Predicting the Helpfulness of Online Product Reviews

Felicia Hjalmarsson

September, 2021

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the Bachelor's degree in Software Engineering. The thesis is equivalent to 10 weeks of full time studies.

The author declare that they are the sole author of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Felicia Hjalmarsson
E-mail: fehj18@student.bth.se

University advisor:
Mikael Svahnberg
Department of Software Engineering

Faculty of Computing	Internet	:	www.bth.se
Blekinge Institute of Technology	Phone	:	+46 455 38 50 00
SE–371 79 Karlskrona, Sweden	Fax	:	+46 455 38 50 57

Abstract

Review helpfulness prediction has attracted growing attention of researchers that proposed various solutions using Machine Learning (ML) techniques. Most of the studies used online reviews from Amazon to predict helpfulness where each review is accompanied with information indicating how many people found the review helpful. This research aims to analyze the complete process of modelling review helpfulness from several perspectives. Experiments are conducted comparing different methods for representing the review text as well as analyzing the importance of data sampling for regression compared to using non-sampled datasets. Additionally, a set of review, review meta-data and product features are evaluated on their ability to capture the helpfulness of reviews. Two Amazon product review datasets are utilized for the experiments and two of the most widely used machine-learning algorithms, Linear Regression and Convolutional Neural Network (CNN). The experiments empirically demonstrate that the choice of representation of the textual data has an impact on performance with tf-idf and word2Vec obtaining the lowest Mean Squared Error (MSE) values. The importance of data sampling is also evident from the experiments as the imbalanced ratios in the unsampled dataset negatively affected the performance of both models with bias predictions in favor of the majority group of high ratios in the dataset. Lastly, the findings suggest that review features such as unigrams of review text and title, length of review text in words, polarity of title along with rating as review meta-data feature are the most influential features for determining helpfulness of reviews.

Keywords: Review helpfulness prediction, product reviews, machine learning, data sampling, regression

Contents

Chapter 1

Introduction

The quality of online product reviews today are highly inconsistent and the countless number of available reviews can result in information overload for consumers. Presenting the most helpful reviews, instead of all, can greatly ease purchase decision making.

In order to present users with helpful reviews, major E-commerce websites like Amazon.com encourage users to vote whether a review is helpful or not. However, most reviews do not have helpfulness votes. In fact, fewer than 20% of the reviews in Amazon Review Dataset have 5 or more votes while only 0.44% have 100 or more votes [1]. Reviews that have been recently posted will not have any votes and will therefore not be presented to consumers to receive the assessments that are required in order to properly estimate helpfulness. Additionally, reviews with few votes are often not trusted. Consequently, there is a need to be able to automatically evaluate the helpfulness score of all reviews as soon as the review is written.

1.1 Background

Review helpfulness prediction has attracted growing attention of researchers that proposed various solutions using Machine Learning (ML) techniques. The goal of ML is to discover meaningful patterns and learn from data [2]. The learned information acquired from the training data is stored in some knowledge representation structure which is typically in the form of a model. A learning algorithm will then specify how to update the learned information with new experience (i.e. training data) to optimize the performance of the task [2]. In this context, learning is the process of converting experience into expertise or knowledge [3].

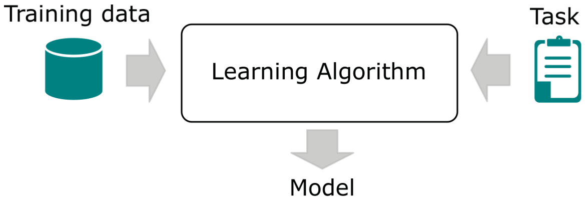


Figure 1.1: A generic machine learning method

Supervised learning, which is the most commonly used approach for review helpfulness prediction, refers to the task of learning a function from training examples, based on their attributes, which are the *inputs* to the learning algorithm, and labels, which are the correct *outputs*. Each training example is a pair of input and output of the underlying unknown function. The goal of supervised learning is ultimately to return a function f, given a set of examples of Y, that best approximates Y [2]. The following equation is a supervised learning algorithm in the most basic form:

$$Y = f(x) \tag{1.1}$$

Through machine-learning it is possible to learn underlying relationships between data which can be used to explore which attributes make the review more helpful. In review helpfulness prediction tasks, these attributes are usually divided into three categories. The first are attributes connected to the review itself such as the length of the review. Many of the previous studies have studied features of this category as they have hypothesized that helpfulness is an underlying property of the text [4][5]. The second group of attributes is related to the reviewer who wrote the review. Reviewer characteristics such as expertise, reputation and non-anonymity are among this category of attributes [16][17][11]. The final and less highlighted one are product attributes, which could include product type or other meta-data of the product being reviewed [4][19].

1.1.1 Machine learning in NLP and text mining

When dealing with unstructured textual data, such as reviews, there are special techniques needed to analyse this data. By leveraging text mining and Natural Language Processing (NLP) techniques it is possible to extract meaning out of text [6]. In the Oxford English Dictionary, text mining is defined as *"the process or practice of examining large collections of written resources in order to generate new information."*. The goal is to discover relevant information in text by transforming the text into data that is suitable for analysis. To achieve this, a variety of analysis methodologies such as natural language processing (NLP), are applied. In NLP and text mining communities, the study of review helpfulness is concentrated on finding features contained in the review text that are most influential for automatic helpfulness prediction.

In order to perform machine learning on text documents, the text content needs to be converted into numerical data that can be processed by a computer. There are several methods to represent written text in numeric form and the quality of the representation could impact the performance of many applications [6]. The process of transforming text into usable data is usually performed by building a language model that assigns probabilities, frequencies or other numbers to words, sentences etc. Some of the most commonly used techniques for modelling language are 1-hot encoding, N-grams, Bag-of-words, tf-idf, Word2vec and GloVe.

1.1.2 Classification and Regression

The task of predicting helpfulness of reviews has been treated as both a classification and a regression task in prior research. Both of these two learning tasks deal with the prediction of the value of one variable (the target) based on the values of the other variables (attributes or features) [2]. What differentiates the two tasks is the target variable, the value which is to be predicted. If the target is categorical, such as “helpful” or “unhelpful”, then the given task is called classification. If the target is numerical, such as 0.87, the task is called regression.

1.1.3 Data sampling

Data sampling is a method of collecting only a subset of the data for the analysis. When working with a relatively large dataset it might not be possible to build a machine learning model with all the data due to the computational limit on the computer - selecting a sample to use is therefore a necessary step. Additionally, imbalanced datasets could skew the performance of machine-learning algorithms, a problem which has been addressed for classification tasks in prior studies with proposed data sampling methods.

1.1.4 Definition of Review Helpfulness

Reviews can be defined as peer-generated comments written to evaluate a product or service posted in an e-commerce platform. The concept of review “helpfulness” is quite complex yet understanding what it refers to is particularly important. A positively written review increases the perceived value of a product and a critical review decreases it. Both types of reviews are opportunities to perform customer service as it allows for customers to assess the product’s value. The central idea of review “helpfulness” is therefore defined in this study by whether the review gives useful product assessment and buying decisions to other customers.

