headings [66], subsections of the review/reviewer profile [67], and definition of product [68].

For more informative structural elements, several studies merge the counting statistics of different language units. For example [69] calculates in a review the ratio of unique terms also known as lexical diversity or vocabulary richness [70] calculates linguistic richness for individual reviews specified as the number of words, including punctuation, to that of unique words [71] calculates in a review the ratio of short words (less than four letters). The average number of sentences per paragraph (average length of paragraphs) [60], the average number of words per sentence (average length of sentences) [72], the average number of characters per word (average length of words) [73] and the average number of characters per word (average length of words) are another example [73] measures the standard deviation of reviews for terms and phrases.

There are also less common structural features used. The writers draw up a set of concepts correlated with Pros-Cons (called paragraph separators) and count the frequency of the concepts in the study [36]. The collection includes nouns and noun phrases widely used by clients to summarize a product's benefits and drawbacks, such as "The Good", "The Bad", "Thumb up", "Bummer", "Likes" and "Dislikes." Similarly, lists in a study the occurrence of "Pros" and "Cons" [74].

In written evaluations, the syntax explores the role of syntax. Present research examines syntactic features in examination documents, such as tenses, sections of speech, spelling and grammar accuracy, and patterns of words [35].

In particular, the number/ratio of open-class words [75], such as nouns [76], verbs [70], adjectives [70] and adverbs [76], is commonly used for the allocation of parts of speech. [54] The number of modal verbs and correct nouns, usually technical words, commodity brands, definitions, etc., is counted. Preposition [76], personal pronouns [76], foreign terms [69], symbols [69], numbers [77], punctuation [69], interjections [54], modal particles [78], and mimic words [78] are other sections of expression and their variations.

The degree to which reviews are correctly published is calculated by a collection of syntactic characteristics. The number/ratio of capitalized characters and words (commonly used for emphasis) [60], sentences beginning with capital letters [79], upper case characters [80], and lower case characters [80] are an important predictor. [80] Measures the ratio in a study of upper case to lower case characters. [81] Verifies whether an analysis begins with a capital letter. In spelling errors [82] and grammatical errors [83], the other two reasons are there. In a study using off-the-shelf English spell checkers, [79] compile the number/ratio of misspelled words.

The readability tests the degree to which web reviews are read and understood by clients. Even a modest improvement in readability will largely boost the readership of reviews [84], contributing to more chances to earn helpful votes for reviews. Seven current measures for readability also known as comprehensibility [82] have frequently been used to estimate the ease of comprehension, taking advantage of the review's structural details. Although the readability tests are well-researched in English, statistical evidence may be missing in extending them to other languages.

The years of experience required to comprehend a piece of writing was assessed by four readability tests. The Gunning Fog Index (GFI) showed that the target audience interested in the publication of newspapers and textbooks will read text quickly. A more precise and conveniently measured replacement for FOG is created by the Simple Measure of Gobbledygook (SMOG). In a study, all measures involve counting difficult words (of three or more syllables). The Automatic Readability Index (ARI) and the Coleman-Liau Index (CLI, in contrast to the syllable-based readability indices, aim at faster computation and focus only on letters, words, and sentences in reviews [85].

2.1.2 Sentiment and Semantic Analysis

The sentiment uses methods of sentiment analysis to research online reviews' valence i.e., negativity and positivity, emotional status, and subjectivity. The overall attitude displayed by clients towards a commented target is summarized by

sentiment functionality. It is possible to approach the identification of the review sentiment using lexical tools [86]. NRC Word-Emotion Association Lexicon (EmoLex) [87], General Inquirer (GI) [88], SentiWordNet (SWN) [89], Opinion Lexicon (OpiLex) [68], Geneva Affect Label Coder (GALC) [90], Linguistic Inquiry and Word Count (LIWC) [74], AFINN , WordNet-Affect (WA), and Valence and WordNet-Affect (WA) are common wordlists [91].

Training domain-specific classifiers utilizing machine learning algorithms such as Naive Bayes, Logistic Regression, and Help Vector Machine is another method to assessing sentiment. Domain experts annotate the testing samples, so the qualified model typically gains better precision in the identification of emotions [92].

Finally, off-the-shelf software [93], such as Senti Strength [94] and Opinion Finder [92], can also test review sentiment. [83] Initiates a collection of positive and negative adjectives and uses Word Net [95] synchronization to expand seed terms [79] emoticon tests, such as :-) and:-D). Researcher measures the difference between the valence of a comment and the valence of the most reviews shared. Another researcher uses a private-sourced wordlist based on hotel reviews to count positive and negative term occurrences [96].

There are various representations of the detected feelings [81]. Feeling can detect by, the total valence strength of a review, the number/ratio of positive/negative language units (e.g., terms, phrases, latent topics), the number/ratio of objective/-subjective (neutral/non-neutral) phrases, the distribution of predefined categories over a review, the continuity of sentiment between objective/subjective phrases, the one-and two-sidedness of the analysis documents, the variations and varieties alluded to above are identical.

Based on implementations and domains, the threshold that determines positivity and subjectivity will vary. The definition of analysis material is studied by the semantics. Predefined numbers are included in the first four sub-categories, which only loosely classify the analysis documents, leading to some lack of information. By specifically modelling multiple language units (usually terms and phrases) in

analysis documents, semantic features calculate the consumer sentiment in a finer-grained fashion.

BOW versions, such as unigrams [97], bigrams [98] and trigrams [99], are the standards for the encoding of examination semantics. The BOW models encode predominantly n-grams with binary values [99] incidences [100] and regular TFIDF values [97]. Researchers measure the TFIDF-based centroid score for a study. [101] Scientist builds dependency bigrams using grammatical dependencies between words to capture a longer range of semantics.

For review representation, some BOW models only use a subset of vocabulary. For example, interrupt words and phrases that appear fewer than ten times were skipped. Similarly, [88] only words with a minimum of three occurrences are contained. A researcher [102] selected about 4, 000 words with the best TFIDF ratings, instead of, picking the top 3, 000 tokens with the highest term frequencies [103]. The authors employ association analysis [33] in [78] to pick a subset of n-grams. Another researcher [104] pick nouns and noun phrases that refer to the product characteristics manually and organize the words into groups. For construction semantics, script phrases and words highlighted by participants are used in [105]. The authors assess the concreteness of material in [106] that is, the degree to which the explicit and abstract terms are used in reviews.

Furthermore, subject modelling learns knowledge from semantics about helpfulness. In [54], LSA is adopted by the authors to discover latent subjects from reviews. Four dining elements (i.e., taste/food, experience, importance, and location) are defined by [94] from online restaurant reviews. On 5.8 million Amazon product ratings, the writers learn 100-dimensional word embedding's in [54]. [61] Range between 5 and 100 vector lengths with increments of 5. [67] Measures the representation of a review by averaging the embedding of its constituent words; it was also possible to compute review vectors by learning together with word embedding. In [107], the Paragraph Vector model [60] was used to explicitly learn embedding for each sentence of a summary, which are used to infer the two-sidedness of review sentences.

The closeness of reviews in textual relevance is measured by text similarity. The most used option is Cosine similarity, which measures the cosine of the angle between the two representations of the review. The comparison between the TFIDF representation of a consumer review and that of the product specification and that of the editorial review is calculated respectively by [54]. The relation between the review texts and the summary of the product [75] is determined in [63] and that between the review texts and the questions addressed on the product page. The association between a current and its previous analysis is determined by [82].

2.1.3 Metadata, Reviewer and Product Analysis

The Metadata defines a review or an item's metadata, including quantitative measurement and temporal/spatial logs. Such knowledge helps the supplement in interpretation and confirmation, and therefore helpfulness, of the viewpoint of a reviewer.

As a quantitative supplement to the qualitative text definition, prior literature used analysis star ratings [99] to a large degree. The new rating mechanism also uses five-point Likert scales from ("strongly dissatisfied" to "strongly satisfied") to measure the general attitudes of reviewers towards products and/or facets of the object. "O'Mahony [80], for example, considers a number of review sub-ratings on TripAdvisor for Las Vegas and Chicago hotels (e.g., "Rooms", "Cleanliness" and "Business Service").

The primary type of rating information is marked by linear star scores i.e., raw values [108] the writers receive the fraction of one- and five-star ratings in an object [104]. [88] In a research it was checked that whether a review receives a moderate three-star rating; in [109] the same term is referred to as "equivocality". Park et al. [110] separates ratings into positive ratings (four and five stars) and bad ratings (one and two stars). The annual shift in average star ratings, the total number [111], and standard deviation [112] of review ratings for an object was captured [56]. In another research it is [65] reported that the number of ratings for a newly added review having the same ratings.

It is possible to normalize raw ranking ratings [65] into values between 0 and 1. Scoring extremity [113] has also been researched to a significant degree. Extremely favorable reviews may be attributable to product promotion, while damning reviews from market rivals are extremely negative. The opinion of one who deviates from the general opinion can affect the perception of helpfulness. The U-shaped relationship [114] between review ratings and helpfulness is captured by several studies using the square of linear star ratings [110]. Only the quadratic term of mild three-star rating and two most serious one-star and five-star ratings were considered in a research [115].

The age of the review [108], usually in the form of days [116], weeks [117], and years [107], reflects the period of the review after it was published. The age of the review tests the time of a review up to a certain timestamp (usually the date of data collection). The examination age is described by researchers in the papers [118] as days that have passed since 1 January 1960. The number of days after the launching of a commodity is determined by [89]. The number of days when a review appears, on the first page of the review list is counted by [65].

At another place, researchers [22] analyse whether a review has external links, addresses game ratings, and is one of the best review list entries. And it was verified [80] before writing a Trip Advisor review the number of optional blanks being filled i.e., being like and hate sections, personal and intent of visit information, and template questions.

The reviewer characteristics analyses demographics, recorded data, and past activities/behaviours of reviewers. Such details help potential readers to detect whether a reviewer is natural or suspect, the domain is encountered by the author and similarity is shared by reviewers.

The reputation of the reviewer [88] depends on the personal details presented by the user in their profiles, such as the name [110] avatar [109] age/date of birth [119] gender [109] and position [120] of the reviewer. From [119] categorize the age of a reviewer into seven Trip Advisor setting intervals: 12 years and under, 13-17 years, 18-24 years, 25-34 years, 35-49 years, 50-64 years, and 65 years and

over [36] are investigating whether a reviewer is related to a confirmed purchase. [53] Verify if it completes the names or initials are used by reviewers and simple avatars that show their faces. Expertise of reviewers [110] studies the participation rate of reviewers [78] and rankings [88]. A line of work [65] models the amount of helpfulness of Trip Advisor contributors, which is related to the number of reviews posted [117] check whether a reviewer is classified as an Elite for the Yelp platform, [110] count the number of Elite badges (awards) held by a reviewer.

The ranking habits of reviewers are often analysed. For example, to calculate its ranking tendency and accuracy, the mean [119] standard deviation [80], and skewness [76] of the historical star ratings of a reviewer are computed. The local rating deviation [76] tests the degree to which the actual rating of a reviewer was close to its normal rating behaviour. Global performance deviation [121] tests to what degree the rating conduct of a reviewer varies from the general population. The following are other seldom seen traits. Activity period as the number of days written by a reader between the first and last review were described [122]. The period of activity also tests the lapsed days [65], weeks [117] and years [119] before the web site was entered by the reviewer. For instance, [99] obtained the number of years since a reviewer registered a Yelp account and became a member of Yelp Elite.

The product attributes focus on an item's intrinsic properties that are more related, domain-specific, and platform-specific to consumer needs e.g., brand reputation. While not writing any portion of online reviews explicitly, the qualities also affect consumers to interpret the helpfulness of the review.

The popularity of objects is widely debated. Many researches draw on the assumption that further visits/purchases and therefore reviews can be drawn by common products. For items [104] hotels [116] books [56], computer games [109], attractions [76], to name a few, the total amount of reviews/photos each object has been used extensively. [56] Consider both the fraction of books without reviews and the amount of a book's annual examination. [80] Measures the mean and standard

deviation of hotels in the number of ratings. Popularity can also be calculated by ranking and temporal statistics.

Another factor, such as sales [9] and prices [112], is the economic influence of products. [56] Extracts the sales ranks of the two books sold on the two websites of online booksellers, Amazon [104] also collected more than 18 months of commodity sales and retail price figures. [71] Are gathering the price range (from $$ to $$$$) of all Yelp San Francisco hotels. The number of reviews of each price class from all Yelp restaurants in Phoenix City is counted [117]. Sales and pricing figures [120] are also obtained from various types of Amazon goods.

Yet another aspect is the essence of an item. Many research [120] classify objects into products of knowledge and quest [114], based on the ease of collecting product details and the reliance on one's senses for impartial comparison of products. In [116], the two forms are referred to as experiential and utilitarian products, while [123] prefer articulate and practical products. [16] High- and low-priced products are further separated.

The following are other uncommon causes. [108] group hotels into high-class (5- and 4-star) and low-class (3- and 2-star) hotels based on Trip Advisor's hotel star level [65]. [103] compile Trip Advisor, Expedia, and Yelp reviews of Manhattan hotels and create a categorical variable to denote the supplier of information (i.e., the source) of a review.

2.2 Context-Based Approaches

In addition to reading texts, review meaning is a critical element in the synthesis of effective helpfulness. Context-based approaches used in helpfulness prediction are discussed below.

Topic modelling, which is a family branch, implies that a text is ruled by a mixture of secret topics and a set of words in the corpus for each topic. A subject model

decomposes the document-term matrix arising from BOW or n-gram representations into a document-topic matrix and a topic-term matrix to minimize sparsity while retaining much of the semantic context. More abstract semantics such as topical and aspectual details are encoded by the compact vector space. Latent Semantic Analysis (LSA) [63] and Latent Dirichlet Allocation (LDA) [38] are two classical subject modelling techniques.

