ABSTRACT

Title of Thesis: STUDYING PRODUCT REVIEWS USING SENTIMENT ANALYSIS BASED ON INTERPRETABLE MACHINE LEARNING

Pranjal Atrey
Master of Science, 2023

Thesis Directed by: Professor Sanghamitra Dutta and Professor Min Wu
Department of Electrical and Computer Engineering

Consumers' reliance on product reviews and ratings has been making substantial impacts on purchasing behaviors in e-commerce. However, the relationship between reviews and ratings has received limited attention. For instance, a product may have a high rating but average reviews. Such feedback can cause confusion and uncertainty about the products, leading to decreased trust in the product. This thesis carries out a natural-language based machine learning study to analyze the relationship from e-commerce big data of product reviews and ratings. Towards answering this relationship question using natural-language-processing (NLP), we first employ data-driven sentiment analysis to obtain a numeric sentiment score from the reviews, which are then used for studying the correlation with actual ratings. For sentiment analysis, we consider the use of both glass-box (rule-based) and black-box opaque (BERT) models. We find that while the black-box model is more correlated with product ratings, there are interesting counterexamples where the sentiment analysis results by the glass-box model are better aligned with the rating.

Next, we explore how well ratings can be predicted from the text reviews, and if sentiment scores can further help improve classification of reviews. We find that neither opaque nor glass-box classification models yield better accuracy, and classification accuracy mostly improves when BERT sentiment scores are augmented with reviews. Furthermore, to understand what different models use to predict ratings from reviews, we employ Local Interpretable Model-Agnostic Explanations (LIME) to explain the impact of words in reviews on the decisions of the classification models. Noting that different models can give similar predictions, which is a phenomenon known as the Rashomon Effect, our work provides insights on which words actually contribute to the decision-making of classification models, even in scenarios where an incorrect classification is made.

STUDYING PRODUCT REVIEWS USING SENTIMENT ANALYSIS BASED ON INTERPRETABLE MACHINE LEARNING

by

Pranjal Atrey

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2023

Advisory Committee:
Professor Sanghamitra Dutta, Chair/Co-Advisor
Professor Min Wu, Co-Advisor
Dr. Michael Brundage

Acknowledgments

I would like to express my gratitude towards my advisors, Dr. Sanghamitra Dutta and Dr. Min Wu, for their support and guidance throughout my journey at the University of Maryland. Their guidance and feedback has tremendously helped in developing my technical skills. It has been an honor to work with such intelligent individuals.

I would like to thank Dr. Michael Brundage for being part of the committee and taking the time to review my thesis. He has also played an integral part in planning the topic of my thesis and providing support throughout.

I would like to thank my program director, Ms. Emily Irwin, for her support during my graduate degree. She has been a support pillar and has taken the time out to sort out any issues that aroused during my academic journey.

Lastly, I would like to thank my family and friends for their support during the completion of my graduate degree. My parents, Pradeep and Manisha Atrey, have been there to support me in every stage of my journey and I am extremely grateful to have such loving and encouraging parents. Thank you to my sister and brother-in-law, Akanksha Atrey and Meet Shah, for their continuous advice and guidance. Through their experiences, they have given me valuable guidance on how to successfully grow my career. Thank you to my best friend, Saumya Pandey, for pushing me through stressful periods and motivating me to become a better person. This wouldn't have been possible without you all and I am immensely privileged to have you in my life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AM	Automotive
BERT	Bidirectional Encoder Representations for Transformers
CPA	Cell Phones and Accessories
DF	Document Frequency
IDF	Inverse Document Frequency
LIME	Local Interpretable Model-Agnostic Explanations
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes
NLP	Natual Language Processing
OP	Office Products
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency
TF	Term Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner
VG	Video Games

Chapter 1

Introduction

1.1 Motivation

The process of buying and selling goods and services over an electronic platform has become ubiquitous in the last decade [9]. The pervasiveness of smartphones has enabled consumers to utilize popular e-commerce applications, such as Amazon and Etsy, at the touch of their fingertips. Further expanding this need has been the COVID-19 pandemic, which proved e-commerce to be an essential part of everyday life. This rapidly expanding nature of e-commerce platforms has directly interested service providers, researchers, and economists to study consumer behavior surrounding product purchases. When making decisions on which products to purchase, consumers often rely on reviews from previous users of the good or service. Consumers' reliance on product reviews and ratings has evolved e-commerce, leading to substantial impacts on purchasing behaviors. Recent studies show that 83% of customers read online reviews before purchasing a product [2]. While both reviews and ratings provide feedback about a product, reviews focus on textual feedback and ratings represent a numerical score describing the product (as shown in Figure 1.1).

Interestingly, prior work [31] has demonstrated that reviews and ratings are not always

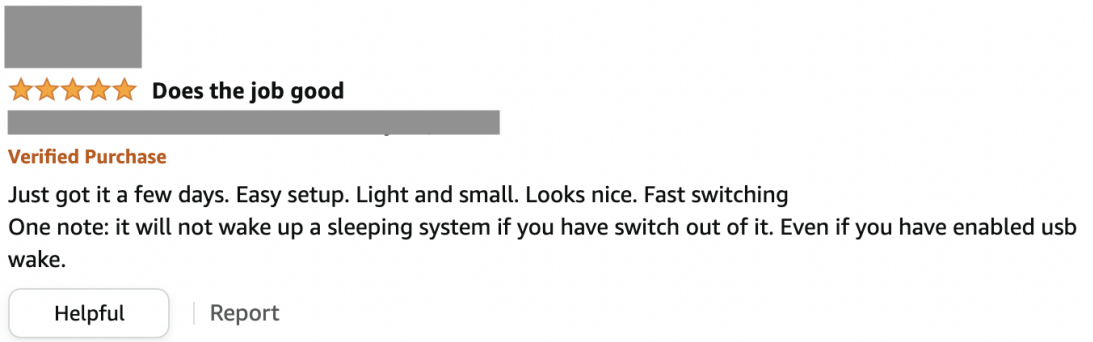


Figure 1.1: An example of text feedback and numerical rating in Amazon product reviews.

correlated. For instance, a product may have a high rating but average reviews, and vice-versa. Such feedback can cause confusion and uncertainty about the products, leading to decreased trust in the product. Hence, it is important to gather a deeper understanding of the relationship between reviews and ratings, and understand when they agree or disagree. As we will show in this thesis, recent advancements in natural language processing (NLP) have enabled a deeper study to understand such relationships [4]. However, the relationship between reviews and ratings has received limited attention. Our work is also closely aligned with Rashomon effect in explainability which refers to the phenomenon of different models performing similar predictions on ML tasks [5].