In a voting system, as the one currently used by Amazon, review helpfulness can be defined as given in equation 1.2.

$$Helpfulness = \frac{\#\, of\ Positive\, Votes}{Total\ \#\, of\ Votes}\ , \tag{1.2}$$

where total votes is the sum of the positive and negative votes. Even though this simple and effortless approach of calculating helpfulness has been adequate and increases an estimated \$2.7B revenue to Amazon.com annually[1], it still has several limitations. The lack of votes for new reviews and the fact that not everyone who reads reviews actually votes on them [5] are major drawbacks to this approach. As a result, the reviews with the most helpfulness votes are not necessarily the most helpful ones. Hence, further investigation and experimenting is necessary to answer the question of what the actual attributes are that make the review helpful to customers.

1.1.5 Motives and value

The increasing popularity of modeling and predicting review helpfulness in recent years can be explained by its importance for consumers and businesses. These days, customers regularly rely on user reviews across all industries (hotels, restaurants,

[1] https://articles.uie.com/magicbehindamazon/

products, movies) to make choices. Considering the huge amount of reviews available on the Web, a review helpfulness prediction system could substantially save people's time by presenting the most helpful reviews. Hence, a successful review helpfulness prediction system could be as useful as a product recommender system.

Many previous studies in the field of review helpfulness [4][5][17][18][20] consider two important issues for predicting the helpfulness of reviews. The first one is determining the variables that are affecting the helpfulness of reviews, such as review, reviewer or product variables. The second issue is adopting a suitable learning algorithm for predicting review helpfulness. Although feature and algorithm selection are important factors in review helpfulness prediction, little research has investigated the complete process of model building from several perspectives. As previous research has been based on direct experiments with proposed influential features or comparisons of machine-learning algorithms, it is missing evaluation of appropriate techniques of data sampling and data preprocessing for each feature.

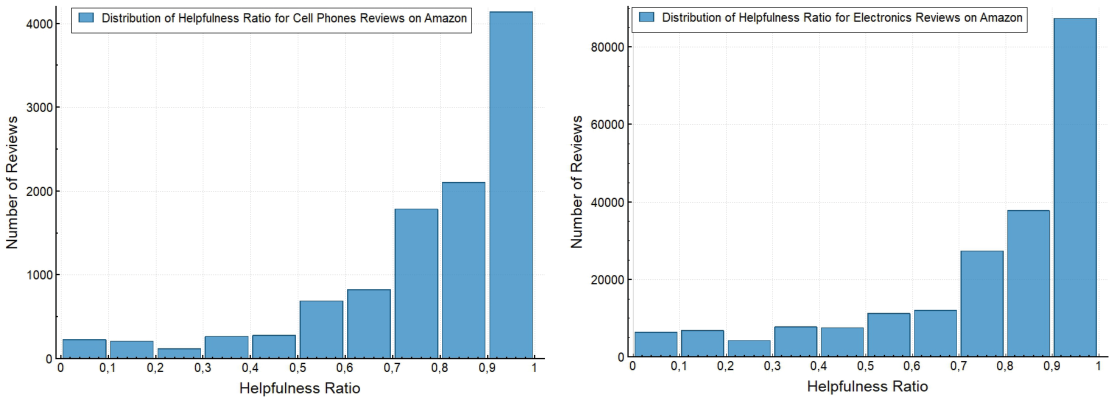
The current research is developed to bridge the gap in literature by shedding more light on this connection. Transforming text into something an algorithm can digest is a complicated process. How to encode such data in an effective and powerful way could have a strong impact on the performance of a model. Yet, different techniques are often used in current research without any explanations or motivation behind it. In terms of converting the text data for modelling, the current research uses several different approaches for embedding the review text which is used as a predictor of review helpfulness. A range of different techniques for processing the textual content of the review are compared in order to identify the best processing option as well as to see the impact of the different choices.

Additionally, this research is inspired by the observation that the distribution of helpfulness ratio scores in the product review datasets often are imbalanced, dominated with reviews with either high or low instances. In Figure 1.2 below, the distribution of helpfulness ratios for two of the datasets in Amazon is shown. As can be seen by the bar plots, both datasets hold large volume of reviews at helpfulness range from 0.9 to 1. Consequently, this might skew the performance of machine learning algorithms which are mainly constructed based on balanced or almost balanced data. While classification problems for imbalanced data have been addressed frequently in previous research [8][9] and are available in lots of papers, regression problems are overlooked most of the time, although dealing with them is significantly different. Data sampling approaches might be proposed to address such tasks.

1.2 Purpose

This bachelor thesis project has two main purposes:

- Contributing to the existing research on how to automatically assess the helpfulness of reviews by answering the question of what the actual attributes are that make the review helpful to customers. The goal is to predict the helpfulness scores for online product reviews from the public Amazon Dataset based on the information contained in the review text as well as review and product metadata, and to understand what makes a review helpful.



(a) (b)

Figure 1.2: Distribution of helpfulness scores in two product categories a) Cell Phones And Accessories and b) Electronics in Amazon Dataset

- Evaluating the impact of word representations and data sampling on the performance of a regression prediction model. Not only are these important aspects for the task of review helpfulness prediction, but the findings could be transferable to other problem definitions as well. In particular, the study will consider the most often used family of bag-of-words models, tf-idf and the recently proposed vector-space models word2vec and GloVe. The influence of data sampling on the performance of a regression model will be evaluated using a stratified sampling method compared with non-sampled datasets.

1.3 Outline

In this research I explore how to automatically assess the helpfulness of online product reviews. To accomplish this, it is necessary to consider the definition of what a useful review is and the conditions of which a useful review has. Clarification and knowledge for that will be required and therefore I will start with a survey of how other studies define a useful review in the related work section.

The following section covers the methodology of the present work: the selected final features for the experiments are presented and the required steps from data collection and preprocessing to the final experimental settings are discussed in depth. Each of the text representation methods included in the experiments are then described as well as the applied data sampling approach. Finally, the machine learning models are presented followed by a discussion of the chosen evaluation metrics.

The results and analysis section presents the outcome of the experiments, as well as comparisons with prior related studies. Then the performance of the various models are discussed with conclusions regarding the effects of data sampling and text representation along with identification of the most influential features.

Chapter 2
Aim, Objectives and Research Questions

2.1 Aim and objectives

The objective of this thesis is to explore the influential indicators of review content, review- and product meta-data that strongly contribute to helpfulness of product reviews. The impact of each type of indicator on perceived helpfulness is examined by utilizing two real-life Amazon product review datasets. Two popular ML algorithms are trained and tested using proposed features and evaluated by three evaluation metrics. The effect of different word representation methods is also examined for their contribution toward helpfulness. The current study contributes to prior studies by looking further into data sampling techniques for handling imbalanced datasets in regression tasks.

The following research questions are addressed:

Research Question 1 (RQ1)

What is the impact of different word representations on model performance for predicting helpfulness of product reviews?

Research Question 2 (RQ2)

How does data sampling techniques influence the performance of predictive modelling for regression?

Research Question 3 (RQ3)

What input features among review and product categories are the most important to capture the helpfulness of a product review?

Chapter 3

Related Work

This chapter will review related research on the topic of review helpfulness prediction. The first section will review the most important attributes that have been confirmed by previous research to have a significant impact on review helpfulness. The following section covers text representation methods that have been used in many of the prior studies. To the best of my knowledge, the current study is the first to investigate the impact of data sampling on predictive modelling for review helpfulness as a regression task. Therefore, this subject will not be included in this chapter.