A further branch of the continuous family is characterized by distributed representations. Each token is mapped into a fixed-length real vector i.e., embedding because of training neural language models [25], wherein each dimension reflects a latent idea shared through tokens. In comparison to local representations, 101 to 103 computing elements are used in an embedding and are thus resistant to dimensional disasters. To capture more sophisticated semantic associations between words, the embedding training process considers the local meaning of a word. Distributional hypothesis [82] in linguistics inspires the intuition, words that appear in the same ways seem to have identical meanings. Therefore, in the trained vector space, identical terms in importance are spatially closer. Through basic algebraic operations, the learned representations often include word analogies [124].

The success of shallow neural networks for word semantic learning was harvested through early embedding training techniques. The Continuous Bag-of-Words (CBOW) model, Skip-Gram model with Negative Sampling (SGNS), and Global Vectors are three classical techniques [125] for studying dense word vectors (Glove). For sub words [126] and other language units, the learning model may also be extended. The (un)weighted average of the qualified vectors of its constituent tokens [127], or the learning along with the token vectors [60], or the creation of another neural model on the token vectors was used to represent the text (or document) [10].

2.3 Word Representation Approaches

Representing texts was of critical importance to many real-world uses, including review helpfulness prediction. A text (e.g., a sentence, paragraph, or document) was encoded in the form of vectors in compliance with Natural Language Processing (NLP) and information retrieval conventions. This section categorized local and continuous representations of text [128] and briefly introduced current approaches used to learn all types of representations.

Vector encoding relies on the one-to-one correspondence of a text and computing elements between physical entities (e.g., characters, words, tokens). A common approach for word representation is the one-hot encoding scheme, also known as the 1-of-N encoding scheme. The scheme first builds and indexes the vocabulary of unique tokens in the corpus, given a set of texts. Each token is represented by a sparse vector of the same length as the vocabulary, one being encoded by the vector into the element indicating the location of the token and otherwise zeros.

By aggregating the one-hot vectors of its constituent tokens, Bag-of-Words (BOW) models represent a text. Three scoring schemes and their variants are frequent used: binary values showing token presence/absence in a file, integers counting token occurrences and the standard Term Frequency Inverse Document Frequency (TFIDF) scheme [127] being the most common choice. TFIDF explains that the value of a token depends on the presence of the token in a text and the number of texts in the token-containing corpus.

BOW models ignore word orders and do thus not differentiate, for example, "is it true and "it is not true" between texts consisting of identical but differently arranged words. By encoding contiguous sequences of n constituent tokens of a text, n-gram models take as input spatial adjacency to alleviate the limitation. Person tokens (n=1), token pairs (n=2), and token triplets (n=3), also known as unigrams, bigrams, and trigrams, respectively, are frequent n-gram choices. It is noted that 1-gram model can be interpreted as a BOW model.

Still, the curse of dimensionality [125] can suffer from n-gram models. A common corpus generally includes a vocabulary of 105-107 tokens, and if higher-level n-grams are used, the vocabulary size will expand exponentially. N-gram models pose large sparsity in addition to computational inefficiency, as many n-grams only have few occurrences and carry little semantic information. Since vector elements are treated independently, synonyms (e.g., "lemon juice" and "lemonade") and hypernyms (e.g., "husky" and "dog") and other semantic relationships cannot be captured by n-gram models.

Some other approaches used pervious labelled product review dataset for helpfulness prediction, or it can be said that it needs information gained from previous ones to predict helpfulness on fresh reviews. To this end, statistical models are first equipped to extract and map representative characteristics of each review to their predicted helpfulness on labeled reviews and then used to fulfil the assignment. Helpfulness labeling also relies on a review's earned votes [6] in the form of "X" of "Y" people think a review is helpful".

Two commonly used calculation approaches are the "X of Y" and "X" helpfulness [6]. The former measures the number of favourable votes that a review gets, while the latter leverages the raw "Yes" votes. [103] translate the reviews into representations of TFIDF and compute as the centroid of the vectors the typical opinion. The helpfulness score is then determined by the similarity of the cosine between the centroid and each examination. Similarly, by measuring the cosine similarity between the LSA representation of a review and that of the word "helpful," [83] design helpfulness.

To make the calculation more intuitive and transparent for consumers, constant utility values can be translated into groups. Dichotomous discretization on the "X of Y" helpfulness is the normal scheme. All reviews, provided a threshold, are labeled as either helpful or unhelpful. A threshold equal to 0.6 balances the false positive and the false negative rate between human annotation and voting data was found [88]. The threshold [89] has been adapted by a wide body of studies; studies [100] have also manually selected the threshold within the range 0.5 - 0.9.

Most experiments trichotomize ongoing helpfulness. The researchers [101] initially consider reviews with the top and bottom 30 percent helpfulness as poor and good, respectively, and then change criteria to ensure that all groups are about equal in scale. To increase the reliability of voting results, the middle section of the review is viewed as questionable and omitted. Similarly, [79] regard as helpful (unhelpful) reviewers whose average review helpfulness scores are in the upper (lower) 40 percent; the remaining reviewers are excluded to eliminate potential biases in voting. Researcher [96] seeks better consistency of evaluation by picking only helpful reviews as those that have the top 1% helpfulness. [129] Built a log-support scoring system based on the voting data to cope with the feedback with high confidence but little support, on which three groups are defined, the helpful positive reviews, the helpful negative reviews, and the unhelpful reviews. Small numbers [57] were also used as thresholds for voting data where only "Yes" votes are eligible. The threshold was set by [63] as the average helpfulness over reviews of each product.

Online reviews used for the prediction of helpfulness are either obtained from main or secondary sources. The former applies to writing computer programs called crawlers or using programming interfaces for applications that scrape knowledge directly from targeted platforms. These approaches allow for high customizability and access to up-to-date analysis records, but in terms of time and cost, they can be challenging; therefore, the size of the data gathered is typically limited. The above applies to off-the-shelf datasets that previous researchers have prepared. For exploring different models, these databases typically have public usability and greater data size; however, the pre-collected reviews can suffer from timeliness and not represent the current pattern in many types of products.

Most current analyses rely on primary data [50], often referred to as ad-hoc datasets. Due to their effect on fact, reviews from three online sites, Amazon [114], Yelp [110] and Trip Advisor [65], are currently more preferred among the research group. These reviews primarily include user-generated thoughts about a range of items, hotels, restaurants, and attractions. Other points of analysis

include tech programs [46] on CNET, electronics [111] gathered from Yahoo shopping, App Store [112] and Google Play [109], IMDB movie ratings [35], Barnes & Noble books [56], Auto Home vehicles [78], Steam video games [24], to name a few. Any reviews from private sources are received. For instance, [104] developed a longitudinal dataset from an e-commerce business selling clothes for girls. Ad-hoc datasets, with a few exceptions, are rarely exchanged with the public [80]. The unavailability of data is one of the key factors that largely hinders reproduction and comparison of outcomes.

In recent years, there has been a growing trend for the study of helpfulness in open-source repositories [88]. [54] Exploit the Amazon analysis data proposed by [38], initially intended for spam identification of views. The Amazon Multi-Domain Sentiment Dataset [32] and the Amazon Review Data [107] include two more common alternatives.

2.4 Research Gap

On the bases of related work, it is concluded that majority of approaches are content based. Only two methods LSA and LDA are used in contextual based approach for review helpfulness prediction. These approaches are commonly used review text for helpfulness prediction. Their latest word embedding and textual analysis methods that may be applied to improve the accuracy of helpfulness model.

Chapter 3

Proposed Methodology

In this chapter, the framework for proposed solution is described, as shown in figure 3.1.

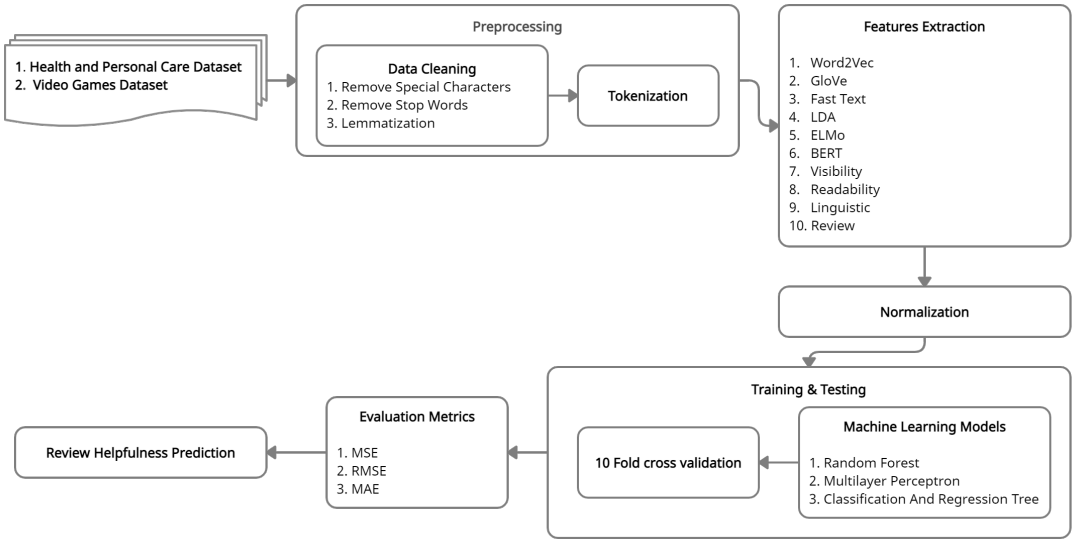

FIGURE 3.1: Block Diagram of Proposed Methodology

The two publicly available Amazon product datasets one is health and personal care dataset, and another is video games dataset were used. Firstly, the pre-processing was applied, pre-processing includes cleaning of reviews and tokenization of review text. After that features were extracted using six methods and then normalized. Machine learning method which were multilayer perceptron, classification and regression tree and random forest, applied on these normalized features

and result evaluate on basis of three metrics. Evaluation metrics were mean square error, root mean square error and mean absolute error. There were two outputs of proposed solution one was influential features and determination of best machine learning model.

This chapter was divided into various sections. The dataset is defined in Section 3.1, section 3.2 describes the features used for our experimental setup, section 3.3 machine learning model, 3.4 evaluation and section 3.5 define tools and languages.

3.1 Dataset Description

Amazon product datasets were used as shown in table 3.1. The datasets for helpful analysis are used in this research. This research interested in textual characteristics and meta-data features of these amazon reviews. These both datasets are further considered for data pre-processing.

TABLE 3.1: Amazon Product Datasets

Sr. #	Product Type	Number of reviews
1.	Health and Personal Care	346357
2.	Video Games	231781

3.1.1 Pre-Processing

Initial activities performed on Amazon product datasets are discussed including the identification and removal of duplicate reviews, the removal of empty text reviews and to remove all the reviews that have scored zero overall votes.

After removal of reviews and empty text, new dataset created is shown in table 3.2.

TABLE 3.2: Processed Datasets

Sr. #	Product Type	Number of reviews
1.	Health and Personal Care	160757
2.	Video Games	167193

3.1.1.1 Data Cleaning

The both data sets are further considered for cleaning process [130].

Stop Words Removal

Stop terms were commonly treated as additional words that do not have an effective effect on the calculation of results. E.g. In English, "the", "is" and "and", easily would identify as stop words. All stop terms in our data collection that greatly boost our investigative performance were removed.

Special Characters Removal

A character that was not an alphabetic or numeric character was a special character. Examples of special characters were punctuation marks and other symbols. Special characters were usually characters that were used for abbreviations, e.g.: @, #, $, %, & etc. Removal of such characters had no effective effects on the evaluation of results. All special characters in our data collection that greatly boost our research results were excluded.

Lemmatisation

Lemmatisation was the algorithmic procedure by which a word's lemma was determined based on its intended context. Lemmatisation, unlike stemming, relies on the proper recognition of the desired part of speech and meaning of word in a sentence, as well as in the wider sense around that sentence, such as adjacent sentences or even an entire document.

3.1.1.2 Word Tokenization

The method of splitting a large sample of text into words was word tokenization. In natural language processing tasks, this was a requirement where each word must be captured and subjected to further study, such as classifying and counting them for a certain sentiment, etc.

3.2 Features

Following features in our methodology were used.

3.2.1 Proposed Features

According to our knowledge we are the first one who used these methods for features generation for review helpful analysis.

3.2.1.1 Word2Vec

Word2vec is a mixture of models used in a corpus to describe distributed representations of words. Word2Vec (W2V) is an algorithm that accepts text corpus as an input [131] and, as seen in the figure 3.2, gives out a vector representation for each word:

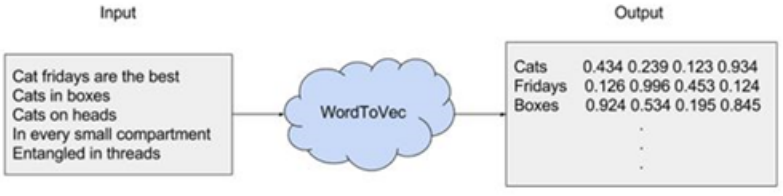

FIGURE 3.2: Word2Vec Text to Vector Representation [131]

Word2Vec is composed of two different learning models, CBOW (Continuous Bag of Words) and Skip-Gram [132].

CBOW

Continuous Bag of Words (CBOW) model can be thought of as learning word embedding by training a model to predict a word given its context.

Skip-Gram

Skip-Gram Model is the opposite, learning word embedding by training a model to predict context given a word.

We used skip-Gram model. Figure 3.3 illustrate the difference between both.