1.2 Thesis Contributions

The objective of this thesis is to study the relationship between product reviews and ratings through the lens of different approaches ranging from glass-box to opaque/black-box models. In doing so, NLP methods are adapted to evaluate various facets of reviews and ratings, including the sentiment of reviews, the nature of the relationship between reviews and ratings, and the contextual contribution of different words in reviews to their corresponding ratings. The relationship

is analyzed via supervised machine learning (ML) techniques. That is, ML models are employed to predict ratings using review text, and their performance is used as a proxy for the relationship.

To this end, the thesis contributions are:

1. Comparing an opaque method called BERT with a glass-box, rule-based method called VADER for reviewing tone of product reviews.

2. Using different classifiers ranging from glass-box linear classifiers to opaque neural network models and understanding their differences using explainability methods such as LIME.

1.3 Thesis Outline

The outline of this thesis is as follows. In Chapter 2, relevant research on the relationship between reviews and ratings is reviewed, and the datasets being studied in this thesis are discussed. Chapters 3 and 4 contain the main contributions of the thesis work that focuses on the relationship between reviews and ratings using NLP techniques. In Chapter 3, (i) both glass-box and black-box sentiment analysis tools are employed to understand the relationship between reviews and ratings, and (ii) attention-based models are employed to convert reviews into low-dimensional representations called embeddings. These embeddings are used alongside corresponding sentiment scores to predict ratings using glass-box and opaque classification models. Chapter 4 extends the work in Chapter 3 by using LIME to explain and understand the contribution of different words in the reviews to classification performance across different NLP models. Finally, Chapter 5 summarizes the work and discusses future research opportunities.

Figure 1.2 shows a detailed view of the work done in this thesis. 10,000 product reviews are converted to numerical values using three embedding methods. Subsequently, two sentiment

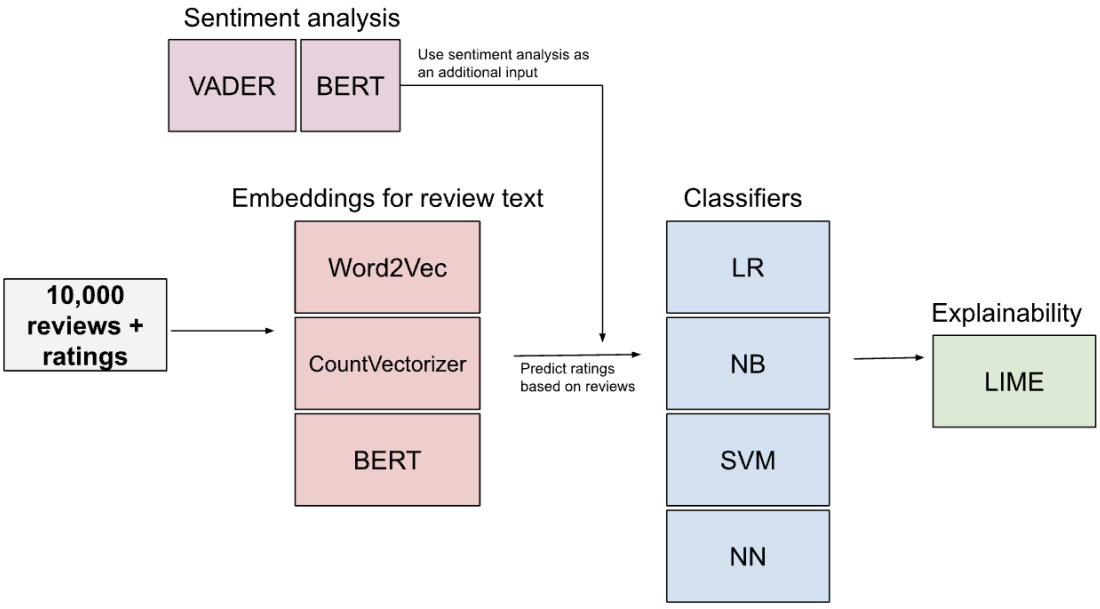


Figure 1.2: Flow diagram of work done in this thesis

analysis methods are compared. Four classification models are further used to assess the performance of predicting ratings based on reviews. Sentiment scores are used as an additional input to assess the change in performance. Finally, LIME is used to explain the impact of using different embedding methods and classification models on specific words in the reviews.

Chapter 2

Related Work and Datasets

In this chapter, first, related work corresponding to the relationship between reviews and ratings in Section 2.1. Next, in Section 2.2, the dataset used in this work is described.

2.1 Related Work

In this section, the work related to the problem addressed in this thesis is discussed from two aspects: (i) the work related to sentiment analysis of text reviews, and (ii) the work that studied the relationship between text reviews and numerical ratings. The work in these two categories is described below in Section 2.1.1 and Section 2.1.2, respectively.

2.1.1 Sentiment Analysis of Text Reviews

Sentiment analysis is a natural language processing technique to identify positive and negative sentiments in text [1]. The simplicity and applicability of traditional sentiment analysis methods have found usability in various domains, including finance [45], politics [46], and education [47].

Many recent works have focused on sentiment analysis of text reviews. Fang and Zhan examined sentiment polarity categorization on Amazon reviews using sentiment analysis [11].

In another study, Salinca [12] performed a classification of business reviews using sentiment analysis on the Yelp Challenge dataset. With the advent of deep learning models, many works have also attempted sentiment analysis for high-dimensional data. Hu et al. [13] used a deep neural network for sentiment analysis. Similarly, Tsao et al. [18] performed a passive analysis of consumer-generated textual data from service reviews. Further, Zhang et al. [20] performed sentiment analysis of e-commerce text reviews and classified them based on the constructed emotional dictionary.

More recent works analyzing the sentiment of textual reviews have considered complex relationships. Murthy et al. [21] performed LSTM-based sentiment analysis to capture the longitudinal aspects of sentiment across large texts. Lee et al. [28] proposed a semi-supervised approach requiring a small training dataset. The authors found this approach to be comparable to supervised learning models that are trained on large datasets. Most recently, Mutinda et al. [32] presented a sentiment classification model, LeBERT, which combines sentiment lexicon, N-grams, BERT, and convolutional neural networks.

Other works have focused on sentiment analysis of reviews in specific application domains, such as Hu et al. [14] for online restaurant reviews, Thet et al. [15], Baid et al. [16], and Devi et al. [19] for movie reviews. A number of survey papers on text-based sentiment analysis that has been published within the past decade such as [22], [23], [24], [26], [29], [25].