3.1 Features

Previous research have highlighted different features when determining the helpfulness of online reviews. These features (also called attributes) are usually divided into three categories:

- Review attributes
- Reviewer attributes
- Product meta-data attributes

The first category of attributes is about the review itself and contains all criteria that are related to the written text of the review, such as the review length, the sentiment expressed in the review text and the semantics of the written text and so on.

The second category is related to characteristics of the person that has written the review and includes features indicating the credibility or qualification of the reviewer. The selected features may vary depending on the e-commerce platform that is utilized and what information is provided in the reviewer Social Profile Information.

The last category is connected to the product the review was written for. Features in this category could include product type or average rating, and meta-data such as product description, image, or price.

3.1.1 Linguistic and sentiment features

Kim et al. [5] investigated how well different class features capture the helpfulness of online reviews for two categories of products in the datasets available from Amazon, MP3 players and digital cameras. Their research focused on five classes of review characteristics, namely structural, lexical, syntactic, semantic and meta-data features. Structural features they experimented with were length of words, number of sentences and average sentence length etc. Lexical features included tf-idf (term frequency and inverse document frequency) of unigrams (single words) and bigrams (pair of words) respectively, the percentage of tokens that are nouns, verbs, adverbs and adjectives were used as syntactic features, positive and negative sentiment words belonging to the semantic class, and the reviewer's rating on the product were utilized as meta-data feature. Their analysis of different features from the performed experiments revealed that the length of the review, its unigrams (tf-idf representation of each word), and its product rating were the most useful determinants to rank the reviews in terms of helpfulness. Apart from the length of the review, other structural features were concluded not to have a significant impact on the prediction of helpfulness as well as features explored in the syntactic class.

Later, another predictive model is presented by Yan et al. [18] hypothesizing that helpfulness is an internal property of the review text. In their experiments they employed two additional semantic features besides text features used in previous related work concerning Structure, Unigram and Geneva Affect Label Coder (GALC) emotion, the latter one being a general lexicon of emotional words. The two semantic features introduced, LIWC and General Inquirer (INQUIRER), are two text analysis tools for mapping text to some semantic tags or to determine the degree of positive or negative emotions, self-references etc. Their reasoning behind the use of the two semantic features was that "a helpful review includes opinions, analyses, emotions and personal experiences, etc". The results from the experiments showed that the semantic feature based models outperformed the baseline features (unigrams and structural features such as number of words, number of sentences etc.) in terms of correlation coefficient.

Some studies made in recent years have tried to extract more sophisticated and interpretable features from the text. Chen et al. [13] approaches the task of helpfulness prediction by leveraging aspect analysis of reviews. Aspect-based features were extracted using topic modelling with the hypothesis that a helpful review will cover many aspects of a product at different emphasis levels. Some of the extracted aspects were *Price*, *Functionality*, *Appearance* and so on. The results supported their hypothesis that reviews covering more aspects tend to be more helpful than reviews covering only a few aspects. Certain aspects did however not contribute to helpfulness and the result were highly dependent on the category of product being reviewed. Additionally, argument-based features are investigated by Liu et al. [12] and their effectiveness for the helpfulness of hotel reviews. These argument-based features, inherent in the reviews, were manually annotated for the study and were used jointly and in comparison to more commonly used structural, unigram and emotional features. The argument-based features were further divided into component-level (the ratio between the number of premises and that of claims etc), token-level (the total number of words in the given component type etc), letter-level (ratio of as a big number of letters and a small number of words) and position-level (the positions of argument components). Among these features the result suggested that token and letter-level feature sets are most effective in identifying helpful reviews. The observations were interpreted as *the larger number of tokens a review contains, the more likely the review is helpful.* The length of reviews was therefore considered to play an important role in the review helpfulness identification.

3.1.2 Review metadata

Qu et al. [10] implemented a CNN (Convolutional Neural Network) model to assess online review helpfulness. The star rating information was used in combination with the review text as determinants for helpfulness, where the star rating was used as a proxy for the sentiment of the review. The experiment conducted demonstrated an increase in accuracy of 2% compared to that of models trained only on review text.

The star rating of a review have consistently been used as a predictor for helpfulness throughout studies. In some research, the extremity of the rating (positive, negative, neutral), have been utilized [17], in which it has been evident that positive and negative ratings are seen as more useful than neutral. In other research, the star rating has been used directly as in Qu et al.'s work.

3.1.3 Reviewer-related features

The impact of reviewers, qualitative and quantitative characteristics of reviews are explored by Huang et al. [16] on helpfulness. The three examined reviewer attributes were cumulative helpfulness (the past average helpfulness), experience (total reviews written) and impact (the number of total votes received on a specific review) which were used as indicators of qualification and credibility of the reviewer. The experiments demonstrated that reviewer experience and reviewer impact does not have significant effect on helpfulness, however, reviewer cumulative helpfulness was detected to be a significant predictor along with product rating. Additionally, word count with a certain threshold was proven effective for the helpfulness of user reviews.

3.1.4 Product features

Some research has incorporated product features as explanatory variables for helpfulness along with additional review features. Mudambi and Schuff [19] investigated the influence of review extremity (star rating), review depth (word count) and product type on perceived helpfulness of the review by using Amazon data. In their study they categorized products as either *search* or *experience* goods which was coded as a binary variable (0 or 1). Their definition of search goods are products that are relatively difficult for consumers to obtain information on the quality prior to purchase as it requires the use of one's senses to evaluate and with attributes that are subjective. Experience goods is defined as the opposite in which attributes are objective and more easily could be compared, obtaining information on the quality is relatively easy prior to interaction with the product. They hypothesize that as different products have differing information needs, the product type feature could effect how both the review extremity and the review depth of reviews are influencing helpfulness. For experience goods, they hypothesis that reviews with extreme ratings are less helpful than reviews with moderate ratings. They also hypothesized that Review depth has a greater positive effect on the helpfulness of the review for search goods than for experience goods. The performed experiments demonstrated support of their hypotheses as review extremity and review length had differing effects on the helpfulness of that review, depending on product type.

Another study that investigated the impact of product determinants on review helpfulness by utilizing Amazon datasets was [4]. This study proposed five product-related indicators among several other features which they further categorized as review content, reviewer, linguistics, readability and visibility features. The proposed product features are percentage of positive reviews, percentage of critical reviews, number of words in product title, number of questions answered and potential score of a product. From the experiments they concluded that the best predictive performance is obtained by review content features as compared to other type of features. The reviewer-type and review visibility-type features deliver comparable performance, but better than product, readability and linguistics-type features using both datasets. This shows that reviewer features are less significant than review content, but better than other features to predict the helpfulness of product reviews. In addition, type of product features presents the lower predictive performance. Number of words in product title is the most effective, and percentage of critical reviews is the least effective indicators of product type. However, the experiments revealed that product-type features are less effective than reviewer and review-type features.

Prior research has invested much effort on identifying important features of helpfulness of product reviews. The current study has considered some of the most influential features found in prior studies, in addition to other features available in the datasets, in order to make a comprehensive study of features. Not considering strong features, such as rating, length of review and unigrams, might result in key insights becoming harder to observe.