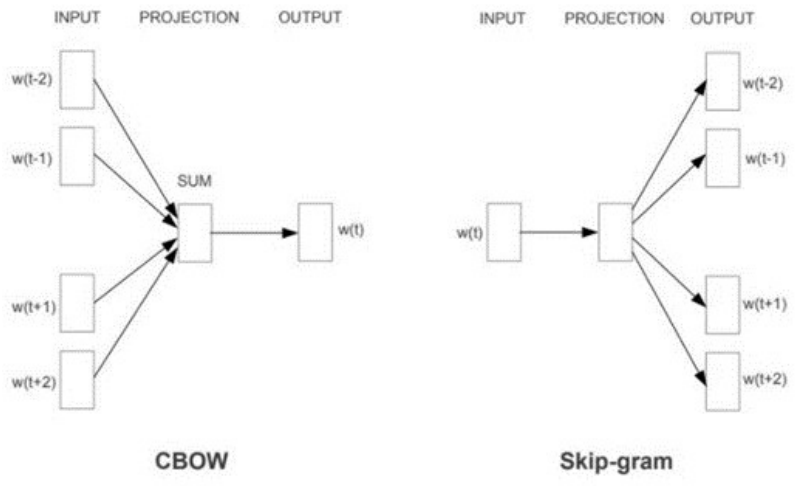

FIGURE 3.3: Continuous Bag of Words and Skip-gram leering Models [132]

3.2.1.2 GloVe

GloVe stands for "Global Vectors". And, to come up with word vectors, GloVe captures both global statistics and local statistics of a corpus. The model utilizes the key benefit of count data, the ability to collect global statistics, while capturing the important linear substructures prevalent in recent log-bilinear prediction-based techniques such as word2vec at the same time. As a result, for unsupervised learning of word representations, GloVe becomes a global log-bilinear regression model that outperforms other models on word analogy, word similarity, and tasks of recognition of named entities. Some advantage of GloVe is Fast Training, Scalable to large corpora and with a small corpus and small vectors GloVe outperform. The GloVe model architecture as shown in figure 3.4. A one-hot representation

of a word is the input. In the model, the word embedding matrices function as weight matrices and thus the model output is a vector of inner products f word vectors.

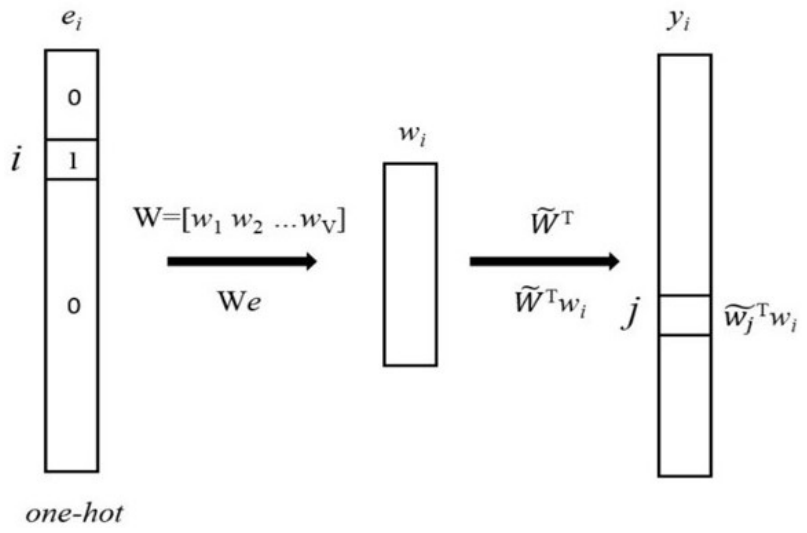

FIGURE 3.4: The GloVe model architecture [133]

3.2.1.3 FastText

FastText is a library for successful learning of word representations and classification of words. It is written in C++ and, during training, supports multiprocessing. FastText enables you to train words and phrases with supervised and unsupervised representations. These representations (embedding) can be used for various data compression implementations, as features in additional models, for candidate selection, or as transfer learning initializers. FastText uses negative sampling, softmax or hierarchical softmax loss functions to enable the training of continuous bag of words (CBOW) or Skip-gram models. Model architecture of FastText for a sentence with N ngram features x1, ... ,xN are show in figure 3.5. These features are embedded and averaged to form the hidden variable [134].

3.2.1.4 Latent Dirichlet Allocation

It is one of the most popular topic modelling methods. Topic modelling offers approaches for storing, comprehending, scanning, and summarizing large electronic

collections automatically. It will assist discovering in the collection the secret themes, classifying the documents into the patterns found and the classification is used to organize/summarize/search the data.

Each document consists of different words, and each topic often has different words belonging to it. The purpose of the LDA is, based on the words in it, to find topics to which a document belongs. As every document is a compilation of words. In the table 3.3, each row represents a different topic and each column represents a different word in the corpus. Each cell contains the probability that the word (column) belongs to the topic(row). By calculating this this probability of words, we find percentage of topic in documents.

TABLE 3.3: Example of word probability in topics

	Word 1	Word 2	Word 3	Word 4	.....
Topic 1	0.01	0.23	0.19	0.03	
Topic 2	0.21	0.07	0.48	0.02	
Topic 3	0.53	0.01	0.17	0.04	

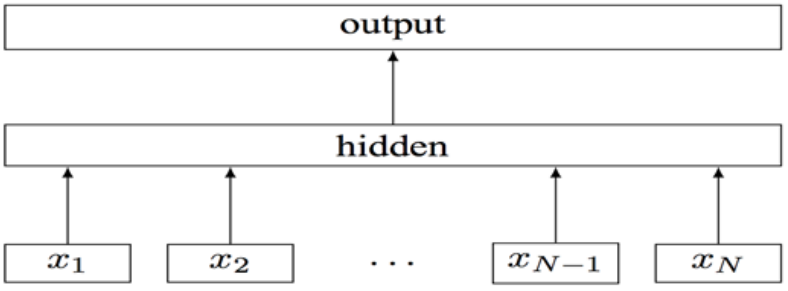

FIGURE 3.5: The FastText model architecture [134]

3.2.1.5 Embeddings from Language Models

To generate word representations, ELMO utilizes a deep, bi-directional LSTM model. ELMo analyses terms in the context that they are used, rather than a dictionary of words and their corresponding vectors and generates vectors on-the-fly by passing text via the deep learning algorithm. It is also character-based,

allowing out-of-vocabulary terms to form representations of the model as shown in figure 3.6.

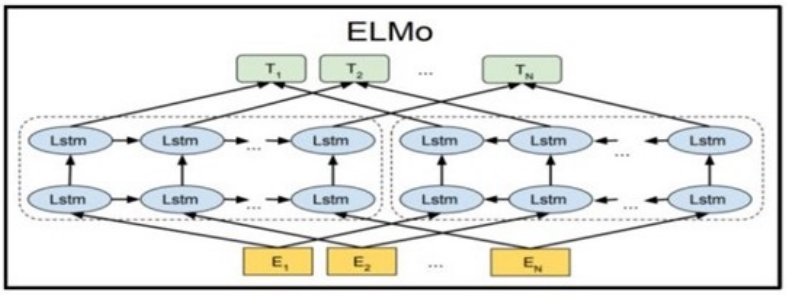

FIGURE 3.6: The ELMO architecture [135]

The ELMo vector assigned to a token or word is simply a function of the entire sentence containing that word, unlike conventional word embeddings, such as word2vec and GLoVe. The same word may then have distinct word vectors in different contexts.

Suppose we have a couple of sentences:

1. I read the book yesterday.
2. Can you read the letter now?

The verb 'read' is in the past tense in the first sentence. And the same verb in the second sentence is translated into the present tense. This is a case of Polysemy in which a word may have several meanings or senses [136]. Some features of ELMo is listed below:

1. ELMo word representations are solely character-based, allowing the network to use morphological hints to form stable representations unseen during training for out-of-vocabulary tokens.
2. It produces word vectors on run time, unlike other word embeddings.
3. It allows embedding of everything you bring in, characters, words, sentences, paragraphs, but it is built in mind for sentence embedding.

3.2.1.6 Bidirectional Encoder Representations for Transformers

BERT is a deep learning model that has produced state-of-the-art study results on a wide range of tasks for natural language processing. On Wikipedia and Books Corpus, it has been pre-trained and needs task-specific fine-tuning. BERT is a bidirectional multi-layer Transformer encoder and have two styles [7] as shown in figure 3.7.

1. BERT base – 12 layers (transformer blocks), 12 attention heads, and 110 million parameters.
2. BERT Large – 24 layers, 16 attention heads and, 340 million parameters

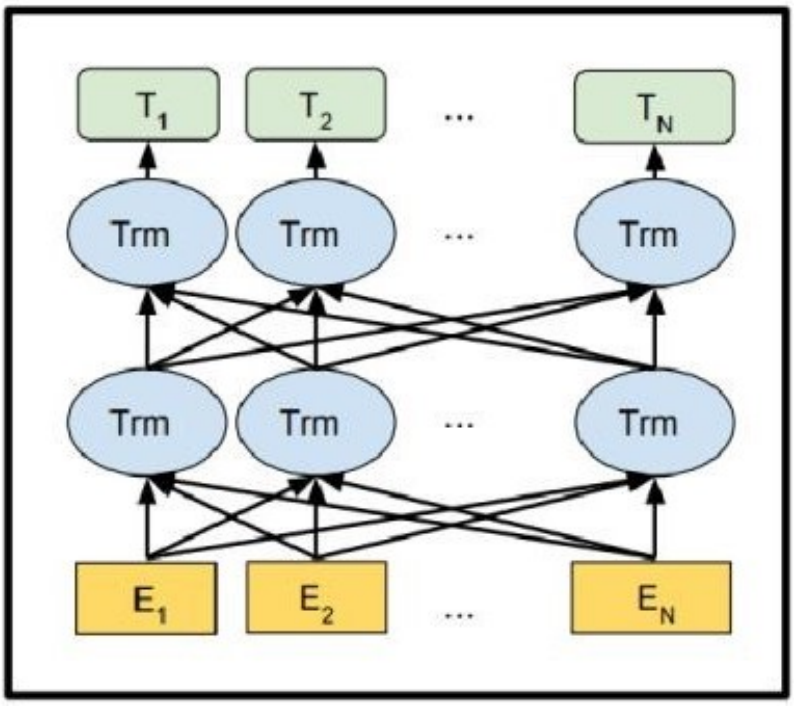

FIGURE 3.7: The BERT model architecture [137]

A word from the embedding layer begins with its embedding representation. To create a new intermediate representation, each layer does some multi-headed attention calculation on the word representation of the previous layer. The size of all these intermediate representations is the same. In the above figure, the embedding representation is E1, the final output is T1 and the intermediate representations

of the same token are Trm. A token would have 12 intermediate representations in a 12-layer BERT model.

3.2.2 Baseline

Following state of the art method were used in base line paper [1] for helpfulness prediction.

3.2.2.1 Visibility Features

The characteristics related to review meta data and length of review are important in prediction of review helpfulness as discussed in related work. The base line paper considered six visibility characteristics. Three indicators are meta data and three are the length of review characteristics. The features are:

1. Review Rating: Rating of the review
2. Elapsed_days: Elapsed days since the posting date
3. Review_sentiment: Sentiment of review in terms of rating
4. Len_chars: Length of a review in characters
5. Len_words: Length of a review in words
6. Len_Sentences: Length of a review in sentences

3.2.2.2 Readability Features

Analysis of readability is to calculate the efforts needed for readers to understand the textual content. In specific, readability measures the amount of education necessary for a reader to readily understand the text [9]. Six famous readability methods selected by base line for comparison:

1. ARI: Automated Readability Index

2. FKGL: Flesch–Kincaid Grade Level
3. SMOG: Simple Measure of Gobbledygook
4. CLI: Coleman–Liau Index
5. GFI: Gunning Fog Index
6. FKRE: Flesch-Kincaid Reading Ease

3.2.2.3 Linguistic Features

From previous research, it is known that by studying the propensities of the language and psychological properties of text, helpful voting behaviour can be better understood. Linguistic features indicate a major correlation in the literature with the helpfulness of the analysis. Language indicators were also considered in the base line paper for a state-of-the-art comparison with the suggested indicators for review helpfulness prediction [9]. The features are:

1. Nouns
2. Adjectives
3. Verbs
4. Adverbs

3.2.2.4 Review Features

Following eleven features are selected by base line paper in textual or review content:

1. Pronoun: percentage of words in the review text that are pronouns
2. Article words: percentage of words in the review text that are article words
3. Prepositions: percentage of words in the review text that are preposition

4. Aux verb: percentage of words in the review text that are aux verb
5. Drives words: percentage of words in the review text that are drive words
6. Words that focus present tense: percentage of words in the review text that focus present tense
7. Relative: percentage of words in the review text that are relative words
8. Space: percentage of words in the review text that are space words
9. Syllables: percentage of words in the review text that are syllables
10. Clout: percentage of words in the review text that are clout
11. Dictionary words: percentage of words in the review text captured by the dictionary

3.3 Machine Learning Models

The Python programming language is used to build the models of helpfulness estimation for Amazon product reviews. Three common techniques in machine learning are used: Multilayer perceptron (MLP), classification and regression trees (CART) and random forest (RandF). For different kinds of experiments using proposed and state-of-the-art baseline characteristics, these ML techniques are trained and checked. For these ML strategies, the built-in packages in Python are used. In all sorts of experiments, 10-fold cross validation is used to test the performance of three ML algorithms.

3.4 Evaluation Metrics

The performance of proposed methodology has been evaluated on the base of three evaluation metrics Root Mean Square Error (RMSE), Mean Square Error (MSE) and Mean Absolute Error (MAE).

3.4.1 Mean Square Error

A calculation of how close a fitted line is to data points is the Mean Squared Error (MSE). You take the distance vertically from the point to the corresponding y value on the curve fit (the error) for each data point and square the value. Then, for all data points, you sum up all such values and divide by the number of points minus two in the case of a fit with two parameters, such as a linear fit. The squaring is achieved so that positive values are not cancel by negative values. The lower the Mean Squared Error, the closer to the data the fit is.