Prior work has also demonstrated the use of LIME explainability in online consumer reviews [25]. However, it has has only demonstrated the general usage of LIME rather than an in-depth study of different classification and embedding models. It has also not focused on specific examples which study the Rashomon effect phenomenon [5] [6] [7].

Key Takeaway: While prior works have considered variants of sentiment analysis techniques to capture the tone of reviews, we compare the difference between a rule-based model called VADER (glass-box) and an opaque model called BERT (black-box) to understand where they agree and disagree with the models. BERT-based sentiment analysis enables the examination of sentence-level sentiments, allowing for better representation. BERT-based sentiment analysis is compared with a rule-based sentiment analysis method called VADER.

2.1.2 Relationship Between Text Reviews and Numerical Ratings

While sentiment analysis has been a great mining tool to evaluate the correlations between text reviews and numerical ratings, sentiment alone does not capture the full contexts of textual data. Instead, some prior works have used alternative methods in NLP to review this relationship. One of the first works in this problem domain was from McAuley and Leskovec [10]. In this work, the authors examined rating dimensions with review topics learned by topic models like LDA. Later, Zhou and Yang looked into how the numerical and textual reviews across three different review types including comparative, suggestive, and regular reviews [17]. The authors found that numerical scores are more important than textual characteristics across three different review types. In another study by Umer et al. [27], the numeric ratings of Google apps were predicted using machine learning classifiers. In their study, it was found that the machine learning-based classifiers were able to predict authentic numeric ratings based on actual user reviews.

Recently, Alantari et al. [30] focused on the fundamental relationship between a consumer's overall empirical evaluation, and the text-based explanation of their evaluation. Further, Almansour et al. [31] studied whether text reviews are always consistent with the combined numeric evaluations,

and found that the quality of the rating scores used for sentiment analysis models is questionable as it might not reflect the sentiment of the associated reviews texts. The authors also emphasized the need to quantify the relationship degree between the text reviews and the scores to understand the quality of rating scores.

Key Takeaway: While prior works have examined varying facets of the relationship between reviews and ratings, our focus is on understanding the mechanics of different models ranging from linear classification to opaque models such as neural networks. To achieve this goal, we employ LIME to understand the contribution of each word toward the rating. This allows us to examine and gather insights on the impact of each word deeply.

2.2 Datasets

The datasets used in this study are retrieved from a larger database consisting of a large crawl of product reviews from Amazon [10]. The database contains 82.83 million unique reviews from approximately 20 million users. The reviews are in text format while the ratings are in numerical format ranging from 1-5.

For the purpose of analyzing product reviews and ratings, 40,000 product reviews and ratings have been extracted from the following four product categories: Cell Phones and Accessories (CPA), Office Products (OP), Automotive (AM), and Video Games (VG), as illustrated in Figure 2.1. Using four different categories enables investigation of the generalizability of the results.

In order to ensure a balanced dataset, each category consists of 2,000 reviews and ratings grouped by each rating star (1-5), consisting of a total of 10,000 reviews. Though there are

Figure 2.1: Summary of the datasets used in this work.

many pre-processed features available in the dataset, this thesis employs only reviews and ratings. Before using the datasets for ML modeling, the datasets have been cleaned. Stop words such as *the* and *and* has been removed from the reviews to capture the more important words, and punctuation marks have also been removed.

Chapter 3

Analyzing the Relationship Between Product Reviews and Ratings

This chapter explores the correlation between product reviews and ratings. Since reviews are textual data and ratings are numerical scores, it is usually not possible to compare them as they are. Sentiment analysis is an approach in natural language processing that focuses on understanding positive and negative sentiments in text. In the context of reviews, a higher sentiment score represents a positive review, while a lower sentiment score represents a negative review. Additionally, another method of understanding the relationship between reviews and ratings is to assess the performance of classification models to predict ratings based on reviews. A higher performance would indicate a stronger relationship as learned by the classification model. Many embedding methods and classification models can be compared via performance metrics.

In this chapter, the following studies are performed:

(i) The correlation between reviews and ratings is studied via BERT-based sentiment analysis and compared against a popular traditional method (in Section 3.1).

(ii) Classification models are used to predict ratings based on reviews as a proxy measure for correlation (in Section 3.2).

(iii) Embeddings of review text are supplemented with BERT-based sentiment scores to augment the performance of the classification models (in Section 3.2).

3.1 Sentiment Analysis

Two types of sentiment analysis methods are explored in this study: (i) Valence Aware Dictionary and sEntiment Reasoner (VADER) [48], and (ii) Bidirectional Encoder Representations for Transformers (BERT) [54]. These two methods are described in Section 3.1.1. Correlation metrics are discussed in Section 3.1.2. Subsequently, experiments and results of sentiment analysis are described in Section 3.1.3.

3.1.1 Sentiment Analysis Methods

VADER is a lexicon and rule-based sentiment analysis method [48]. In addition to a gold-standard sentiment lexicon that is specifically attuned to sentiments expressed in social media, VADER uses the following five general rules to express sentiment intensity:

(1) Punctuation. Exclamation points increase the magnitude of intensity in a sentence. For instance, *"The product was great."* is more intense than *"The product was great!"*.

(2) Capitalization. Using ALL-CAPS to emphasize certain sentiment-related words increases the magnitude of intensity in a sentence. For instance, *"The product was GREAT!"* is more intense than *"The product was great!"*.

(3) Degree modifiers. Certain words can be used to increase or decrease the intensity of a sentence or word. For instance, For instance, *"The product was extremely good!"* is more intense than *"The product was moderately good"*.

(4) Constrastive junctions. Certain words such as *but* can cause the sentiment polarity to

switch. For instance, *"The product was good but the service was horrible."* has mixed sentiment, but the latter half is weighed more towards the overall sentiment.

(5) Tri-grams. By examining the tri-gram preceding a sentiment-specific lexical feature, negation flips can be detected better which changes the polarity of a sentence. For instance, a negated sentence would be *"The product wasn't really all that great"*.

VADER sentiment scores range between -1 and 1, where 1 represents positive text and -1 represents negative text. In order to keep the range of ratings and sentiment scores the same, min-max scaling is used to convert the VADER sentiment scores to range between 1 and 5 in floating point numbers. In addition to the advantages mentioned in the five rules above, VADER has many other advantages; It does not require any training data as it uses a lexicon-based approach and it has evidently worked well with social media and reviews text.