3.2 Review text representation

The previous studies on review helpfulness prediction have used different methods for representing the words in the review text which are further utilized as features. Qu et al. [10] compares the effect of different word embedding initializations by using both GloVe pre-trained word vectors and randomized word embedding initialization. The two methods achieve similar performance. Pre-trained embeddings from GloVe are also used for initialization in Chen et al.'s [15] study in which a CNN model is built to predict helpfulness of reviews across different product categories. Other studies [5][21][14][18][12] leverages tf-idf method for representation of the review text. In Haque et al.'s [22] study, both unigram and bigram lexical features are computed using tf-idf method in addition to embeddings learned from Word2Vec which are considered as semantic features. In another study by Passon et al. [23], both Bag-of-words (BoW) and tf-idf preprocessing variants are used for the experiments in combination with features exploited from MARGOT, an argumentation mining system. Although the focus of this study was to evaluate the effectiveness of features obtained from MARGOT, thus completely ignoring the textual content of the review, the results revealed that tf-idf representation with MARGOT features achieves the best result.

None of the related research motivates the chosen method for representing the review text or compares the performance of other methods. This study differs from other related research by comparing the performance of the most frequently used methods in prior studies and evaluating the impact for this task.

Chapter 4

Method

For the empirical research method I will be conducting experiments by constructing various predictive models. Different experiments will be made for each of the research questions as the conditions will differ. The methodology adapted in this research is presented in Figure 4.1. The first step is to collect reviews from two categories of products from the public Amazon Review Dataset. The two categories are Electronics and Cell Phones and Accessories, each with 1689188 and 194439 data instances respectively. In the next step, these product reviews are further preprocessed, and review content, review- and product meta-data features are computed. The review text is then represented using four different techniques, Bag-of-words, tf-idf, Glove and Word2Vec. These methods are chosen as they are the most frequently used in prior related studies. The data is sampled using stratified sampling method which are compared with non-sampled dataset. The data will be splitted into training, test, validation sets for performance estimations. After this, two regression models including a regularised linear regression and a Convolutional Neural Network (CNN) model are chosen to construct the predictive models for review helpfulness. These regression models are evaluated using mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE) evaluation metrics. Finally, the set of most influential features as well as text representation method are identified for building an effective review helpfulness prediction model. Additionally, the effect of data sampling is evaluated.

In this study, the examined case is the regression problem in which the target variable is review helpfulness. For the purpose of this study, it is defined as the helpful votes divided by the total votes received by a review. The features are classified into three categories: review content, review meta-data and product characteristics. The descriptions of these features are discussed in the next section.

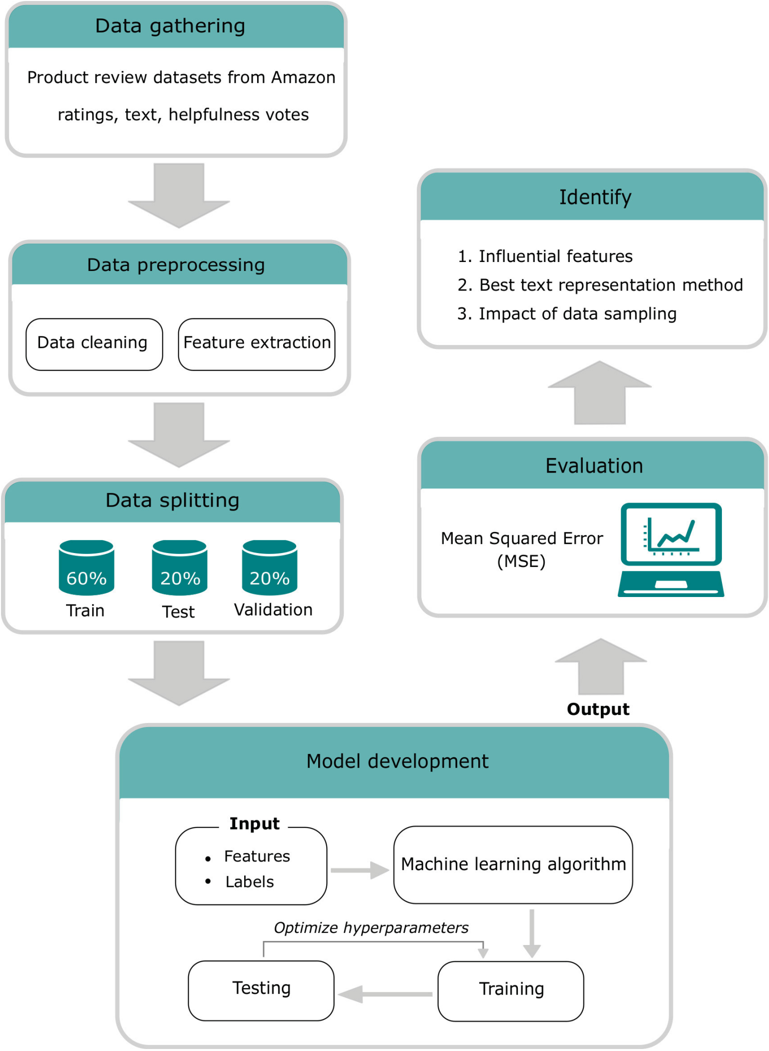


Figure 4.1: Research method

4.1 Features

Features for determining the helpfulness of reviews and the input to the machine-learning algorithms in the experiments, are categorized in three groups: review content, review meta-data and product features. The feature candidates are identified from recent literature and the data available in the datasets. Table 4.1 contains each of the features used in the experiments.

4.1.1 Review content features

As previous studies have proved the importance of review textual/content characteristics for helpfulness analysis [4][5], several content-related features are proposed in this research. The unigrams of a review are each word in it, which will be represented using various methods explained further down in this chapters in section 4.3. The length of review text and title in words and sentences are calculated using

Features		
Review content	UNI	Unigrams of review (words)
	LW	Length (in words)
	LS	Length (in sentences)
	LT	Length of title
	T	Title (words)
	SR	Subjectivity of review text
	ST	Subjectivity of title
	PR	Polarity of review
	PT	Polarity of title
	SE	Spelling errors
Review meta-data	R	Rating
Product	PO	Positive reviews
	CR	Critical reviews
	LN	Length of name
	LD	Length of description

Table 4.1: Selected features for each category with abbreviations

python. Subjectivity measures how much personal opinion vs factual information is contained in the text. Polarity measures the sentiment, or emotion of the text, which can be either negative, neutral or positive. Both subjectivity and polarity are quantified as a score which is calculated using textBlob, a language tool for NLP. The score describing subjectivity is between 0 and 1, the higher the score, the more personal opinions rather than factual information the text contains. The polarity score lies between -1 and 1, with -1 defining a negative sentiment and 1 a positive sentiment. The number of spelling errors contained in the text are calculated by using the python spell checker library.

4.1.2 Review meta-data features

The reviewer-assigned ratings of products on Amazon.com are values ranging from 1 to 5. A rating of 1 indicating the lowest satisfaction and 5 the highest. In this study the review rating will be used directly as a feature extracted from the datasets.

4.1.3 Product features

To investigate the impact of product determinants on review helpfulness, four product-related indicators are proposed in this study. The percentage of positive reviews for a product is calculated as well as the percentage of critical reviews. A review with a rating of at least 3 stars is considered as positive in this study and less than 3 stars is considered as a negative review. Additional proposed features are length of product name and length of product description.