The formula is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 \tag{3.1}$$

Where:

n = number of documents

x = Actual value

y = Predicted value

3.4.2 Root Mean Square Error

The standard deviation of the residuals is Root Mean Square Error (RMSE) (prediction errors). Residuals are a measure of how far data points are out from the regression line; RMSE is a measure of how these residuals are spread out. It shows you, in other words, how concentrated the data is along the best fit line.

The RMSE is the square root of average value of the square of the residual (actual - predicted).

The formula is:

$$Rootmeansquarederror(RMSE|RMSD) = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \tag{3.2}$$

Where:

y = Prediction,

x = Actual Value

The bar above the squared differences is the mean (like $\bar{x}$).

3.4.3 Mean Absolute Error

A model assessment metric used in regression models is Mean Absolute Error. The mean absolute error of a formula about a test set is the mean of the absolute values of the individual errors of estimation over all the test set instances. Each prediction error is the difference between the instance's true value and the expected value.

The formula is:

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - \lambda(x_i))}{n} \tag{3.3}$$

3.5 Tools and Languages

For the evaluation of our experimental results, we use following tools and techniques:

- Python – is used for the implementation of all algorithms
- Microsoft Excel – is used to store all calculated results
- Google CoLab
- Weka

Chapter 4

Experiments and Results Analysis

This chapter presents various number of experiments and their results are discussed. This chapter is divided into three sections. Section 4.1 describes the details of experimental setup, then section 4.2 presents feature-wise analysis of proposed and baseline features and their comparisons and section 4.3 discuss impacts of features selection.

4.1 Experimental Setup

I used Python (version 3.6) programming language and its libraries for implementation, first of all we converted Json dataset file to excel file. Json library is used to read data from Json file and xlswriter library is used to write data on excel file. We read then review data from excel file by using xlrd library and in preprocessing step we used NLTK library. All the readability, Linguistic, Visibility and Review scores are computed via the Textstat library. We used Gensim library for implementation of Word2Vec, GloVe and Fast Text. We used the default setting provided by the official released toolkit. LDA topic modelling is developed using Scikit-learn, Keras in Tensor flow is used to implement ELMo and BERT. For all classifier implementation we used SKLearn library. Google Colaboratory is used as implementation environment. Google Colaboratory is a free online cloud

based Jupyter notebook environment that allows us to train our machine learning and deep learning models on CPUs, GPUs, and TPUs. The main hardware configuration is as follows:

1. Intel(R) Xeon(R) CPU @ 2.30GHz,
2. 32GB RAM.
3. 108GB Hard Disk
4. Use GPU - Tesla T4, for ELMo and BERT

4.2 Experiment 1: Feature-wise Analysis

In this section, first we conducted experiments to analyse the impact of features normalization using three ML methods. The methods are random forest, multilayer perceptron and classification and regression tree on each feature analysis techniques. The evaluation matric that we used in this experiment is mean square error (MSE).

Figure 4.1 illustrates the results of random forest, MLP and CART using Health and personal care dataset considering MSE as evaluation matric. There is total ten feature analysis methods, four from base line paper which are review, linguistic, readability and visibility. Other six are our purposed. All ten features are not normalized. Using CART in review feature analysis method MSE is 0.3053, in linguistic MSE is 0.3215, in readability MSE is 0.3222, in visibility MSE is 0.3178, in GloVe MSE is 0.2081, in fast text MSE is 0.1948, in word2vec MSE is 0.2189, in LDA MSE is 0.1921, and in BERT MSE is 0.1349 and in ELMo MSE is 0.1297.Second classifier used is MLP. In review feature analysis method MSE is 0.2167, in linguistic MSE is 0.2309, in readability MSE is 0.2325, in visibility MSE is 0.2259, in GloVe MSE is 0.1792, in fast text MSE is 0.1709, in word2vec MSE is 0.1807, in LDA MSE is 0.1158, and in BERT MSE is 0.1112 and in ELMo MSE is 0.1062. The results of random forest classifier are best than MLP and CART.

In review feature analysis method MSE is 0.1705, in linguistic MSE is 0.1816, in readability MSE is 0.1849, in visibility MSE is 0.1764, in GloVe MSE is 0.1621, in fast text MSE is 0.1597, in word2vec MSE is 0.1696, in LDA MSE is 0.1009, and in BERT MSE is 0.0946 and in ELMo MSE is 0.0918.

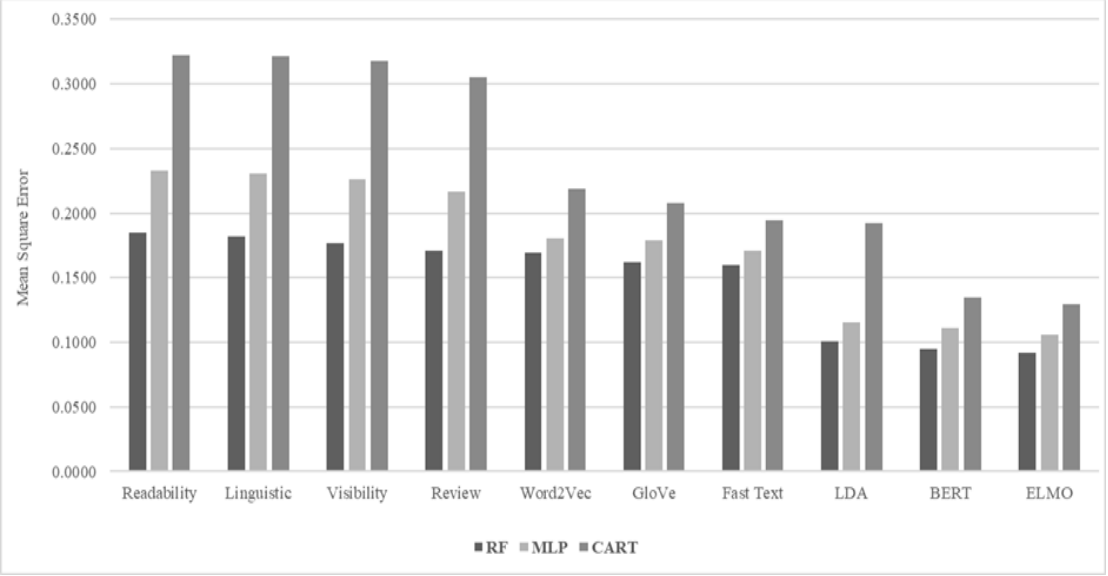

FIGURE 4.1: Feature analysis without normalization using Health and Personal Care dataset

Figure 4.2 illustrates the results of random forest, MLP and CART using Health and personal care dataset considering MSE as evaluation matric but in experiment we normalized all features before using ML methods. Using CART in review feature analysis method MSE is 0.2483, in linguistic MSE is 0.2473, in readability MSE is 0.2643, in visibility MSE is 0.2508, in GloVe MSE is 0.1730, in fast text MSE is 0.1660, in word2vec MSE is 0.1614, in LDA MSE is 0.1491, and in BERT MSE is 0.1327 and in ELMo MSE is 0.1283.Second classifier used is MLP. In review feature analysis method MSE is 0.1793, in linguistic MSE is 0.1532, in readability MSE is 0.1994, in visibility MSE is 0.1939, in GloVe MSE is 0.1431, in fast text MSE is 0.1309, in word2vec MSE is 0.1269, in LDA MSE is 0.1122, and in BERT MSE is 0.1101 and in ELMo MSE is 0.1060. The results of random forest classifier are best than MLP and CART. In review feature analysis method MSE is 0.1488, in linguistic MSE is 0.1290, in readability MSE is 0.1493, in visibility MSE is 0.1488, in GloVe MSE is 0.1280, in fast text MSE is 0.1236, in word2vec MSE is

0.1235, in LDA MSE is 0.0997, and in BERT MSE is 0.0934 and in ELMo MSE is 0.0887.

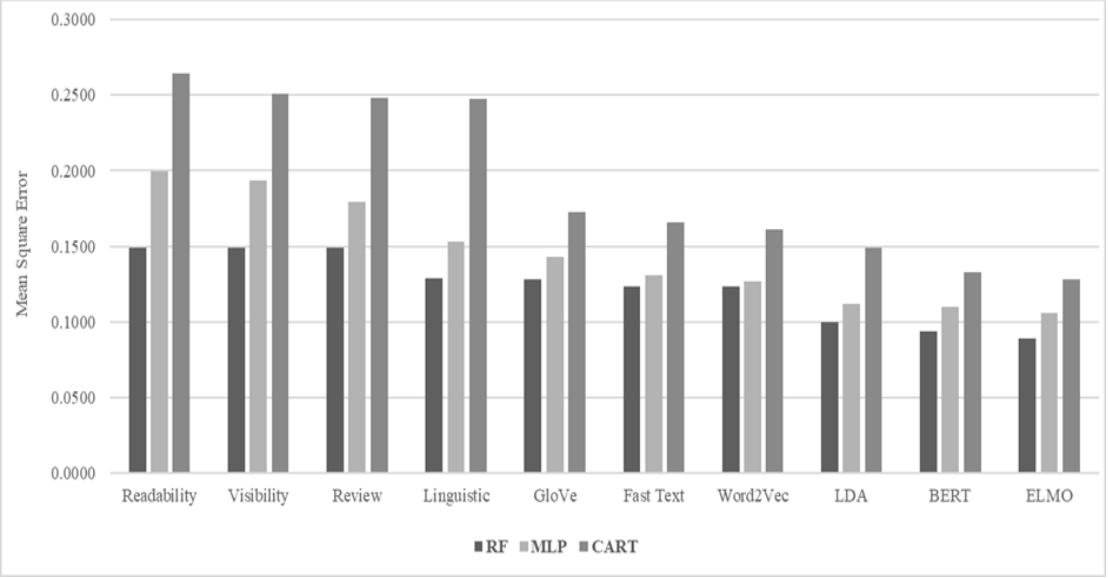

FIGURE 4.2: Normalized Feature analysis using Health and Personal Care dataset

Figure 4.1 and Figure 4.2 shows the impact normalization. From above experiments we analysed that normalizing the features have positive impact on ML models results. So we normalized our all feature for next experiments.

Then after feature normalization we conducted various experiments to analyse the impact of six features types on review helpfulness predication using three ML methods. The methods are random forest, multilayer perceptron and classification and regression tree on each feature analysis techniques. Random forest is a supervised learning algorithm that can be used to classify and forecast data. A forest, as we all know, is made up of trees, and more trees equal a more robust forest. Similarly, the random forest algorithm constructs decision trees from data sets, extracts predictions from each, and then votes on the best solution. It's an ensemble approach that's different than a single decision tree because it averages the results to reduce over-fitting. A multilayer perceptron (MLP) is a feed forward artificial neural network that produces a series of outputs from a collection of inputs. Several layers of input nodes are connected as a directed graph between the input and output layers in an MLP. Back propagation is used by MLP to train the

network. MLP is a form of deep learning. A Classification and Regression Tree (CART) is a machine learning predictive algorithm. It describes how the values of a target variable can be predicted from other values. It's a decision tree in which each branch represents a split in a predictor variable and each node at the end represents a target variable prediction. The CART algorithm is an essential decision tree algorithm that forms the basis of machine learning. It also acts as the base for other sophisticated machine learning algorithms such as bagged decision trees, random forest, and boosted decision trees. Purpose of using three models is to compare the results of each model on the basics of purposed evaluation matrices and select best one.

The evaluation matrices that we used are mean square error (MSE). In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the errors; the average squared difference between what is expected and the estimated values MSE is a probability function that reflects the squared error loss's estimated value. The MSE is determined by averaging the square of the difference between the data's original and predicted values. The second evaluation metric is mean absolute error (MAE). We recognize that an error is the total difference between the real or true values and the predicted values. If the consequence has a negative symbol, it is discarded in absolute difference. (MAE = True values – Predicted values). MAE takes the average of this error from every sample in a dataset and gives the output. MAE measures the absolute average distance between the real data and the predicted data, but it fails to punish large errors in prediction. MSE measures the squared average distance between the real data and the predicted data. Third evaluation matric which is root mean square. RMSE is the standard deviation of the errors which occur when a prediction is made on a dataset. This is the same as MSE, but the origin of the value is taken into consideration when assessing the model's accuracy. RMSE is the square root of MSE. Also, this metrics solves the problem of squaring the units. For validation purpose we use 10-fold cross validation.

We used four base line feature analysis method which are review, linguistic, readability and visibility from our base line paper [1] for compression with our purposed

feature analysis methods.

In all experiments on both dataset random forest classifier performs best. After random forest MLP performance is better and CART perform is less then random forest and MLP classifier.

Figure 4.3 illustrates the results of random forest, MLP and CART using video games dataset considering MSE as evaluation matric. There is total ten feature analysis methods, four from base line paper which are review, linguistic, readability and visibility. Other six are our purposed. Using CART in review feature analysis method MSE is 0.1825, in linguistic MSE is 0.1824, in readability MSE is 0.1813, in visibility MSE is 0.1801, in GloVe MSE is 0.1727, in fast text MSE is 0.1654, in word2vec MSE is 0.1660, in LDA MSE is 0.1382, and in BERT MSE is 0.1229 and in ELMo MSE is 0.1158. The performance of our purposed feature analysis methods is best than base line features analysis methods and ELMo performance is best among all in classifier.

Second classifier used is MLP and its results are better than CART. In review feature analysis method MSE is 0.1481, in linguistic MSE is 0.1350, in readability MSE is 0.1348, in visibility MSE is 0.1272, in GloVe MSE is 0.1185, in fast text MSE is 0.1111, in word2vec MSE is 0.1111, in LDA MSE is 0.1011, and in BERT MSE is 0.1002 and in ELMo MSE is 0.0901. Also, in MLP our purposed feature analysis methods performed better than base line feature analysis methods.