The other sentiment analysis method explored in this study is BERT [54], which was introduced by researchers from Google. BERT is a machine learning model used for NLP tasks. It focuses on pre-training deep bidirectional representations which read text input as a sequence of words at once. In other words, the context of a word can be understood by learning about its surroundings (left and right of the word) as opposed to reading sequentially. The pre-trained BERT model can be fine-tuned with an additional output layer for tasks such as sentiment analysis. For sentiment analysis, HuggingFace *bert-base-multilingual-uncased* model fine-tuned for sentiment analysis on product reviews is used. It predicts the sentiment of the review as a number of stars between 1 and 5. The fine-tuned model uses product reviews ranging in 6 different languages as training input. Since the model is uncased, it does not make a difference between capitalized and uncapitalized words. It uses factors such as punctuation to

better understand the context of words and phrases.

In comparison, VADER and BERT are two different kinds of sentiment analysis methods. BERT requires pre-training while VADER uses a rule-based approach and does not require pre-training. Hence, BERT takes significantly more time to compute sentiment scores compared to VADER. Specifically, for the purpose of identifying sentiment in product reviews, BERT is expected to perform better than VADER as it is more tailor-made for product reviews compared to VADER. BERT has been trained on product reviews whereas VADER uses a lexicon attuned for social media context. Additionally, BERT has the ability to understand contextual information as well.

3.1.2 Measuring Correlation Between Sentiment Scores and Ratings

The relationship between ratings and product reviews is studied by analyzing the correlation between sentiment scores of text reviews and corresponding ratings. Pearson's correlation, which is shown in Equation 3.1 is used to measure the strength of the linear relationship between the two variables. It has a range between -1 and 1, where 1 represents a total positive correlation, 0 represents no correlation, and -1 represents a total negative correlation.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

In the above equation,

r represents the correlation coefficient between measurements x and y,

n denotes the number of sample points, denoted by x_i for x and y_i for y, $1 \leq i \leq n$, and

$\bar{x}$ and $\bar{y}$ are the sample mean for x, and y, respectively.

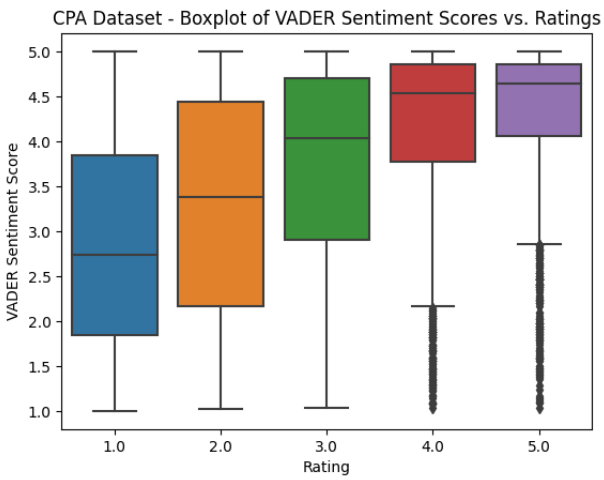
3.1.3 Experimental Results

Table 3.1: Correlation between sentiment scores of reviews and ratings

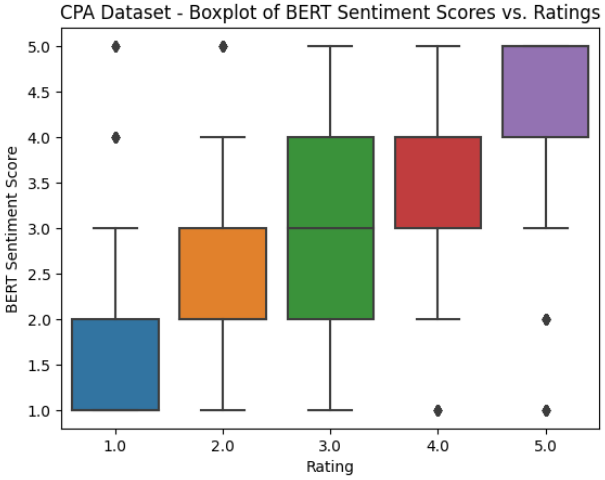
Dataset	Correlation with VADER	Correlation with BERT
Cell Phone and Accessories	0.4414	0.7675
Office Products	0.5228	0.8202
Automotive	0.4320	0.7871
Video Games	0.3972	0.7653

Table 3.1 depicts the correlations between the ratings and both types of sentiment scores in the four datasets used in this experiment. It can be seen that the correlations between the actual ratings and BERT sentiment scores are much higher than the ones with VADER. There is a moderate positive correlation between the VADER sentiment scores and ratings as the correlations are mostly between 0.3 and 0.5. There is, however, a strong positive correlation between the BERT sentiment scores and ratings as the correlations are between 0.5 and 1.

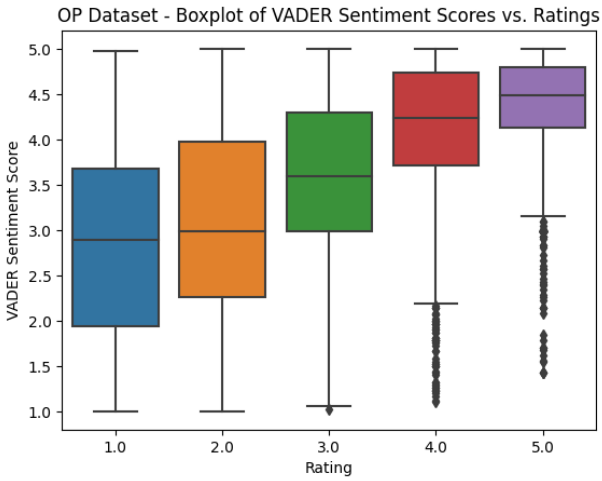
The relationship between reviews and ratings can also be observed via box plots, as shown in Figure 3.1. Box plots are a standardized way of visualizing the distribution of a data set. They focus on 5 points; the "minimum", first quartile (Q1), median, third quartile (Q3), and "maximum". The interquartile range is covered in the box, which is the 25th to 75th percentile. Box-plots of both VADER and BERT sentiment scores are plotted against ratings for two datasets: Cell Phones and Accessories (Figures 3.1a and 3.1b), and Office Products (Figures 3.1c and 3.1d). From the box plots, it can be observed that VADER sentiment scores are more skewed than BERT sentiment scores. For instance, the interquartile range for reviews with a 1-star rating in the CPA dataset for VADER sentiment scores is between around 2 and around 3.8. On the other hand,