4.2 Data-collection and preprocessing

Two categories of products from the public Amazon Review Dataset are utilised for the experiments. Electronics and Cell Phones And Accessories with 1689188 and 194439 data instances respectively. Reviews with less than 5 total votes are filtered out since these reviews may have a bias voting ratio.

Category	Reviews
Cell Phones and Accessories	10619
Electronics	207965

Table 4.2: Number of reviews of each product type after the preprocessing

The helpfulness ratios are calculated by dividing the helpfulness votes by the total votes. The total votes are also extracted and will be used as sample weights during training of the models. Reviews with more total votes will have higher weights when training the models.

For the cleaning of the textual data, the following steps will be performed:

- Special characters are removed
- Text is converted to all lowercase
- Contractions are expanded
- Stop words are removed (such as “the“, “a“, and “is“)
- Detected spelling errors are corrected (after the count of spelling errors is extracted as a feature)
- Stemming of the words

4.3 Text representation methods

This section will introduce the four text representation methods used in the experiments: Bag-of-words, td-idf, GloVe and Word2Vec.

4.3.1 Bag-of-words

A bag-of-words model (BoW) is a representation of text that specifies the occurrence of words within a text. The counting of each of the words contained in a text document is then represented as fixed-length vectors. The process of converting text into bag-of-words involve the following steps:

1. A vocabulary of unique words needs to be defined, which will be all the words found in all review texts of the dataset.
2. Count how many times each of the words from the determined vocabulary are present in each text document.

The insight of the method is that documents that are similar will have similar word counts. From the content alone we can learn something about the meaning of the document. As the model is only concerned with whether words occur in the document, some contextual information such as the order, structure or meanings of words, is discarded.

4.3.2 TF-IDF

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. This is calculated by multiplying two different metrics:

- The *term frequency* (how many times a word appears in a document).
- The *inverse document frequency* of the word across a set of documents (how common or rare a word is in the entire document set). The closer it is to 0, the more common a word is. This metric can be calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

Multiplying these two numbers results in the TF-IDF score of a word in a document. If a word is very common and appears in many documents, this number will approach 0. Otherwise, it will approach 1. The higher the score, the more relevant that word is in that particular document.

4.3.3 GloVe and Word2Vec

Word embeddings are word vector representations where words with similar meaning have similar representation. GloVe (Global Vectors) and Word2Vec are two of the most popular word embeddings and have proven their usefulness in many areas of NLP [7]. Because the two models differ in the way they are trained, their resulting word vectors have subtly different properties. Pretrained models for both these embeddings are accessible online and they are easy to incorporate for the experiments. All individual words are derived from a large dataset and are represented as real-valued vectors in a predefined vector space,

4.4 Data sampling

The dataset will be sampled using Stratified Sampling method, where the dataset is divided into subgroups based on their helpfulness ratio, i.e. in intervals of 0.1 ranging from 0-0.1 to 0.9-1.0 as demonstrated in Figure 4.2. Each group will then be sampled using Random Sampling method. A random selection of 150 reviews from each subgroup will result in a sample size of 1500 reviews.

4.5 ML models and evaluation metrics

Although the focus of this research is not intended to compare different machine-learning algorithms, I will make use of two different algorithms for the experiments:

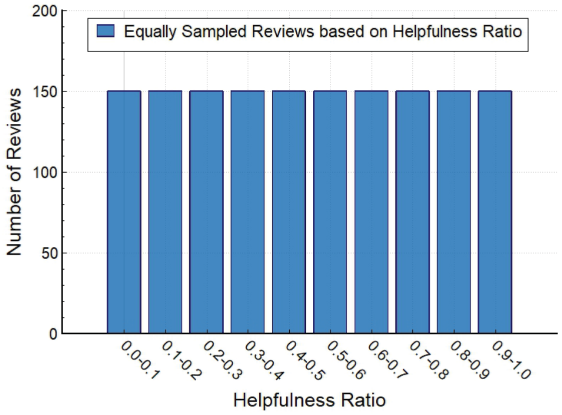


Figure 4.2: Equally sampled reviews for Cell Phones and Accessories dataset

Linear Regression and Convolutional Neural Networks. These algorithms are very different yet both widely-used in previous, related research [15][10]. The assumption is that it will contribute to producing a more comprehensive study with well-supported conclusions.

4.5.1 Ridge regression

A variant of Linear Regression, Ridge Regression, will be used in cases when data sampling is applied to the dataset. It is basically a regularized linear regression model and has the same usage. When we have a complex model with many independent variables, in this case each word in the review text among other features, there could arise a problem when only using a sampled number of instances. If there are more predictors than cases, it is no longer possible to estimate a unique, optimal solution. Regularization is applied to deal with this problem using Ridge Regression.

4.5.2 CNN

The architecture of the CNN model is presented in Figure 4.3. I have two convolutional layers followed by a max pooling layer. Then two dense layers and an output layer with sigmoid activation function.

As mentioned previously, the total votes for each review are used as additional weights for both models. Reviews with more total votes have higher weights than reviews with less votes.

4.5.3 MSE, RMSE and MAE

Three of the most popular metrics for evaluating predictions made on regression problems are MSE, RMSE and MAE [24]. These metrics involve calculating an error score to describe the predictive accomplishment of a model. The MSE, or Mean Squared Error, is calculated as the mean or average of the squared differences between predicted and actual target values in a dataset. The Root Mean Squared Error, or RMSE, is an extension of the mean squared error and is calculated as the square root of the MSE. MAE, or Mean Absolute Error, is as the name suggests,

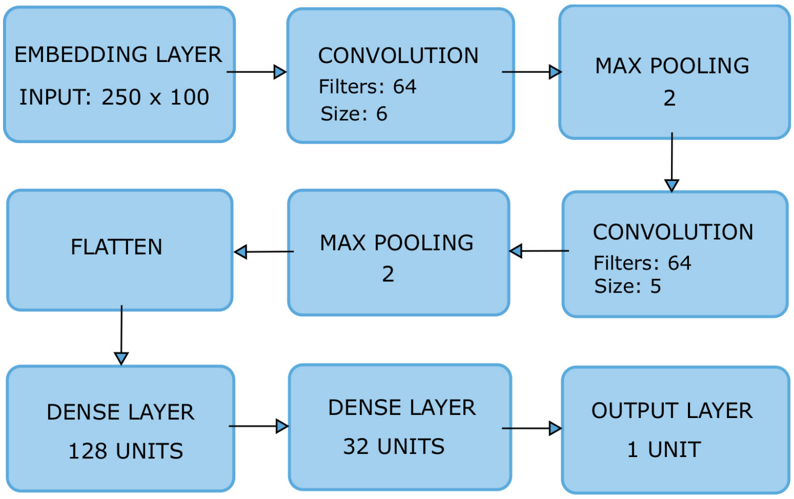


Figure 4.3: Architecture of CNN model

calculated as the average of the absolute error values. A perfect value for all of these metrics is 0.0, which means that all predictions matched the expected values exactly.

Chapter 5

Results and Analysis

In this study, various experiments were conducted to investigate predictive modelling of review helpfulness from several perspectives. The performance was compared using two machine learning regression models and popular MSE, RMSE and MAE evaluation metrics. The focus of the experiments are text representation analysis, sampling-importance analysis and proposed feature significance analysis on review helpfulness.