The results of random forest classifier are best than MLP and CART. In review feature analysis method MSE is 0.1071, in linguistic MSE is 0.1069, in readability MSE is 0.1061, in visibility MSE is 0.0998, in GloVe MSE is 0.0987, in fast text MSE is 0.0938, in word2vec MSE is 0.0939, in LDA MSE is 0.0934, and in BERT MSE is 0.0871 and in ELMo MSE is 0.0786. Using random forest ELMo has minimum MSE.

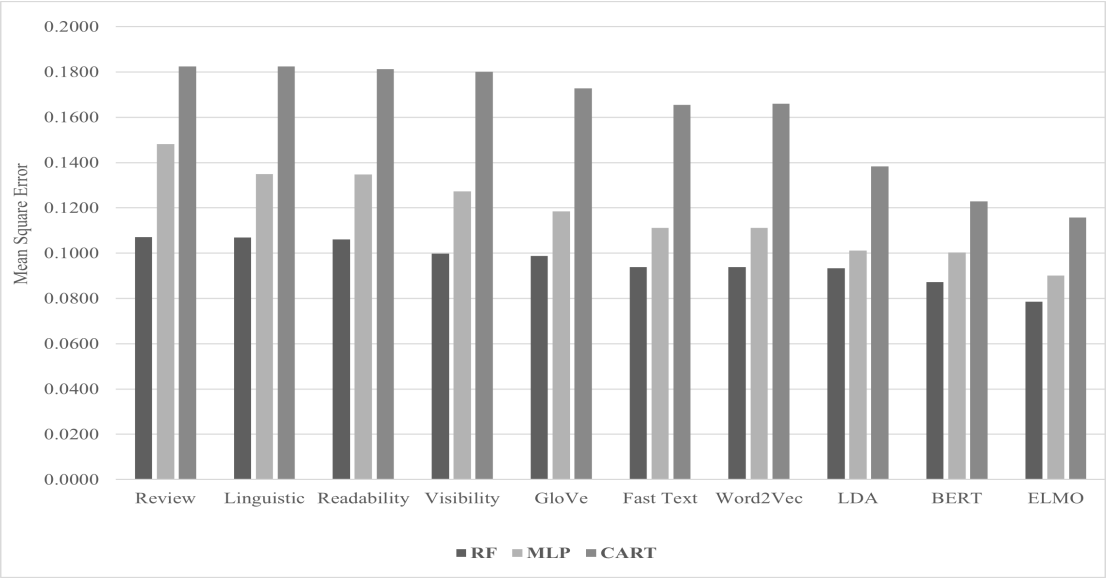

FIGURE 4.3: Feature analysis using video games dataset (three classifier's comparison via MSE)

Figure 4.4 also use video games dataset but consider MAE as evaluation matric. Numbers of feature analysis methods are same as above. Using CART in review feature analysis method MAE is 0.3242, in linguistic MAE is 0.3222, in readability MAE is 0.3182, in visibility MAE is 0.3272, in GloVe MAE is 0.3155, in fast text MAE is 0.3051, in word2vec MAE is 0.3065, in LDA MAE is 0.2818, in BERT MAE is 0.2749 and in ELMo MAE is 0.2625. Considering MAE as evaluation matric the performance of our purposed feature analysis methods is again best as MSE than base line features analysis methods and ELMo performance is best among all in CART classifier.

MLP results are better than CART again. In review feature analysis method MAE is 0. 3246, in linguistic MAE is 0. 3182, in readability MAE is 0. 3161, in visibility MAE is 0. 3061, in GloVe MAE is 0. 2861, in fast text MAE is 0. 2716, in word2vec MAE is 0. 2716, in LDA MAE is 0.2696, in BERT MSE is 0. 2679 and in ELMo MAE is 0. 2609. Also, in MLP our purposed feature analysis methods performed better than base line feature analysis methods same as CART.

The results of random forest classifier are best than MLP and CART. In review feature analysis method MAE is 0.2567, in linguistic MAE is 0.2564, in readability MAE is 0.2559, in visibility MAE is 0.2429, in GloVe MAE is 0.2512, in fast text

MAE is 0.2395, in word2vec MAE is 0.2397, in LDA MAE is 0.2377, in BERT MSE is 0.2297 and in ELMo MAE is 0.2213. Using random forest ELMo has minimum MAE.

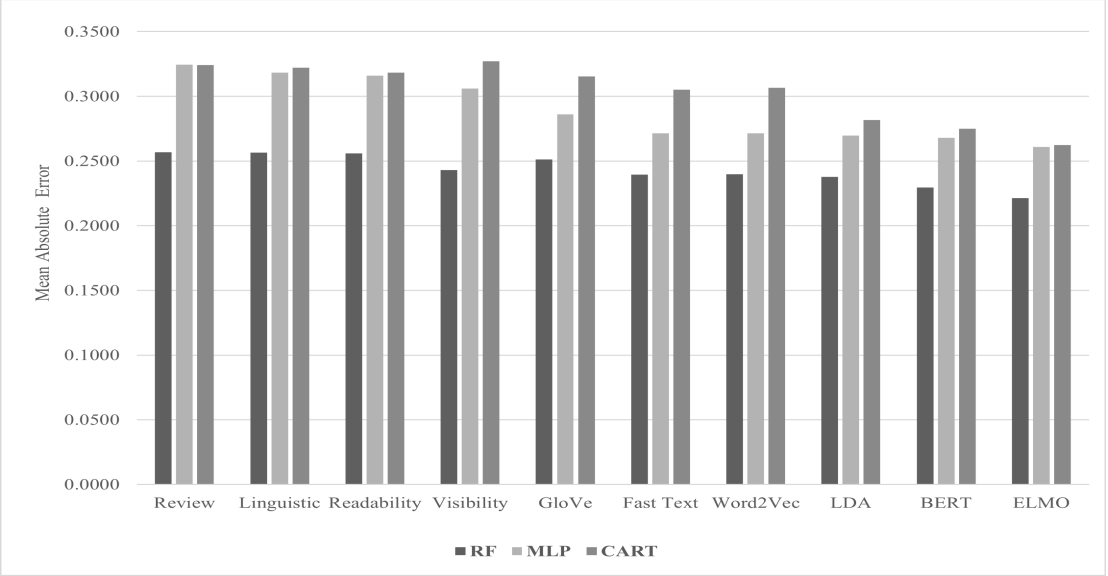

FIGURE 4.4: Feature analysis using video games dataset (three classifiers' comparison via MAE)

Figure 4.5 considered RMSE as evaluation matric using video games dataset. Using CART in review feature analysis method RMSE is 0.4272, in linguistic RMSE is 0.4271, in readability RMSE is 0.4258, in visibility RMSE is 0.4244, in GloVe RMSE is 0.4156, in fast text RMSE is 0.4067, in word2vec RMSE is 0.4074, in LDA RMSE is 0.3718, and in BERT RMSE is 0.3505 and in ELMo RMSE is 0.3403. Using MLP in review feature analysis method RMSE is 0.3849, in linguistic RMSE is 0.3674, in readability RMSE is 0.3671, in visibility RMSE is 0.3567, in GloVe RMSE is 0.3442, in fast text RMSE is 0. 3333, in word2vec RMSE is 0.3333, in LDA RMSE is 0.3179, in BERT MSE is 0. 3166 and in ELMo RMSE is 0. 3002. Using random forest RMSE in review feature analysis method is 0.3272, in linguistic RMSE is 0.3269, in readability RMSE is 0.3257, in visibility RMSE is 0.3159, in GloVe RMSE is 0.3142, in fast text RMSE is 0.3063, in word2vec RMSE is 0.3064, in LDA RMSE is 0.3056, and in BERT RMSE is 0.2952 and in ELMo RMSE is 0.2803. Performance of random forest and ELMo is best in all cases using video games dataset.

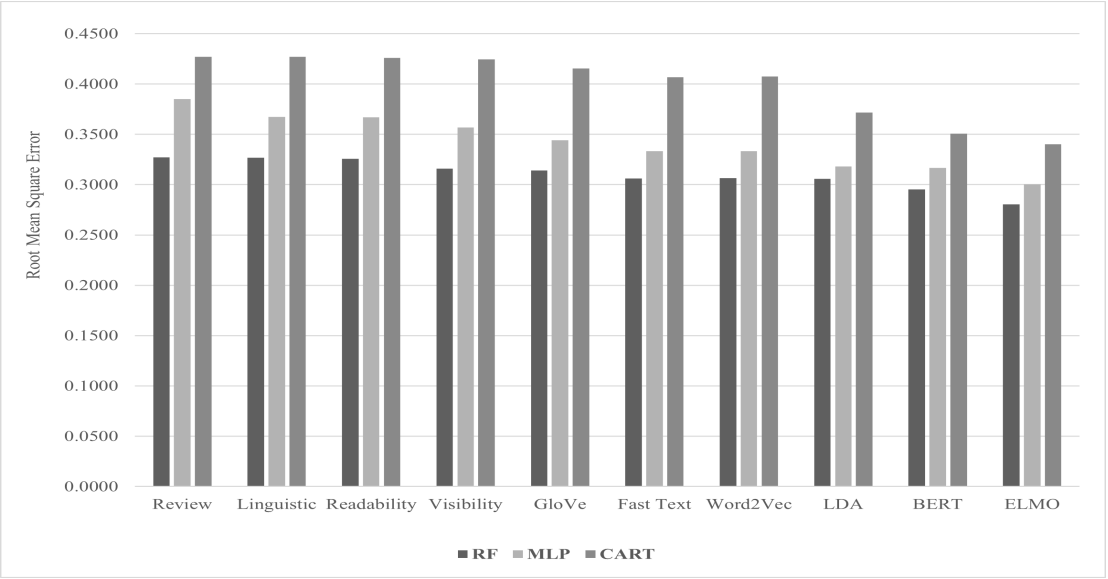

FIGURE 4.5: Feature analysis using video games dataset (three classifiers' comparison via RMSE)

Figure 4.6 use health and personal care dataset considering MSE as evaluation matric. Number of feature analysis methods are same as we used in video games dataset. First, we will explain result of CART classifier and result are arrange in descending order by MSE. In readability feature analysis method MSE is 0.2643, in visibility MSE is 0.2508, in review MSE is 0.2483, in linguistic MSE is 0.2473, in GloVe MSE is 0.1730, in fast text MSE is 0.1660, in word2vec MSE is 0.1614, in LDA MSE is 0.1491, and in BERT MSE is 0.1327 and in ELMo MSE is 0.1283.MLP results are better than CART. In readability feature analysis method MSE is 0.1994, in visibility MSE is 0.1939, in review feature analysis method MSE is 0.1793, in linguistic MSE is 0.1532, in GloVe MSE is 0.1431, in fast text MSE is 0.1309, in word2vec MSE is 0.1269, in LDA MSE is 0.1122, and in BERT MSE is 0.1101 and in ELMo MSE is 0.1060.The results of random forest classifier are best than MLP and CART. In readability feature analysis method MSE is 0.1493, in visibility MSE is 0.1488, in review MSE is 0.1488, in linguistic MSE is 0.1290, in GloVe MSE is 0.1280, in fast text MSE is 0.1236, in word2vec MSE is 0.1235, in LDA MSE is 0.0997, and in BERT MSE is 0.0934 and in ELMo MSE is 0.0887. Performance of ELMo is best using random forest classifier.

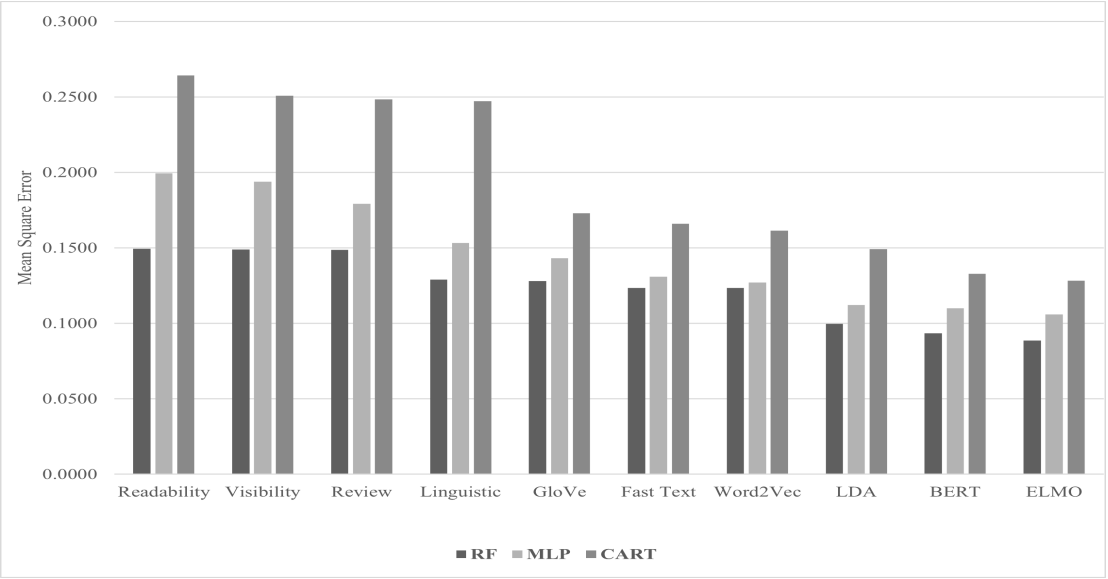

FIGURE 4.6: Feature analysis using health and personal care dataset (three classifiers' comparison via MSE)

Figure 4.7 use health and personal care dataset considering MAE as evaluation matric. Using CART in readability feature analysis method MAE is 0.4184, in visibility MAE is 0.4077, in review MAE is 0.3911, in linguistic MAE is 0.3901, in GloVe MAE is 0.3112, in fast text MAE is 0.3067, in word2vec MAE is 0.3021, in LDA MAE is 0.2876, in BERT MAE is 0.2655 and in ELMo MAE is 0.2515. The performance of our purposed feature analysis methods is best than base line features analysis methods and ELMo performance is best among all in CART classifier. Using MLP in readability feature analysis method MAE is 0. 3833, in visibility MAE is 0. 3803, in review MAE is 0. 3575, in linguistic MAE is 0. 3254, in GloVe MAE is 0. 3067, in fast text MAE is 0. 2971, in word2vec MAE is 0. 2869, in LDA MAE is 0.2717, in BERT MSE is 0. 2594 and in ELMo MAE is 0. 2433. Also, in MLP our purposed feature analysis methods performed better than base line feature analysis methods. Using random forest in readability feature analysis method MAE is 0.3196, in visibility MAE is 0.3181, in review MAE is 0.3175, in linguistic MAE is 0.2918, in GloVe MAE is 0.2908, in fast text MAE is 0.2833, in word2vec MAE is 0.2831, in LDA MAE is 0.2557, in BERT MSE is 0.2378 and in ELMo MAE is 0.2329. Using random forest ELMo has minimum MAE.