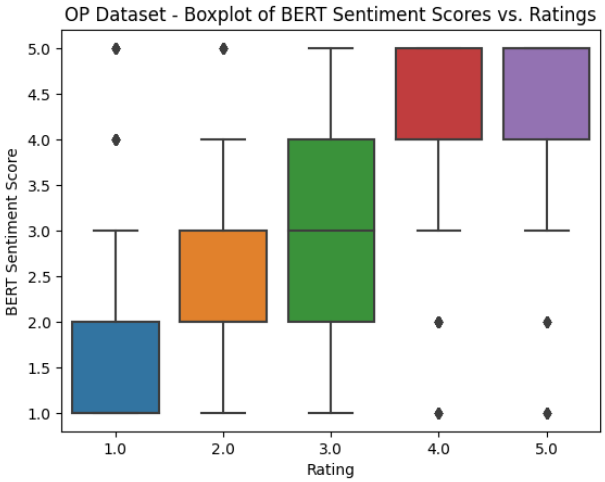
(a) VADER Boxplot in CPA dataset



(b) BERT Boxplot in CPA dataset



(c) VADER Boxplot in OP dataset



(d) BERT Boxplot in OP dataset

Figure 3.1: Box plots of sentiment scores vs. ratings in CPA and OP datasets

the interquartile range for reviews with a 1-star rating in the CPA dataset for BERT sentiment scores is between around 1 and 2. The visualizations depict that the correlations are significantly more accurate for BERT sentiment scores than VADER sentiment scores. In addition, it can also be seen that there are many outliers in the data for all rating levels. Outliers refer to sentiment scores that are not closely related to the actual ratings and are denoted by dots on the box plots. Such outliers signify the importance of assessing whether sentiment analysis is a strong method to understand the relationship between reviews and ratings. Hence, it is important to understand the underlying issues behind the difference in reviews and ratings.

Lastly, another way of assessing sentiment analysis methods is to plot median and mean sentiment values for all ratings, as shown in Figure 3.3. Similar to the box-plots in Figure 3.1, medians and means of both VADER and BERT sentiment scores are plotted against ratings for two datasets: Cell Phones and Accessories (Figures 3.3a and 3.3b), and Office Products (Figures 3.3c and 3.3d). An ideal plot would depict a linear relationship with an identity line between the median/mean and ratings. Based on the results, it can be seen that the relationship between median/mean BERT scores and ratings is more linear as compared to the median/mean VADER scores. In fact, Figure 3.3a depicts a perfect identity line for median BERT scores.

Based on the computed correlations and visualizations, it can be concluded that BERT sentiment analysis is more correlated with ratings than VADER sentiment analysis. As mentioned previously, BERT is expected to be more accurate than VADER as it is more tailor-made for sentiment analysis in product reviews compared to VADER, and has the ability to understand contextual information in text. However, it requires training models which can be time-consuming. Conversely, VADER understands some of the contexts based on the rules but fails to understand all possible contexts. Since it is rule-based, there is no previous training required. Though both

```
Example #1:
Review: I purchased this item sold directly from Amazon...It is as
"Genuine Motorola" as it gets...It works perfectly...AT&T; wanted
$29.00..Amazon wanted $8.38...Beware of some other sellers!!as
their items are not genuine and ruin your battery if they even
work at all!!
Rating: 5
BERT Sentiment Score: 1
VADER Sentiment Score (scaled): 4.06

Example #2:
Review: The case for the 8310. This is the real deal the oem case.
The case turns phone off and on
Rating: 5
BERT Sentiment Score: 1
VADER Sentiment Score (scaled): 3
```

Figure 3.2: Examples in the dataset where VADER scores are closer to ratings than BERT scores

sentiment analysis have their own advantages, BERT is more correlated with the ratings VADER based on performance.

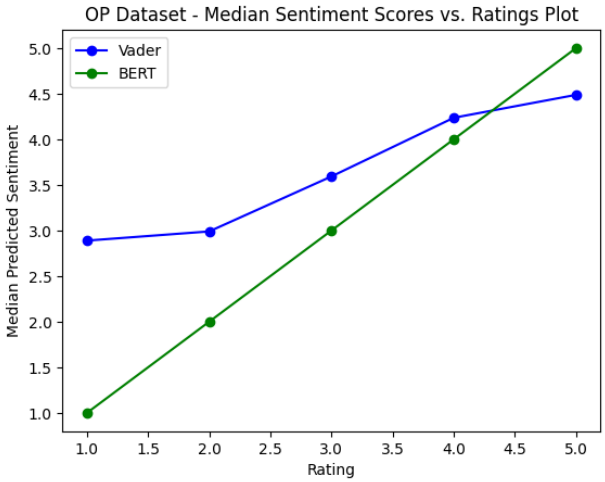
Based on the experiments, it can be inferred that though the relationship between reviews and ratings is moderately strong, there are certain cases where reviews and ratings are not aligned. In fact, there are many instances when VADER provides more accurate results than BERT. For instance, Figure 3.2 showcases some examples where VADER is providing a sentiment score closer to the actual ratings compared to BERT. This depicts that sentiment analysis solely cannot be used to understand the relationship between reviews and ratings. Hence, classification models will also be explored to study the relationship in Section 3.2. However, it is possible that genuinely ratings are not correlated with sentiment scores in some scenarios.

3.2 Classifying Ratings Based on Reviews

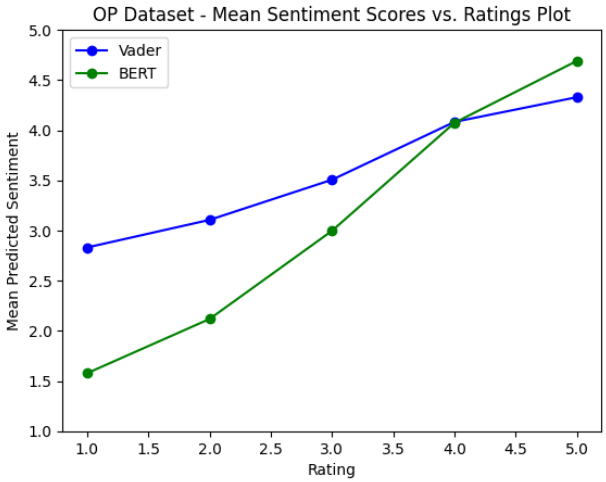
In this section, classification models are used to predict numerical ratings based on text reviews. Traditional ML models are not able to take textual data as input. Hence, vectorization

(a) Median plot in CPA dataset

(b) Mean plot in CPA dataset



(c) Median plot in OP dataset



(d) Mean plot in OP dataset

Figure 3.3: Plots of means and medians of sentiment scores vs. ratings

and embedding methods are needed to convert the raw textual data into a vector format of real numbers which ML models support. Multiple types of vectorization and embedding methods are discussed in Section 3.2.1. Different classification models used to assess performance are discussed in Section 3.2.2. Subsequently, evaluation metrics and corresponding experimental results of classifying ratings based on reviews are described in Section 3.2.3 and 3.2.4 respectively.