In the first experiment, two regression models were trained and tested on two datasets using unigrams of the review text as features. The performance of the models were compared using MSE, RMSE and RRSE evaluation metrics. The experimental results reveal the outperformance of the Ridge regression model which presents the lowest mean square error value as compared to the CNN model for both datasets, Cell Phones and Accessories and Electronics. Additionally, the Cell Phones and Accessories dataset presented better results as compared to the Electronics dataset. The ridge regression model were therefore utilized for the remaining set of experiments since it delivered the best results.

5.1 Text representation analysis

This section describes the results of the experiments conducted to examine the effectiveness of four types of representation of the review text to predict the helpfulness of product review using two datasets. Ridge regression is utilised as a regression model to conduct these experiments. The selected text representation methods are bag-of-words, tf-idf and two pre-trained embeddings Word2Vec and GloVe. Table 5.1 demonstrates the predictive performance of each method with and without sampling of the datasets using MSE, RMSE and MAE evaluation metrics. It is to be noted that lower values of these metrics indicate higher prediction accuracy.

The experimental results demonstrate that bag-of-words (BoW) method for representing the review text achieves the worst results for both datasets except for the sampled case using the Cell Phones and Accessories dataset. In this case, pre-trained word embeddings from GloVe obtained higher MSE value. This shows that frequency of each word is not the best representation of the review text for this task and that common words with the highest frequency may have little predictive power over the target variable.

Although there are similarities between tf-idf and bag-of-words by counting frequencies of words, their results differ quite a bit, having the worst and the best performances. The extension of tf-idf to weigh the occurrence of a word in a single document in relation to the entire corpus, proved useful for this task.

The choice of input word vector representation, between word2vec and GloVe, was shown to have an impact on performance. For this task, word2Vec performed consistently better.

Tf-idf and Word2Vec performed better overall with tf-idf having the best results.

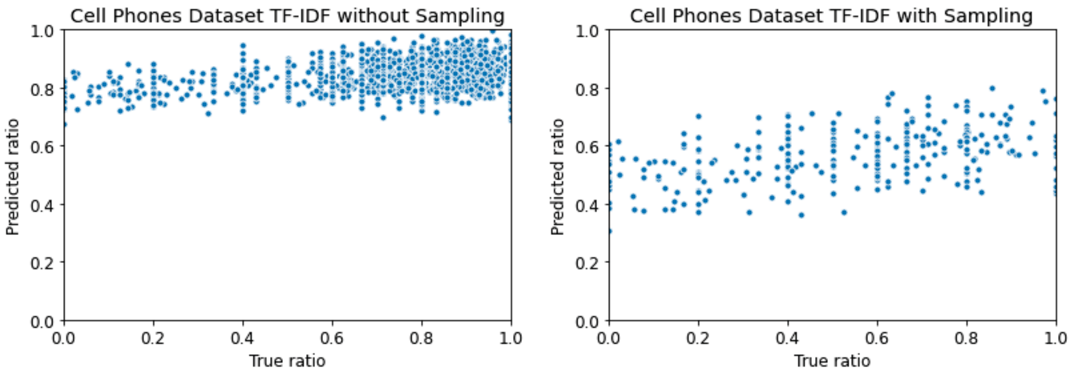
Cell Phones And Accessories						
Text Representation	Non-sampled			Sampled		
	MSE	RMSE	MAE	MSE	RMSE	MAE
BoW	0.0492	0.2218	0.1601	0.0704	0.2653	0.2193
TF-IDF	0.0491	0.2216	0.1595	0.0661	0.2572	0.2091
GloVe	0.0497	0.2230	0.1616	0.0787	0.2807	0.2347
Word2Vec	0.0483	0.2198	0.1601	0.0732	0.2706	0.2265
Electronics						
Text Representation	Non-sampled			Sampled		
	MSE	RMSE	MAE	MSE	RMSE	MAE
BoW	0.0567	0.2381	0.2781	0.0905	0.3008	0.2379
TF-IDF	0.0425	0.2062	0.1286	0.0751	0.2741	0.2273
GloVe	0.0517	0.2274	0.1493	0.0802	0.2833	0.2368
Word2Vec	0.0430	0.2073	0.1274	0.0795	0.2820	0.2368

Table 5.1: Comparison of prediction performance for text representation methods

5.2 Data sampling importance

In this section, the importance of data sampling in predictive modelling is examined and evaluated from the experiments performed using words of the review text as features. Figure 5 shows the results of sampling vs non-sampling method for assessing helpfulness. From these results it is evident that non-sampling approach generates a better MSE, RMSE and MAE than using a sampled dataset. When representing the unigrams using the tf-idf method, the non-sampled datasets for Cell Phones and Accessories and Electronics, present an MSE of 0.0491 and 0.0425 respectively. In the sampled case, the MSE presented is 0.0661 and 0.0751. Figure 5.1 shows a plotted graph of the prediction results using sampled and non-sampled dataset. As can be seen by the graph, there is greater variance in the predictions when using data sampling method, while the predictions from using unsampled dataset tend to be around 0.8-0.9. This kind of "naive" results obtained, with bias predictions in favor of the majority group of ratios, is due to the imbalanced dataset. Both datasets hold large volume of reviews at helpfulness range from 0.9 to 1 which indicates that the density of helpfulness for both datasets is skewed toward the right.

When comparing only the low ratios in the test dataset and only the high ratios separately, the MSE is shown in Table 5.2. These results show that reviews with lower



(a) (b)

Figure 5.1: Plotted predictions against actual helpfulness ratio for a) unsampled dataset and b) sampled dataset

ratios under 0.5 are more difficult to predict for both the sampled and unsampled datasets and have higher MSE than reviews with higher ratios. However, for the unsampled data, the MSE is notably worse with around 0.31 for both models. As could be seen from the graph in Figure 5.1 as well, where the low ratios are predicted to be much higher.

From the results it is evident that the unbalanced helpfulness ratios in the unsampled datasets affect the performance of both Linear Regression and CNN model, with biased predictions in favor of the majority group of high ratios in the dataset. There is low variance in the predictions which are mostly around 0.8. This kind of “naive” predictions, which is due to the imbalanced dataset, resulted in better Mean Squared Error (MSE), determined over the whole test dataset, than the sampled dataset in both Linear Regression and CNN model.

.	Cell Phones And Accessories		Electronics	
	Sampled	Non-sampled	Sampled	Non-sampled
Overall	0.0663	0.0751	0.0515	0.0425
High rate	0.0470	0.0551	0.0208	0.0111
Low rate	0.0877	**0.3010**	0.0942	**0.3120**

Table 5.2: MSE of prediction results of high helpfulness ratios (>0.5) and low ratios (<=0.5) in sampled and non-sampled datasets

5.3 Feature analysis

This section covers the experiments performed to examine proposed input features. The objectives of these experiments are to investigate how much each feature contributes to review helpfulness and which feature contributes the most. The importance of the proposed review, review meta-data and product features using the Cell

Phones and Accessories dataset is presented in Figure 5.2. The significance is evaluated on the basis of mean square error. The result demonstrated that the rating, unigrams of the review text and polarity of the review title are the most important features. The rating is the most influential feature which could be considered as proxy for sentiment of the review. This indicates that reviews which have high positive or negative sentiment attract more helpful votes. The polarity of title is the second most influential feature of the review type which indicates that more positive or negative titles of reviews intrigue more people to read the review, which will therefore receive more helpful votes. Among the product-related features, the percentage of positive and negative reviews demonstrates the best performance. The length of product name and description are the least effective features of product type.