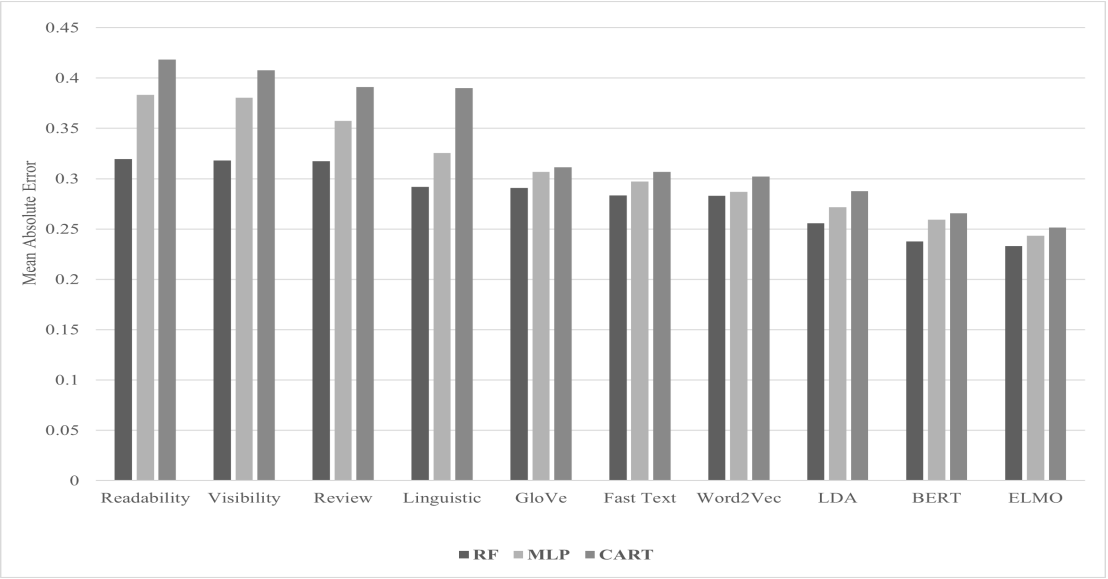

FIGURE 4.7: Feature analysis using health and personal care dataset (three classifiers' comparison via MSE)

Figure 4.8 illustrates the results of random forest, MLP and CART using health and personal care dataset considering RMSE as evaluation matric. Using CART in readability feature analysis method RMSE is 0.5141, in visibility RMSE is 0.5008, in review RMSE is 0.4983, in linguistic RMSE is 0.4973, in GloVe RMSE is 0.4159, in fast text RMSE is 0.4074, in word2vec RMSE is 0.4017, in LDA RMSE is 0.3861, and in BERT RMSE is 0.3643 and in ELMo RMSE is 0.3582. Using MLP its results are better than CART. In readability feature analysis method RMSE is 0. 4465, in visibility RMSE is 0.4403, in review RMSE is 0.4234, in linguistic RMSE is 0.3914, in GloVe RMSE is 0. 3783, in fast text RMSE is 0.3618, in word2vec RMSE is 0.3563, in LDA RMSE is 0.3349, in BERT MSE is 0.3318 and in ELMo RMSE is 0.3255. The results of random forest classifier are best than MLP and CART. In readability feature analysis method RMSE is 0.3864, in visibility RMSE is 0.3858, in review RMSE is 0.3857, in linguistic RMSE is 0.3591, in GloVe RMSE is 0.3578, in fast text RMSE is 0.3515, in word2vec RMSE is 0.3514, in LDA RMSE is 0.3157, and in BERT RMSE is 0.3056 and in ELMo RMSE is 0.2978. Using random forest ELMo has best performance.

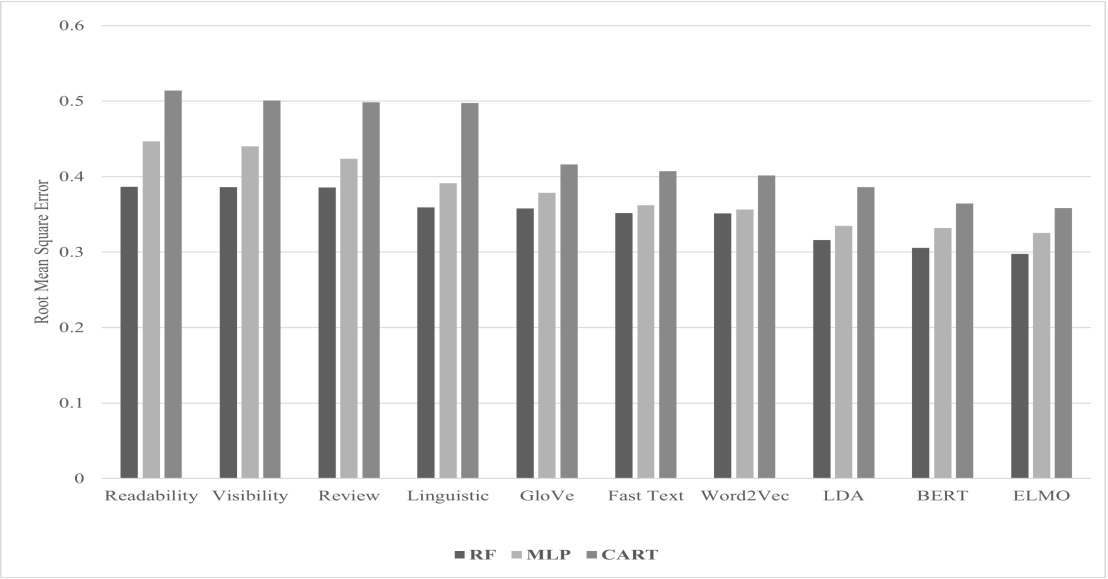

FIGURE 4.8: Feature analysis using health and personal care dataset (three classifiers' comparison via RMSE)

RQ 1 Can latest review contextual features improve the helpfulness predication along state-of-the-art base line by using random forest machine learning method?

As discussed above the latest contextual features is perform better than state of art base line by using random forest machine learning method.

4.3 Experiment 2: Impact of Feature Selection

We used wrapper backward elimination method for selection. Wrapper approaches evaluate a subset of features using a machine learning algorithm that uses a search technique to search for the space of available feature subsets, assessing each subset depending on the quality of the algorithm's results. Oder of wrapper methods working is, first wrapper methods check for a subset of features, it selects a subset of features from the available ones using a search function. Then in second step create a machine learning model, in this process, a pre-selected subset of features is used to train a machine learning algorithm. Finally, we use a selected metric to evaluate the newly trained ML model. The whole process is restarted with a new collection of features, a new machine learning model, and so on. We stop until the

desired condition is satisfied, and then in the optimization process, we select the best subset with the best result.

We must eventually stop looking for a subset of features. To do so, we'll need to set certain pre-determined conditions for when the search should end. Few examples of these standards are model performance reduce, model performance increases and a predefined number of features is reached.

Wrapper method has four search methods which are Forward Feature Selection, Backward Feature Elimination, Exhaustive Feature Selection and Bidirectional Search. We used Backward Feature Elimination. In backward feature selection we start with all the features in the dataset and then evaluate the algorithm's performance. After that, backward feature elimination eliminates one feature at a time at each iteration, resulting in the highest performing algorithm based on an evaluation criterion. This feature is also known as the least important of the options left. And so on, before a certain criterion is met, feature after feature is removed.

Wrapper methods has two main advantages. They detect the interaction between variables, and they find the optimal feature subset for the desired machine learning algorithm. Wrapper methods are normally more significantly predictive than filter methods [138].

When we apply wrapper back elimination feature selection using health and personal care dataset and random forest as ML model. when we apply wrapper back elimination feature selection method, in GloVe methodology MAE is improved by 16.36%, MSE is 25.85% improved and RMSE is 13.89% improved. Fast text has less error than GloVe and after feature MAE is improves by 4.05%, MSE improves by 10.67% and RMSE improves by 5.49%. In word2vec methodology there are 3.36% improvement in MAE, 9.39% improvement in MSE and 4.80% improvement in RMSE after feature selection. After applying wrapper back elimination feature selection LDA methodology, improvement in MAE, MSE and RMSE are 9.07%, 14.94% and 7.76% respectively. In BERT methodology MAE is improved

by 2.22%, MSE is improved by 8.24% and RMSE is improved by 4.18% improvement. In ELMo methodology after feature selection MAE, MSE and RMSE are 5.71%, 10.82% and 7.12% respectively. Detail values are shown in Table 4.1.

TABLE 4.1: Comparison with selected features using health & personal care dataset & RF as ML model

	Before Features Selection			**After Features Selection**		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
Word2Vec	0.2831	0.1235	0.3514	0.2728	0.1119	0.3345
Fast Text	0.2833	0.1236	0.3515	0.2718	0.1104	0.3322
GloVe	0.2908	0.1280	0.3578	0.2432	0.0949	0.3081
LDA	0.2557	0.0997	0.3157	0.2375	0.0848	0.2912
BERT	0.2378	0.0934	0.3056	0.2325	0.0857	0.2928
ELMo	0.2329	0.0887	0.2978	0.2196	0.0791	0.2812

When we apply wrapper back elimination feature selection using health and personal care dataset and MLP as ML model. when we apply wrapper back elimination feature selection method, in GloVe methodology MAE is improved by 7.79%, MSE is 23.27% improved and RMSE is 12.39% improved. Fast text has less error than GloVe and after feature MAE is improves by 6.32%, MSE improves by 7.41% and RMSE improves by 3.58%. In word2vec methodology there are 6.48% improvement in MAE, 8.99% improvement in MSE and 4.71% improvement in RMSE after feature selection. After applying wrapper back elimination feature selection LDA methodology, improvement in MAE, MSE and RMSE are 2.31%, 8.02% and 4.09% respectively. In BERT methodology MAE is improved by 8.96%, MSE is improved by 11.26% and RMSE is improved by 5.81% improvement. In ELMo methodology after feature selection MAE, MSE and RMSE are 7.63%, 4.52% and 2.27% respectively. Detail values are shown in Table 4.2.

TABLE 4.2: Comparison with selected features using health & personal care dataset & MLP as ML model

	Before Features Selection			**After Features Selection**		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
GloVe	0.3067	0.1431	0.3783	0.2828	0.1098	0.3314
Fast Text	0.2971	0.1309	0.3618	0.2783	0.1212	0.3482
Word2Vec	0.2869	0.1269	0.3563	0.2683	0.1153	0.3395
LDA	0.2717	0.1122	0.3349	0.2654	0.1032	0.3212
ELMo	0.2594	0.1060	0.3255	0.2396	0.1012	0.3181
BERT	0.2433	0.1101	0.3318	0.2215	0.0977	0.3125

Now using health and personal care dataset and CART as ML model. when we apply wrapper back elimination feature selection method, in GloVe methodology MAE is improved by 1.70%, MSE is 5.66% improved and RMSE is 5.21% improved. Fast text has less error than GloVe and after feature MAE is improves by 4.98%, MSE improves by 3.55% and RMSE improves by 5.20%. In word2vec methodology there are 4.56% improvement in MAE, 3.71% improvement in MSE and 2.06% improvement in RMSE after feature selection. After applying wrapper back elimination feature selection LDA methodology, improvement in MAE, MSE and RMSE are 4.90%, 6.63% and 4.14% respectively. In BERT methodology MAE is improved by 4.29%, MSE is improved by 7.38% and RMSE is improved by 2.25% improvement. In ELMo methodology after feature selection MAE, MSE and RMSE are 4.05%, 8.10% and 4.21% respectively. Detail values are shown in Table 4.3.

TABLE 4.3: Comparison with selected features using health & personal care dataset & CART as ML model

	Before Features Selection			**After Features Selection**		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
GloVe	0.3112	0.1730	0.4159	0.3059	0.1632	0.3942
Fast Text	0.3067	0.1660	0.4074	0.2914	0.1601	0.3862
Word2Vec	0.3021	0.1614	0.4017	0.2883	0.1554	0.3811
LDA	0.2876	0.1491	0.3861	0.2735	0.1392	0.3701
BERT	0.2655	0.1327	0.3643	0.2541	0.1229	0.3561
ELMO	0.2515	0.1283	0.3582	0.2413	0.1179	0.3431

Using video games dataset and random forest as ML model, in GloVe methodology MAE is improved by 7.68%, MSE is 14.18% improved and RMSE is 7.35% improved. Compare to GloVe word2vec generate better results. In word2vec after feature selection there are 9.01% improvement in MAE, 17.03% improvement in MSE and 8.90% improvement in RMSE. Then next method is fast text. Fast text gives better results than word2vec and GloVe. When apply feature selection on fast text methodology MAE is improves by 9.22%, MSE improves by 17.59% and RMSE improves by 9.21%. After applying wrapper back elimination feature selection LDA methodology, improvement in MAE, MSE and RMSE are 0.88%, 6.74% and 3.43% respectively. In BERT methodology MAE is improved by 3.91%,

MSE is improved by 9.75% and RMSE is improved by 5.04%. In ELMo methodology after feature selection MAE, MSE and RMSE are 7.90%, 9.92% and 5.06% respectively. Detail values are shown in Table 4.4.