3.2.1 Embedding Methods

The process of using embeddings involves tokenization, which is breaking down sentences into individual words and converting them to vectors based on certain rules and/or methods. In this thesis, we use three types of embeddings as follows.

(1) CountVectorizer is a vectorizer that converts text into vector format based on the frequency of words in the given corpus. It creates a matrix where rows represent the documents and columns represent tokens. The cell values indicate the frequency of each token in each document. CountVectorizer is computationally efficient and simple to use. However, it fails to understand the semantic relationship and context between words.

(2) Term Frequency - Inverse Document Frequency (TF-IDF) is another text embedding technique that transforms text into a usable vector [49]. It is a combination of two concepts: term frequency (TF) and document frequency (DF). Term frequency refers to the number of occurrences of a word in a document. Document frequency refers to the number of documents containing a specific word. Inverse document frequency (IDF) is referred to as the weight of a term, that reduces the weight of a term if the term occurs throughout all the

documents. The TF-IDF is calculated as shown in Equation 3.2.

$$\textit{TF-IDF} = TF * IDF \tag{3.2}$$

where, TF is the number of times a term appears in the document divided by the total number of terms in the document, and IDF is the log of the number of documents in the corpus divided by the number of documents in the corpus that contain the term. In addition to considering the frequency of words present (e.g., CountVectorizer), TF-IDF also takes into account the importance of the words. This allows the model to be less complex since less important words are being removed.

(3) Word2Vec is a frequently used word embedding method which focuses on which words occur with other words more often [50]. The semantic closeness between certain words is mathematically close to the vector values of those words. Unlike in TF-IDF, Word2Vec uses an unsupervised learning process in which unlabeled data is trained via neural networks to create a Word2Vec model that creates word vectors. Unlike most embedding methods, the vector size does not have to be as much as the number of unique words in the corpus. The size of the vector can be parameterized based on the size of the dataset and requirements. For instance, a corpus of 1000 words would require a vector size of 1000 in One Hot Encoding, which is a popular word embedding method. However, Word2Vec allows low-dimensional representation of vector size to define a larger corpus. In the example, a vector size of 100 would suffice for a corpus of 1000 words, therefore avoiding complex computations of vectors. The Word2Vec model used in this study employs a pre-trained model which consists of Google News corpus (3 billion running words) vectorized into 3

million 300-dimension English word vectors.

3.2.2 Classification Models

Four types of classification models are employed for predicting ratings using vectorized reviews as follows.

(1) A support vector machine (SVM) separates a given dataset to the best of its ability by using multiple hyperplanes to split the data and choosing the best hyperplane available [51]. To find the optimal hyperplane, an SVM looks at the support vectors of the classes that are being compared, and uses the margin between these support vectors to determine if the hyperplane is satisfactory; the larger the vector margins between the classes the better. Smaller vector margins indicate a poor choice of the hyperplane. Having a good hyperplane helps the classifier define which data belongs to a class, thus providing a more accurate result.

(2) Logistic regression (LR) is a binary classification method that uses a logistic function to model the dependent variable [52]. LR is primarily designed for binary classification. As this study involves multi-class classification, the LR classifier is trained on just the examples belonging to one class versus all the examples of all other classes. Similarly, all other classes are also independently classified against all others. The multi-class classification problem is viewed as multiple binary classification problems. After the classifiers have learned to distinguish their chosen class from other classes, the classifiers are run and the most confident LR classifier gets used.

(3) Naive Bayes (NB) classification is based on using probabilistic ML models [53]. They are

quite popular in text classification problems. The classifier is based on Bayes' theorem, as shown in Equation 3.3.

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{3.3}$$

Where, $P(\theta|\mathbf{D})$ denotes the posterior probability, $P(\theta)$ is the prior probability, $P(\mathbf{D}|\theta)$ is the likelihood, and $P(\mathbf{D})$ is the evidence.

Using the Bayes theorem above, we can find the probability of θ happening, given the occurrence of **D**. Unlike discriminative classifiers, like logistic regression, NB does not learn which features are most important to differentiate between classes. However, it is quite fast as the probabilities can be directly computed without any iterations. Multinomial NB is used in the experiments as there are multiple classes to be classified.

(4) Neural networks have proven to perform well in classification tasks. A neural network architecture has been used in this study for classification. The neural network is trained using a learning rate of $1e-4$ with two hidden layers of size 512 and 128, respectively, with ReLu activations embedded in between. The output layer is of size 5 and uses *softmax* activation. The training is executed using Adam optimizer with 20 epochs.

Classification models can also be seen as opaque and glass-box models. Logistic regression and naive bayes are viewed as glass-box models while SVMs and neural networks are considered to be opaque models [8].

3.2.3 Evaluation Metrics

Embedding methods and classification models are assessed with the four datasets. The classifier aims to classify ratings based on vectors representing the review text. All models are trained using 70% (7,000) of the data and tested using the remaining 30% (3,000). The following performance metrics are used to analyze the results of the classification models:

- Accuracy is the most straightforward way to analyze the performance of classifiers. It requires actual and predicted classes for each data point. Accuracy can be computed by dividing the number of correct predictions by the total number of predictions.

- Confusion matrices are visual depictions of the results where each column in the matrix represents a class. A n-class classifier would produce a nxn matrix to represent the performance of the classifier. They group the results into four categories:

  - True positive (TP), when both actual and predicted values are 1

  - True negative (TN), when both actual and predicted values are 0

  - False positive (FP), when the actual value is 0 but the predicted value is 1

  - False negative (FN), when the actual value is 1 but the predicted value is 0

- F1-score is the harmonic mean of precision and recall. Precision is a measure of how accurate a model's positive predictions are while recall recall explains how many of the actual positive cases we were able to predict correctly with our model. The equations to calculate precision, recall, and F1-score are shown in Equations 3.4, 3.5, and 3.6.

$$Precision = \frac{TP}{TP + FP} \tag{3.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.5}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.6}$$

3.2.4 Experimental Results

Table 3.2 depicts the accuracy results of three embedding methods (discussed in Section 3.2.1) and four classification models (discussed in Section 3.2.2) used on the four datasets. The accuracies are mostly between 40% and 60%.