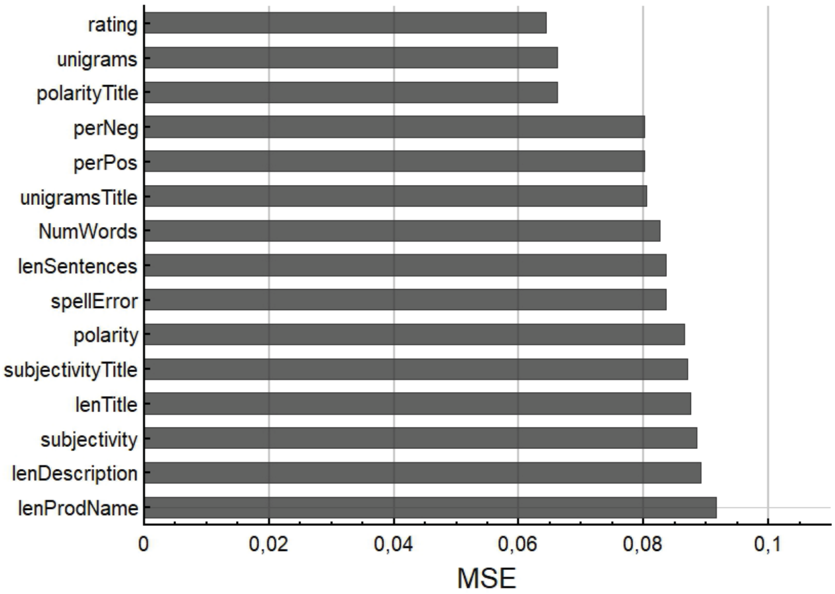


Figure 5.2: Feature importance using Cell Phones and Accessories dataset

For the Electronics dataset, the importance of proposed features is shown in Figure 5.3 in which the Mean square error metric is utilized to analyze the performance of each feature. The tf-idf representation of the review text is again the most influential feature and outperforms, obtaining the lowest MSE value. Other influential review type features are review title, length of review text in words and length in sentences with the review title as the second most important feature. The rating and polarity of title are not identified as influential to the same degree as by using the Cell Phones and Accessories dataset in which the rating was the most influential feature. The effectiveness of the various product features also differs between the two datasets. In the Electronics dataset, the length of product name and length of product description demonstrates the best performance.

Among the three categories of features - review, review meta-data and product - review features are the most influential for both datasets. However, for the Cell Phones and Accessories dataset, review meta-data demonstrates comparable significance in capturing review helpfulness. The experimental results of feature category importance are presented in Table 5.4. Using a hybrid combination of the most in-

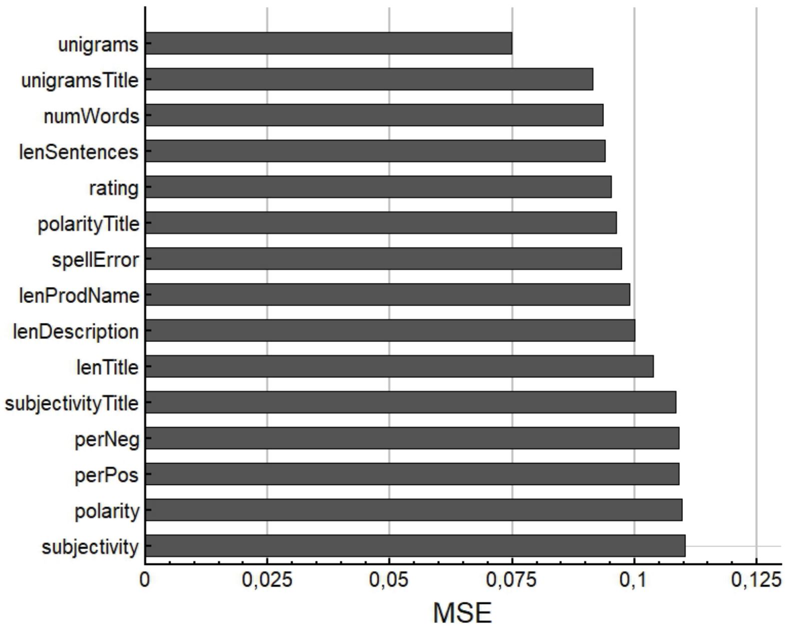


Figure 5.3: Feature importance using Electronics dataset

fluential features for each dataset achieves the best results and using all the features generates the second best result.

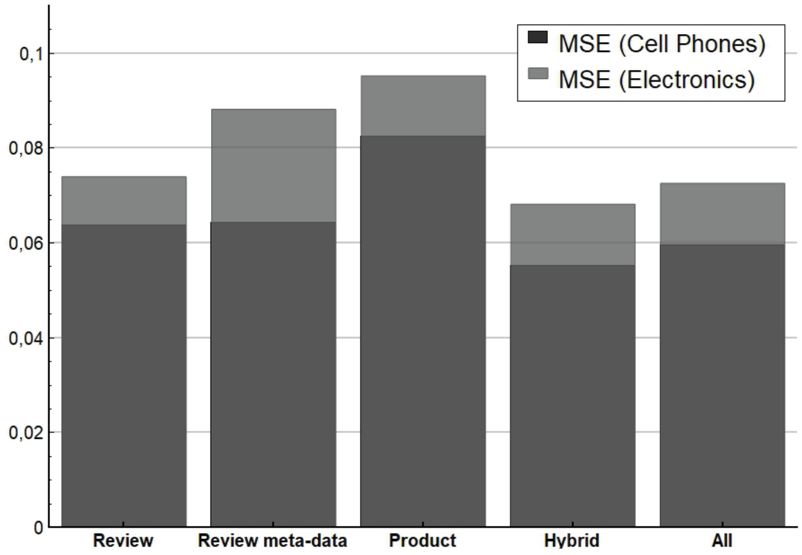


Figure 5.4: Feature category performance analysis using both dataset

Chapter 6
Validity Threats

There are several limitations to this study and some points which could be considered as validity threats. Several of the intended features to be used in this study are not considered due to data availability issues. These features include reviewer type features and some additional product-related features such as price and product image. Using these features would contribute to a more comprehensive study of the impact of product-related features to predict helpfulness.

Additionally, the current study did not verify the reliability and validity of the review features extracted by some of the NLP tools used such as textBlob and NLTK, for calculating the subjectivity and polarity scores of the review text and title as well as counting spelling errors. Although these tools have been utilized in prior research, further studies are necessary to verify whether they work well for analyzing review text. Also, the code used for the experiments has not been reviewed by a peer which could also be a validity threat to the results.

Another point concerns the experimental input data, the Amazon product review datasets, and the validity of these reviews. Several articles[1][2] have exposed problems of brand hijacking and fake reviews on Amazon where the reviews were written for entirely different products. These inaccuracies in the input data could potentially affect the validity of the experimental results. If a review was actually written for a different product than the one presented to the user reading the review, the incorporated product-related features would not be accurate. In this case, the results of the experiments using these features would be affected to some extent, and not necessarily all of the experiments conducted. However, if the faulty information of the product would influence the perceived helpfulness of the review due to the mismatch of the product information and the written review, this would be another threat adding noise in the data. These issues are not taken into consideration due to the complex nature of such assessment.