TABLE 4.4: Comparison analysis with selected features using video games dataset and RF as ML model

	Before Features Selection			**After Features Selection**		
	MAE	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
GloVe	0.2512	0.0987	0.3142	0.2319	0.0847	0.2911
Word2Vec	0.2397	0.0939	0.3064	0.2181	0.0779	0.2791
Fast Text	0.2395	0.0938	0.3063	0.2174	0.0773	0.2781
LDA	0.2377	0.0934	0.3056	0.2356	0.0871	0.2951
BERT	0.2297	0.0871	0.2952	0.2207	0.0786	0.2803
ELMo	0.2213	0.0786	0.2803	0.2038	0.0708	0.2661

RQ 2 Which type of features (Word2Vec, GloVe, Fast text, LDA, BERT and ELMo) are the most contributing features for helpfulness predication of product review?

In both experiments ELMo results are best then all type of other features (Word2Vec, GloVe, Fast text, LDA and BERT). So ELMo is the most contributing features for helpfulness predication of product review.

Chapter 5

Conclusion and Future Work

In this chapter the research which was done and discussed in the previous four chapters is concluded. And this chapter also includes the future perspectives of this research that on which factors there is a need to do more research in future.

Conclusion

The objective of this work was to investigate influential set of significant features to improve the accuracy for review helpfulness and apply a more robust machine learning model for predictive model construction. In this study, we have used six types of features (Word2vec, glove, fast text, LDA, BERT, and ELMo) for feature analysis and review helpfulness prediction. According to our best knowledge we are the first one that consider these types of features on Amazon dataset. We used two different Amazon's review datasets which are health and personal care and video games dataset. The three different classifiers (CART, MLP and Random Forest) were trained and tested on both datasets. The findings of the current research revealed that random forest ML model performance is better considering all six types of features and on both datasets. By features point of view ELMo features performance is best considering the all the evaluation metrics (RMSE, MSE, MAE). While performance of remaining five types features is better than state-of-the-art base line features. To compare with base line, we were also implemented

and use base line features on same two Amazon datasets (health and personal care and video games). Moreover, we have used wrapper backward elimination method for features selection. In feature selection we only used random forest ML method for training and testing as it performed best in pervious features analyses step. The features selection step improved all types features result significantly.

Future Work

Future research may consider using hybrid evolutionary algorithms to improve the predictive accuracy of review helpfulness model. Semantic and sentimental variables can be explored to introduce influential determinants for the helpfulness of online reviews. In the context of product reviews, the bag-of-words may not necessarily be the ideal representation. For example, With the bag-of-words paradigm, two reviews such as "awesome hotel in an awful town" and "awful hotel in an awesome town" are portrayed in the same way, despite the fact that they convey radically opposing opinions. In the same way, we may compare the performance of unigrams, bigrams, and trigrams to determine whether they lead to any improvements. This machine learning task, in general, may be used to help with tasks like recommender systems, sentiment summarization, as text summarization, identification of the influential reviewers, opinion extraction and spam filtering etc.

Bibliography

[1] M. Malik and A. Hussain, "An analysis of review content and reviewer variables that contribute to review helpfulness," *Information Processing & Management*, vol. 54, no. 1, pp. 88–104, 2018.

[2] Invesp, "The importance of online customer reviews [infographic]." https://www.invespcro.com/blog/the-importance-of-online-customer-reviews-infographic/, 2006. (Accessed on 06/15/2021).

[3] 4, "Ups pulse of the online shopper: A customer experience study." https://www.comscore.com/Insights/Presentations-and-Whitepapers/2014/UPS-Pulse-of-the-Online-Shopper-A-Customer-Experience-Study, September 2014. (Accessed on 06/15/2021).

[4] "Episerver: 2020 online shopping habits and retailer strategies ¿ thecustomer." https://thecustomer.net/episerver-2020-online-shopping-habits-and-retailer-strategies/?cn-reloaded=1.

[5] "Digital buyers worldwide 2021 — statista." https://www.statista.com/statistics/251666/number-of-digital-buyers-worldwide/, 2021. (Accessed on 06/15/2021).

[6] N. Kitonyi, "Uk online shopping and e-comme - gurufocus.com." https://www.gurufocus.com/news/492058/uk-online-shopping-and-ecommerce-statistics-for-2017, March 2017. (Accessed on 06/15/2021).

[7] S. Lee and J. Y. Choeh, "Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees," *Management Decision*, 2017.

[8] "Census." https://www.census.gov/. (Accessed on 06/15/2021).

[9] eMarketer Editors, "Digital investments pay off for walmart in ecommerce race - insider intelligence trends, forecasts & statistics," February 2019. (Accessed on 06/15/2021).

[10] A. Murthy, "How many e-commerce companies are there? — pipecandy." https://blog.pipecandy.com/e-commerce-companies-market-size/, April 2021. (Accessed on 06/15/2021).

[11] "Kpmg international cooperative. global online consumer report." https://home.kpmg/xx/en/home/insights/2020/06/consumers-and-the-new-reality.html. (Accessed on 06/15/2021).

[12] D. Shaw, "Announcing the 2018 local search ranking factors survey - moz." https://moz.com/blog/2018-local-search-ranking-factors-survey, November 2018. (Accessed on 06/15/2021).

[13] R. Kats, "Surprise! most consumers look at reviews before a purchase - insider intelligence trends, forecasts & statistics." https://www.emarketer.com/content/surprise-most-consumers-look-at-reviews-before-a-purchase, February 2018. (Accessed on 06/15/2021).

[14] "Blog - fan and fuel." https://fanandfuel.com/blog/, April 2021. (Accessed on 06/15/2021).

[15] K. Hollar, "The impact of business software reviews — capterra." https://www.capterra.com/b2b-software-reviews-infographic, May 2015. (Accessed on 06/15/2021).

[16] R. Murphy, "Local consumer review survey: How customer reviews affect behavior." https://www.brightlocal.com/research/

local-consumer-review-survey/, December 2020. (Accessed on 06/15/2021).

[17] R. Murphy, "Local consumer review survey 2018 - brightlocal." https://www.brightlocal.com/research/local-consumer-review-survey-2018/, December 2018. (Accessed on 06/15/2021).

[18] X. Li and L. M. Hitt, "Price effects in online product reviews: An analytical model and empirical analysis," *MIS quarterly*, pp. 809–831, 2010.

[19] F. S. Khoo, P. L. Teh, and P. B. Ooi, "Consistency of online consumers' perceptions of posted comments: An analysis of tripadvisor reviews," *Journal of ICT*, vol. 16, no. 2, pp. 374–393, 2017.

[20] D. Gen, "Infographic: Content is on the rise." https://www.demandgenreport.com/resources/reports/2017-content-preferences-survey-report/, 2017. (Accessed on 06/15/2021).

[21] M. Querzoli, "New study shows user-generated content tops marketing tactics by influencing 90 percent of shoppers' purchasing decisions." https://www.prnewswire.com/news-releases/.html. (Accessed on 06/15/2021).

[22] "https://www.emarketer.com/article/moms-place-trust-other-consumers/1007509." https://www.emarketer.com/Article/Moms-Place-Trust-Other-Consumers/1007509, February 2010. (Accessed on 06/15/2021).

[23] I. MediaCT, "Social influence: Marketing s new frontier - pdf free download." https://docplayer.net/14680646-Social-influence-marketing-s-new-frontier.html, March 2014. (Accessed on 06/15/2021).

[24] "No online customer reviews means big problems in 2017 - fan and fuel." https://fanandfuel.com/

no-online-customer-reviews-means-big-problems-2017/, December 2016. (Accessed on 06/15/2021).

[25] R. Murphy, "Local consumer review survey 2016 - brightlocal." https://www.brightlocal.com/research/local-consumer-review-survey-2016/, November 2016. (Accessed on 06/15/2021).

[26] "Q1 shopping index: Global digital commerce grew 58 percent, stimulus checks boost u.s. sales - salesforce news." https://www.salesforce.com/news/stories/, April 2021. (Accessed on 06/15/2021).

[27] "Sec." https://www.sec.gov/. (Accessed on 06/15/2021).

[28] "Rivaliq." https://www.rivaliq.com/blog/customer-reviews/. (Accessed on 06/15/2021).

[29] "Powerreviews." https://www.powerreviews.com/blog/what-makes-a-review-helpful/. (Accessed on 06/15/2021).

[30] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou, "What reviews are satisfactory: novel features for automatic helpfulness voting," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 495–504, 2012.

[31] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *proceedings of the 25th international conference on world wide web*, pp. 507–517, 2016.

[32] Yelp, "Yelp." https://www.yelp.com/dataset/challenge. (Accessed on 06/15/2021).

[33] J. Li, M.-T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," *arXiv preprint arXiv:1506.01057*, 2015.

[34] M. Fan, Y. Feng, M. Sun, P. Li, H. Wang, and J. Wang, "Multi-task neural learning architecture for end-to-end identification of helpful reviews," in

2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 343–350, IEEE, 2018.

[35] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," in *2008 Eighth IEEE international conference on data mining*, pp. 443–452, IEEE, 2008.

[36] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 334–342, 2007.

[37] Q. Cao, W. Duan, and Q. Gan, "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach," *Decision Support Systems*, vol. 50, no. 2, pp. 511–521, 2011.

[38] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 international conference on web search and data mining*, pp. 219–230, 2008.

[39] "Amazon.com. spend less. smile more.." https://www.amazon.com/. (Accessed on 06/15/2021).

[40] "Restaurants, dentists, bars, beauty salons, doctors - yelp." https://www.yelp.com/. (Accessed on 06/15/2021).

[41] "Tripadvisor: Read reviews, compare prices & book." https://www.tripadvisor.com/. (Accessed on 06/15/2021).

[42] R. Kashyap and A. Ponnam, "Conceptualising a formative model for online review helpfulness: Proposal," *The Marketing Review*, vol. 19, no. 1-2, pp. 107–125, 2019.

[43] J. Tang, H. Gao, X. Hu, and H. Liu, "Context-aware review helpfulness rating prediction," in *Proceedings of the 7th ACM Conference on Recommender Systems*, pp. 1–8, 2013.

[44] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 10, pp. 1498–1512, 2010.

[45] M. P. O'Mahony and B. Smyth, "Using readability tests to predict helpful product reviews," in *Paper presented at RIAO 2010 the 9th international conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, Paris, France, April 28-30, 2010*, 2010.

[46] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pp. 423–430, 2006.

[47] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *2011 IEEE 11th international conference on data mining workshops*, pp. 81–88, IEEE, 2011.

[48] X. Yan, J. Wang, and M. Chau, "Customer revisit intention to restaurants: Evidence from online reviews," *Information Systems Frontiers*, vol. 17, no. 3, pp. 645–657, 2015.

[49] M. A. Hamilton and K. L. Nowak, "Information systems concepts across two decades: An empirical analysis of trends in theory, methods, process, and research domains," *Journal of communication*, vol. 55, no. 3, pp. 529–553, 2005.

[50] G. O. Diaz and V. Ng, "Modeling and prediction of online product review helpfulness: a survey," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 698–708, 2018.

[51] Amazon, "Amazon's "not helpful" button missing? — random thoughts - randocity!." https://randocity.com/2019/04/13/amazons-not-helpful-button-missing/, April 2019. (Accessed on 06/15/2021).

[52] R. Hanna, "Amazon on a positive note: The end of downvoting — sellics." https://sellics.com/blog-amazon-on-a-positive-note-the-end-of-downvoting/. (Accessed on 06/15/2021).

[53] Z. Liu and S. Park, "What makes a useful online review? implication for travel product websites," *Tourism management*, vol. 47, pp. 140–151, 2015.

[54] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 51–57, 2006.

[55] X. Sun, M. Han, and J. Feng, "Helpfulness of online reviews: Examining review informativeness and classification thresholds by search products and experience products," *Decision Support Systems*, vol. 124, p. 113099, 2019.

[56] I. Pentina, A. A. Bailey, and L. Zhang, "Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp review persuasiveness and purchase intentions," *Journal of Marketing Communications*, vol. 24, no. 2, pp. 125–145, 2018.

[57] R. Filieri, "What makes an online consumer review trustworthy?," *Annals of Tourism Research*, vol. 58, pp. 46–64, 2016.

[58] R. Rietsche, D. Frei, E. Stöckli, and M. Söllner, "Not all reviews are equal-a literature review on online review helpfulness," 2019.

[59] M. Li, L. Huang, C.-H. Tan, and K.-K. Wei, "Helpfulness of online product reviews as seen by consumers: Source and content features," *International Journal of Electronic Commerce*, vol. 17, no. 4, pp. 101–136, 2013.

[60] M. K. Baowaly, Y.-P. Tu, and K.-T. Chen, "Predicting the helpfulness of game reviews: A case study on the steam store," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 5, pp. 4731–4742, 2019.

[61] M. E. Haque, M. E. Tozal, and A. Islam, "Helpfulness prediction of online product reviews," in *Proceedings of the ACM Symposium on Document Engineering 2018*, pp. 1–4, 2018.