Based on the results, it can be seen that TF-IDF embeddings are performing better than CountVectorizer and Word2Vec regardless of the classification model. Compared to CountVectorizer, this is explainable as TF-IDF also takes into account the importance of words in the entire corpus. TF-IDF can perform better than Word2Vec in certain cases. Such factors that may impact the results are using multi-class vs. binary classification, imbalanced vs. balanced datasets, and the type of data being classified. It cannot be concluded that one classification model is performing better than the others. Based on different datasets and embedding methods, all classification models have better accuracies in certain cases. For instance, the LR classifier has the best accuracy with CountVectorizer in the OP dataset. However, the SVM classifier has the best accuracy with TF-IDF embeddings in the same dataset. This suggests that no single classifier solely can be said to yield the best accuracy overall. A trend that is seen in the table is that using

Table 3.2: Classification of ratings based on reviews

	Vectorization	Logistic Regression	Naive Bayes	Support Vector Machine	Neural Network Architecture
CPA	CountVectorizer	44.77%	45.03%	46.00%	49.00%
	TF-IDF	48.07%	44.00%	48.03%	48.80%
	Word2Vec	43.67%	36.10%	44.67%	40.10%
OP	CountVectorizer	52.87%	52.30%	50.73%	53.60%
	TF-IDF	53.80%	54.30%	**55.70%**	54.57%
	Word2Vec	47.53%	31.33%	49.43%	45.80%
AM	CountVectorizer	46.23%	48.13%	44.87%	47.13%
	TF-IDF	49.93%	49.26%	49.77%	48.13%
	Word2Vec	42.47%	28.10%	44.77%	39.67%
VG	CountVectorizer	47.97%	46.60%	44.10%	50.43%
	TF-IDF	50.23%	43.07%	49.23%	51.63%
	Word2Vec	45.13%	30.7%	46.83%	42.70%

the NB classifier with Word2Vec embeddings gives significantly lower accuracies compared to other models. The accuracies are between 28% and 37%, which is around 15% less accuracy compared to other results. We hypothesize the reason behind this trend can be explained by the key assumption of NB classifiers. The NB classifier assumes that all input features in the model are conditionally independent, or unrelated to any of the other features. Since Word2Vec focuses more on the contextual understanding of words and phrases compared to its surrounding, using a NB classifier may not be helpful as it would consider the vectors to be independent and fail to capture the context. Additionally, it cannot be concluded whether glass-box models such as LR and NB perform better than opaque models such as SVM and neural networks. The accuracy results fail to recognize one type of model which performs better for all cases.

Using TF-IDF embeddings with the SVM classifier for the OP dataset yields the most accuracy of 55.70%. Figure 3.4 depicts the confusion matrix for that classification. All the percentages totaled in the matrix add up to 55.70%, which is the accuracy of the classifier. As can

be seen from the matrix, the diagonal values represent instances when the true label matches the predicted label. The ratings are classified mostly into their respective classes or the surrounding rating classes. For instance, the 2-star rating reviews are being classified mostly into 1-star, 2-star, and 3-star rating classes. However, there has been incorrect classification in all classes as no single value in the matrix is 0.

Precision, recall, and F-1 scores have been compared for the most and least accuracies in Table 3.3. As depicted in the table, the F-1 scores for SVM with TF-IDF are significantly higher than the F-1 scores for NB with Word2Vec. A larger F-1 score represents excellent precision and recall. To iterate, the F-1 score is the harmonic mean of the precision and recall. The F-1 scores for both examples shown are relatively higher for the 1-star and 5-star ratings. This means that the classifier finds it easier to correctly classify 1-star and 5-star ratings compared to other ratings. Such an observation is expected as it's easier to state whether a review is really good or really bad compared to predicting that a review is somewhat good or somewhat bad.

Table 3.3: Precision, recall, and F1-scores of most and least accuracy classifiers

	SVM with TF-IDF (55.70% Accuracy)			NB with Word2Vec (28.10% Accuracy)		
Rating	Precision	Recall	F1-Scores	Precision	Recall	F1-Scores
1	0.58	0.58	0.58	0.47	0.15	0.23
2	0.46	0.45	0.46	0.23	0.75	0.36
3	0.46	0.51	0.48	0.40	0.08	0.13
4	0.56	0.52	0.54	0.32	0.07	0.11
5	0.73	0.72	**0.72**	0.34	0.35	**0.35**

BERT sentiment scores are used as an additional input to the classification model to assess any change in performance. It is expected that the sentiment scores will cause the accuracy of

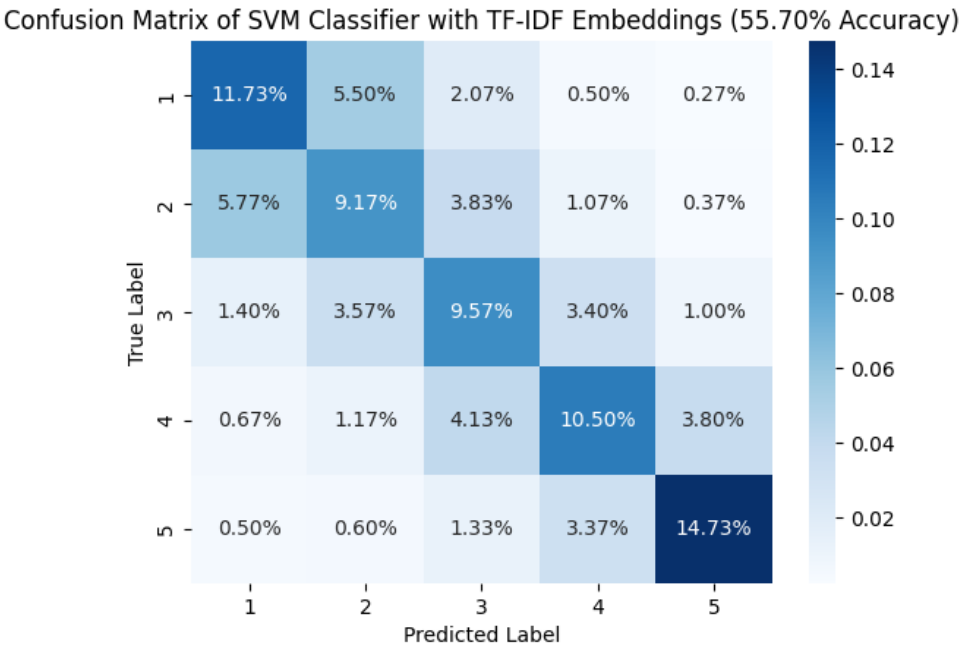


Figure 3.4: Confusion matrix of SVM classifier with TD-IDF embeddings yielding 55.70% accuracy

the classifier to improve. Table 3.4 depicts the results of the accuracies with BERT sentiment score as an additional input. The change in accuracy after adding the sentiment scores is also shown in parentheses beside the accuracies in the table. The classification accuracies mostly improve. The largest increase in accuracy is 9.93% while the largest decrease in accuracy is 5.26%. The decrease in accuracy is mostly when NB classification is used. Specifically, using TF-IDF embeddings with the NB classifier always results in a decrease in accuracy when BERT sentiment scores are used as an additional input.