Finally, as the experiments were implemented only using two category datasets from Amazon.com datasets, the possibilities to generalize are more limited. The category of product could contribute to the helpfulness of reviews and the reviews for these two categories of products may not be representative for all product reviews. In order to get more general results, this study should be expanded by obtaining additional datasets from other categories and possibly from other e-business companies as well.

[1] https://kleinbottle.com/#AMAZON\%20BRAND\%20HIJACKING

[2] https://www.consumerreports.org/customer-reviews-ratings/hijacked-reviews-on-amazon-can-trick-shoppers/

Chapter 7

Conclusions and Future Work

In this paper, the important problem of predicting the helpfulness of reviews was considered by examining three research questions. Extensive experiments on the Amazon product review dataset have been conducted with the aim of shedding some light on some of the major factors affecting the perceived helpfulness as well as the modelling task. The study in this paper has focused on the online product reviews, but the approach is general enough to be adapted to other domains as well.

What is the impact of different word representations on model performance for predicting helpfulness of product reviews? The choice of representation of the textual data was confirmed by the experiments to have an impact on performance. However, as the results were not completely consistent between the two datasets, further research is required in order to conclude which representation is better suited for this task. The results do however indicate that tf-idf is considerably more effective than bag-of-words which is the least effective of the selected methods. Of the pre-trained embeddings, word2Vec performed consistently better than GloVe for both datasets. However different representations may perform better for different tasks.

How does data sampling techniques influence the performance of predictive modelling for regression? The importance of data sampling in predictive modelling for regression tasks was evaluated using equal-data sampling technique. Even though an overall lower MSE value was obtained from using non-sampled dataset, the results were revealed to be "naive" and highly biased due to the imbalanced dataset dominated by high helpfulness ratios. Selecting a well-balanced sample from the large dataset covering all the variations found in the large dataset by using an effective and unbiased data sampling technique was proved to be effective for addressing such problems.

What input features among review and product are the most important to capture the helpfulness of a product review? Previous studies revealed that review and review meta-data characteristics are the most important determinants for helpfulness prediction [4][5][18]. Our research supports the findings of prior studies and demonstrates the outperformance of the proposed review and review meta-data features. Among the many features evaluated, unigram tf-idf yielded strong predictive power for both dataset, demonstrating the effectiveness of encoding semantics for helpfulness prediction. Due to some inconsistency between the two datasets, the results are however not sufficient to conclude which of the other features are the most effective within the three categories of features. Among the product related-features, the result from each of the datasets was completely contradicting. Using the Cell Phones and Accessories dataset, the most influential product features were the percentage of positive and negative reviews, however, for Electronics, the length of product name and description demonstrated higher significance for predicting review helpfulness. Further research is necessary to conclude what product features capture helpfulness better.

Several significant extensions can be made in future studies. One of the future extensions is to explore other platforms. Since social media has become an important marketing channel, a prospective future work would be to compare review characteristics written in social media with reviews posted on online shopping sites. Another possible extension would be to study some of the proposed features in more depth. For example, the influence of word count could further be studied by investigating if there is a threshold for the number of words in its effect on review helpfulness. Finally, as review manipulation has become an increasing problem for platforms such as Amazon, predicting the validity of reviews might be even more important than predicting helpfulness in future studies.

References

[1] J. McAuley and J. Leskovec, “Hidden factors and hidden topics: understanding rating dimensions with review text”, In *In RecSys*, 2013, pp. 165–172

[2] A. Ławrynowicz and V. Tresp, “Introducing Machine Learning”, In book: *Perspectives on Ontology Learning*, 2014, vol. 18, pp. 35-50.

[3] S. Shalev-Shwartz and S. Ben-David, “Introduction”, In book: *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, USA. 2014, pp. 19. [Online]. Available: http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning

[4] M.S.I. Malik, “Predicting users’ review helpfulness: the role of significant review and reviewer characteristics”, In *Soft Comput.*, 2020, vol. 24, pp. 13913–13928.

[5] S. 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*, ser. EMNLP ’06, 2006, pp. 423-430

[6] Liddy, E.D. “Natural Language Processing”. In *Encyclopedia of Library and Information Science*, 2001, 2nd Ed. NY. Marcel Decker, Inc.

[7] M. Ekman, *Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, NLP, and Transformers using TensorFlow.* Addison-Wesley Professional, 2021.

[8] V. Ganganwar, “An Overview of Classification Algorithms for Imbalanced Datasets”. In *International Journal of Emerging Technology and Advanced Engineering*, 2012, vol. 2(4), pp. 42-47.

[9] G. Lemaître, F. Nogueira, and C.K. Aridas, “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”, In *Journal of Machine Learning Research*, 2017, vol. 18(1), pp. 1-5.

[10] X. Qu, X. Li and J.R. Rose., “Review Helpfulness Assessment based on Convolutional Neural Network”, *arXiv preprint arXiv:1808.09016*, 2018.

[11] Y. Liu, X. Huang, A. An and X. Yu, “Modeling and Predicting the Helpfulness of Online Reviews”, In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM ’08, 2008, pp. 443–452.

[12] 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”, In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, Copenhagen, Denmark, 1358–1363.

[13] Y. Yang, C. Chen and F. S. Bao, “Aspect-Based Helpfulness Prediction for

Online Product Reviews", In *International Conference on Tools with Artificial Intelligence (ICTAI)*, San Jose, CA, USA, 2016, pp. 836-843.

[14] Y. Almutairi and M. Abdullah, "IRHM: Inclusive Review Helpfulness Model for Review Helpfulness Prediction in E-commerce Platform", In *Journal of Information Technology Management*, 2020, pp. 184-197.

[15] C. Chen, M. Qiu, Y. Yang, J. Zhou, J. Huang, X. Li, and F. S. Bao, "Multi-Domain Gated CNN for Review Helpfulness Prediction", In *The World Wide Web Conference*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 2630–2636.

[16] A. H. Huang, K. Chen, D. C. Yen and T. P. Trang, "A study of factors that contribute to online review helpfulness", In *Computers in Human Behavior*, 2015, vol. 48, pp. 17-27.

[17] A. Ghose and P.G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics", In *IEEE Transactions on Knowledge and Data Engineering*, 2011, vol. 23(10), pp. 1498–1512.

[18] Y. Yang, M. Qiu, Y. Yan and F.S Bao, "Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews", In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL-2015)*, 2015, vol. 2, pp. 38-44.

[19] S. Mudambi, and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com", In *MIS Quarterly*, 2010, vol. 34, pp. 185-200.

[20] Y. Park, "Predicting the Helpfulness of Online Customer Reviews across Different Product Types", In *Sustainability*, 2018, vol. 10, pp. 17-35.

[21] Q. Cao, W. Duan W and Q. Gan, (2011) "Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach", In *Decision Support Syst 50*, 2011, vol. 50, pp. 511–521.

[22] 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*, 2018, pp. 1-4.

[23] M. Passon, M. Lippi, G. Serra and C. Tasso, "Predicting the Usefulness of Amazon Reviews Using Off-The-Shelf Argumentation Mining", In *Proceedings of the 5th Workshop on Argument Mining*, 2018, pp. 35-39.

[24] A. Botchkarev, "A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms", In *Interdisciplinary Journal of Information, Knowledge, and Management*, 2019, vol. 14, pp.45-76.