[62] A. Qazi, K. B. S. Syed, R. G. Raj, E. Cambria, M. Tahir, and D. Alghazzawi, "A concept-level approach to the analysis of online review helpfulness," *Computers in Human Behavior*, vol. 58, pp. 75–81, 2016.

[63] S. Saumya, J. P. Singh, A. M. Baabdullah, N. P. Rana, and Y. K. Dwivedi, "Ranking online consumer reviews," *Electronic commerce research and applications*, vol. 29, pp. 78–89, 2018.

[64] E. Bjering, L. J. Havro, and Ø. Moen, "An empirical investigation of self-selection bias and factors influencing review helpfulness," *International Journal of Business and Management*, vol. 10, no. 7, p. 16, 2015.

[65] Y.-H. Hu and K. Chen, "Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings," *International Journal of Information Management*, vol. 36, no. 6, pp. 929–944, 2016.

[66] S. Yang, J. Yao, A. Qazi, *et al.*, "Does the review deserve more helpfulness when its title resembles the content? locating helpful reviews by text mining," *Information Processing & Management*, vol. 57, no. 2, p. 102179, 2020.

[67] M. Akbarabadi and M. Hosseini, "Predicting the helpfulness of online customer reviews: The role of title features," *International Journal of Market Research*, vol. 62, no. 3, pp. 272–287, 2020.

[68] J. E. Fresneda and D. Gefen, "A semantic measure of online review helpfulness and the importance of message entropy," *Decision Support Systems*, vol. 125, p. 113117, 2019.

[69] L. B. Maroun, M. M. Moro, J. M. Almeida, and A. P. C. Silva, "Assessing review recommendation techniques under a ranking perspective," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, pp. 113–123, 2016.

[70] J. P. Singh, S. Irani, N. P. Rana, Y. K. Dwivedi, S. Saumya, and P. K. Roy, "Predicting the "helpfulness" of online consumer reviews," *Journal of Business Research*, vol. 70, pp. 346–355, 2017.

[71] L. Zhu, G. Yin, and W. He, "Is this opinion leader's review useful? peripheral cues for online review helpfulness," *Journal of Electronic Commerce Research*, vol. 15, no. 4, p. 267, 2014.

[72] R. Dong, M. Schaal, M. P. O'Mahony, and B. Smyth, "Topic extraction from online reviews for classification and recommendation," in *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[73] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: A randomized experiment," *Science*, vol. 341, no. 6146, pp. 647–651, 2013.

[74] D. Dey and P. Kumar, "A novel approach to identify the determinants of online review helpfulness and predict the helpfulness score across product categories," in *International Conference on Big Data Analytics*, pp. 365–388, Springer, 2019.

[75] X. Zheng, S. Zhu, and Z. Lin, "Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach," *Decision Support Systems*, vol. 56, pp. 211–222, 2013.

[76] M. Malik and K. Iqbal, "Review helpfulness as a function of linguistic indicators," *Int J Comput Sci Netw Secur*, vol. 18, no. 1, pp. 234–240, 2018.

[77] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in *Proceedings of the 19th international conference on World wide web*, pp. 691–700, 2010.

[78] Y. Liu, C. Jiang, Y. Ding, Z. Wang, X. Lv, and J. Wang, "Identifying helpful quality-related reviews from social media based on attractive quality theory," *Total Quality Management & Business Excellence*, vol. 30, no. 15-16, pp. 1596–1615, 2019.

[79] S.-T. Li, T.-T. Pham, and H.-C. Chuang, "Do reviewers' words affect predicting their helpfulness ratings? locating helpful reviewers by linguistics styles," *Information & Management*, vol. 56, no. 1, pp. 28–38, 2019.

[80] M. P. O'Mahony and B. Smyth, "A classification-based review recommender," in *Research and development in intelligent systems XXVI*, pp. 49–62, Springer, 2010.

[81] J. L. Barbosa, R. S. Moura, and R. L. d. S. Santos, "Predicting portuguese steam review helpfulness using artificial neural networks," in *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pp. 287–293, 2016.

[82] Z. Zhang, Y. Ma, G. Chen, and Q. Wei, "Extending associative classifier to detect helpful online reviews with uncertain classes," in *2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology (IFSA-EUSFLAT-15)*, pp. 1134–1139, Atlantis Press, 2015.

[83] Y. Liu, J. Jin, P. Ji, J. A. Harding, and R. Y. Fung, "Identifying helpful online reviews: a product designer's perspective," *Computer-Aided Design*, vol. 45, no. 2, pp. 180–194, 2013.

[84] W. H. DuBay, *Smart Language: Readers, Readability, and the Grading of Text.* ERIC, 2007.

[85] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, vol. 11, no. 3, pp. 205–217, 2012.

[86] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 51–57, 2006.

[87] M. Mousavizadeh, M. Koohikamali, and M. Salehan, "The effect of central and peripheral cues on online review helpfulness: A comparison between functional and expressive products," 2015.

[88] Y. Yang, C. Chen, and F. S. Bao, "Aspect-based helpfulness prediction for online product reviews," in *2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI)*, pp. 836–843, IEEE, 2016.

[89] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3751–3759, 2015.

[90] H. Liu, Y. Gao, P. Lv, M. Li, S. Geng, M. Li, and H. Wang, "Using argument-based features to predict and analyse review helpfulness," *arXiv preprint arXiv:1707.07279*, 2017.

[91] C. Strapparava, A. Valitutti, *et al.*, "Wordnet affect: an affective extension of wordnet.," in *Lrec*, vol. 4, p. 40, Citeseer, 2004.

[92] P.-J. Lee, Y.-H. Hu, and K.-T. Lu, "Assessing the helpfulness of online hotel reviews: A classification-based approach," *Telematics and Informatics*, vol. 35, no. 2, pp. 436–445, 2018.

[93] S. P. Eslami, M. Ghasemaghaei, and K. Hassanein, "Which online reviews do consumers find most helpful? a multi-method investigation," *Decision Support Systems*, vol. 113, pp. 32–42, 2018.

[94] Y. Luo and X. Xu, "Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp," *Sustainability*, vol. 11, no. 19, p. 5254, 2019.

[95] W. Fellbaum, "An electronic lexical database (language, speech, and communication)," 1998.

[96] L. Martin and P. Pu, "Prediction of helpful reviews using emotions extraction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[97] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pp. 423–430, 2006.

[98] W. Xiong and D. Litman, "Automatically predicting peer-review helpfulness," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 502–507, 2011.

[99] D. R. Radev, H. Jing, M. Styś, and D. Tam, "Centroid-based summarization of multiple documents," *Information Processing & Management*, vol. 40, no. 6, pp. 919–938, 2004.

[100] M. Passon, M. Lippi, G. Serra, and C. Tasso, "Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining," *arXiv preprint arXiv:1809.08145*, 2018.

[101] M. Mertz, N. Korfiatis, and R. V. Zicari, "Using dependency bigrams and discourse connectives for predicting the helpfulness of online reviews," in *International Conference on Electronic Commerce and Web Technologies*, pp. 146–152, Springer, 2014.

[102] S.-Y. Hwang, C.-Y. Lai, J.-J. Jiang, and S. Chang, "The identification of noteworthy hotel reviews for hotel management," *Pacific Asia Journal of the Association for Information Systems*, vol. 6, no. 4, p. 1, 2014.

[103] Z. Xiang, Q. Du, Y. Ma, and W. Fan, "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," *Tourism Management*, vol. 58, pp. 51–65, 2017.

[104] P. Zhao, J. Wu, Z. Hua, and S. Fang, "Finding ewom customers from customer reviews," *Industrial Management & Data Systems*, 2019.

[105] T. L. Ngo-Ye, A. P. Sinha, and A. Sen, "Predicting the helpfulness of online reviews using a scripts-enriched text regression model," *Expert Systems with Applications*, vol. 71, pp. 98–110, 2017.

[106] S. Shin, N. Chung, Z. Xiang, and C. Koo, "Assessing the impact of textual content concreteness on helpfulness in online travel reviews," *Journal of Travel Research*, vol. 58, no. 4, pp. 579–593, 2019.

[107] B. Lutz, N. Pröllochs, and D. Neumann, "Understanding the role of two-sided argumentation in online consumer reviews: A language-based perspective," *arXiv preprint arXiv:1810.10942*, 2018.

[108] E. Y. Wang, L. H. N. Fong, and R. Law, "Review helpfulness: The influences of price cues and hotel class," in *Information and Communication Technologies in Tourism 2020*, pp. 280–291, Springer, 2020.

[109] S. Karimi and F. Wang, "Online review helpfulness: Impact of reviewer profile image," *Decision Support Systems*, vol. 96, pp. 39–48, 2017.

[110] S. Park and J. L. Nicolau, "Asymmetric effects of online consumer reviews," *Annals of Tourism Research*, vol. 50, pp. 67–83, 2015.

[111] D. Yin, S. D. Bond, and H. Zhang, "Anxious or angry? effects of discrete emotions on the perceived helpfulness of online reviews," *MIS quarterly*, vol. 38, no. 2, pp. 539–560, 2014.

[112] D. Yin, S. Mitra, and H. Zhang, "Research note—when do consumers value positive vs. negative reviews? an empirical investigation of confirmation bias in online word of mouth," *Information Systems Research*, vol. 27, no. 1, pp. 131–144, 2016.

[113] C. Vo, D. Duong, D. Nguyen, and T. Cao, "From helpfulness prediction to helpful review retrieval for online product reviews," in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, pp. 38–45, 2018.

[114] M. M. Susan and S. David, "What makes a helpful online review? a study of customer reviews on amazon. com," *MIS Quarterly*, vol. 34, no. 1, pp. 185–200, 2010.

[115] Y. Wang, J. Wang, and T. Yao, "What makes a helpful online review? a meta-analysis of review characteristics," *Electronic Commerce Research*, vol. 19, no. 2, pp. 257–284, 2019.

[116] Y. Pan and J. Q. Zhang, "Born unequal: a study of the helpfulness of user-generated product reviews," *Journal of retailing*, vol. 87, no. 4, pp. 598–612, 2011.

[117] J. Li, E. Ngai, *et al.*, "An examination of the joint impacts of review content and reviewer characteristics on review usefulness—the case of yelp. com," 2016.

[118] M. Siering, J. Muntermann, and B. Rajagopalan, "Explaining and predicting online review helpfulness: The role of content and reviewer-related signals," *Decision Support Systems*, vol. 108, pp. 1–12, 2018.

[119] L. Kwok and K. L. Xie, "Factors contributing to the helpfulness of online hotel reviews," *International Journal of Contemporary Hospitality Management*, 2016.

[120] L. M. Willemsen, P. C. Neijens, F. Bronner, and J. A. De Ridder, ""highly recommended!" the content characteristics and perceived usefulness of online consumer reviews," *Journal of Computer-Mediated Communication*, vol. 17, no. 1, pp. 19–38, 2011.

[121] S. Mukherjee, K. Popat, and G. Weikum, "Exploring latent semantic factors to find useful product reviews," in *Proceedings of the 2017 SIAM international conference on data mining*, pp. 480–488, SIAM, 2017.

[122] M. Malik and A. Hussain, "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Computers in Human Behavior*, vol. 73, pp. 290–302, 2017.

[123] M. Mousavizadeh, M. Koohikamali, and M. Salehan, "The effect of central and peripheral cues on online review helpfulness: A comparison between functional and expressive products," 2015.

[124] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

[125] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[126] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[127] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[128] T. Mikolov, J. Dean, Q. Le, T. Strohmann, and C. Baecchi, "Learning representations of text using neural networks," in *NIPS Deep learning workshop*, pp. 1–31, 2013.

[129] Y.-C. Zeng, T. Ku, S.-H. Wu, L.-P. Chen, and G.-D. Chen, "Modeling the helpful opinion mining of online consumer reviews as a classification problem," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014*, 2014.

[130] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," in *2008 Eighth IEEE international conference on data mining*, pp. 443–452, IEEE, 2008.

[131] Z. Ali, "A simple word2vec tutorial. in this tutorial we are going to... — by zafar ali — medium." https://medium.com/@zafaralibagh6/a-simple-word2vec-tutorial-61e64e38a6a1, January 2019. (Accessed on 06/15/2021).

[132] J. Gilyadov, "Word2vec explained." https://israelg99.github.io/2017-03-23-Word2Vec-Explained/, March 2017. (Accessed on 06/15/2021).

[133] C. Liu, P. Zhang, T. Li, and Y. Yan, "Semantic features based n-best rescoring methods for automatic speech recognition," *Applied Sciences*, vol. 9, no. 23, p. 5053, 2019.

[134] R. Zhu, D. Yang, and Y. Li, "Learning improved semantic representations with tree-structured lstm for hashtag recommendation: An experimental study," *Information*, vol. 10, no. 4, p. 127, 2019.

[135] K. Purohit, "Learn how to build powerful contextual word embeddings with elmo — by karan purohit — saarthi.ai — medium." https://medium.com/saarthi-ai/elmo-for-contextual-word-embedding-for-text-classification, June 2019. (Accessed on 06/15/2021).

[136] P. JOSHI, "What is elmo — elmo for text classification in python." https://www.analyticsvidhya.com/blog/2019/03/learn-to-use-elmo-to-extract-features-from-text/#:~:text=ELMo%20is%20a%20novel%20way,as%20well%20as%20the%20indu, March 2019. (Accessed on 06/15/2021).

[137] Y. Seth, "Bert explained – a list of frequently asked questions – let the machines learn." https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/, June 2019. (Accessed on 06/15/2021).

[138] Y. Charfaoui, "Hands-on with feature selection techniques: Wrapper methods — by younes charfaoui — heartbeat." https://heartbeat.fritz.ai/hands-on-with-feature-selection-techniques-wrapper-methods, January 2020. (Accessed on 06/15/2021).