3.3 Summary

In this chapter, the relationship between reviews and ratings was studied via ML and NLP methods such as sentiment analysis and classification. Two sentiment analysis methods, VADER and BERT, were studied and compared. Classification models were used to predict ratings based

Table 3.4: Using BERT sentiment scores as additional input to classifiers

	Vectorization	Logistic Regression	Naive Bayes	Support Vector Machine	Neural Network Architecture
CPA	CountVectorizer	49.40% (+4.63)	46.13% (+1.10)	54.00% (+8.00)	47.66% (-1.34)
	TF-IDF	53.70% (+5.63)	37.87% (-6.13)	53.47% (+5.44)	48.40% (-0.40)
	Word2Vec	53.40% (+9.73)	39.73% (+3.63)	53.46% (+8.79)	41.03% (+0.93)
OP	CountVectorizer	57.03% (+4.16)	51.70% (-0.60)	57.90% (+7.17)	53.87% (+0.27)
	TF-IDF	58.97% (+5.17)	50.03% (-4.27)	56.40% (+0.7)	53.80% (-0.77)
	Word2Vec	56.07% (+8.54)	32.80% (+1.47)	57.30% (+7.87)	46.23% (+0.43)
AM	CountVectorizer	50.63% (+4.40)	47.03% (-1.10)	54.70% (+9.83)	48.60% (+1.47)
	TF-IDF	54.60% (+4.67)	44.00% (-5.26)	53.93% (+4.16)	48.27% (+0.14)
	Word2Vec	52.30% (+9.83)	29.10% (+1.00)	53.90% (+9.13)	40.56% (+0.89)
VG	CountVectorizer	50.83% (+2.86)	48.20% (+1.60)	**54.03% (+9.93)**	51.27% (+0.84)
	TF-IDF	54.33% (+4.10)	43.80% (+0.73)	53.16% (+3.93)	50.93% (-0.70)
	Word2Vec	53.23% (+8.10)	30.73% (+0.03)	53.27% (+6.44)	42.80% (+0.10)

on reviews, with a comparison between multiple embedding methods and ML models. Sentiment scores were additionally used with reviews to classify ratings. The following points summarize the key takeaways in this chapter:

- While both BERT and VADER sentiment analysis methods have a positive correlation with ratings, BERT is more correlated to the actual ratings compared to VADER due to it being more susceptible to product reviews and its ability to capture the context in the text better.
- Though BERT is more correlated than VADER, there are many cases in which VADER may be more ideal to use than BERT. Hence, more analysis is required via classification.
- Compared to the classifiers discussed in Section 3.2.2, it cannot be concluded whether black-box models depict better results than glass-box models. Compared to other embedding methods discussed in Section 3.2.1, TF-IDF embeddings are better suited for converting raw text in the Amazon dataset to numerical vector models.

- Classification accuracy improves for most models when BERT sentiment scores are augmented with review embeddings as input to the classifier.
- Classifiers tend to mostly predict the exact or surrounding (off by 1) ratings.
- Classifiers tend to predict 1-star and 5-star ratings more precisely than other rating classes.

Chapter 4

Assessing Classification Models Using Local Explanations

4.1 Overview

In the previous chapter, sentiment analysis and classification models were used to assess the relationship between reviews and ratings. Though performance metrics discussed in Section 3.2.4 can be helpful in understanding such models, specific words in textual data that impact the predictions are not discussed. ML models are mostly black boxes in nature due to their high complexities. It is important to understand how classifiers in a given model work. For instance, certain words in a review may impact the prediction of a rating more than others. Moreover, different models can give similar predictions, a phenomenon also known as the Rashomon Effect. Understanding the impact of certain words on decision-making can further help interpret such a phenomenon. An approach known as LIME [33] (Local Interpretable Model-agnostic Explanations) has been introduced which provides the ability to explain certain predictions of classification models, specifically for text.

In this chapter, the following studies are performed:

(i) LIME is discussed in details (in Section 4.2).

(ii) Experiments are conducted using LIME with multiple embedding methods and classification

models to gain a deeper understanding of how reviews are classified with an emphasis on which words contribute to the decisions (in Section 4.3).

4.2 Local Interpretable Model-agnostic Explanations (LIME)

Local Interpretable Model-agnostic Explanations (LIME) is an approach that can be used to explain the predictions of any classification or regression model [33]. LIME works with the objective of identifying an interpretable model which can present an interpretable representation that can be trusted locally at any stage of the classification.

The key aspects of LIME include interpretable data representations, fidelity-interpretability trade-offs, sampling for local exploration, and sparse linear explanations. The first aspect, interpretable data representations emphasize that the representations that are used for interpretable explanations must be understandable to humans, no matter what the actual features the model uses. For the second aspect, fidelity-interpretability trade-offs in LIME are accomplished by minimizing a measure of interpretability as well as the complexity with which the interpretations are understood by humans. In the third aspect, the sampling for local exploration works on presenting explanations that are locally trustworthy even though it may be complex to explain the original model globally. The fourth and final aspect of LIME, sparse linear explanations, works on the principle that we can estimate the faithfulness of certain explanations and use it to select a set of explanations across a set of multiple interpretable model classes.

The following steps are used to explain a prediction using LIME:

(1) Choose a prediction which we want to be explained.

(2) Permutations of this instance are created, and model predictions are collected.

(3) LIME then assigns weights to the new created samples based on how closely they match with the data of the original prediction.

(4) A new, less complex, interpretable model is trained on the data variations created using the weights attached to each variation.

(5) The prediction can be explained by this local interpretable model.

Since LIME was presented the first time in 2016, there have been many derivative works based on it. For instance, Bramhall et al. [34] redefined the linear relationships presented by LIME as quadratic relationships, and presented a variation of LIME as qLime. Further, Zafar and Khan [35] proposed a deterministic version of LIME, called DLIME. Also, Zhao et al. [36] presented a Bayesian variation of LIME, called BayLime. Recently, Zhong et al. [37] proposed an enhanced Bayesian version of LIME, called EBLIME.

In the past few years, LIME has been used in many applications, such as medical [38], [39], [40], [41], time-series forecasting [42], object detection [43], and image classification [44]. In this work, LIME is used for explainable analysis of the product reviews.

4.3 Experimental Results

The embedding method and classification model are specified into a pipeline which is then used in the LIME explainer to understand certain predictions. Due to sequential models not being supported for LIME text explainers, the neural network architecture discussed in Section 3.2.2 is not used for experiments in this section. In this section, observations are made with LIME on explaining predictions.

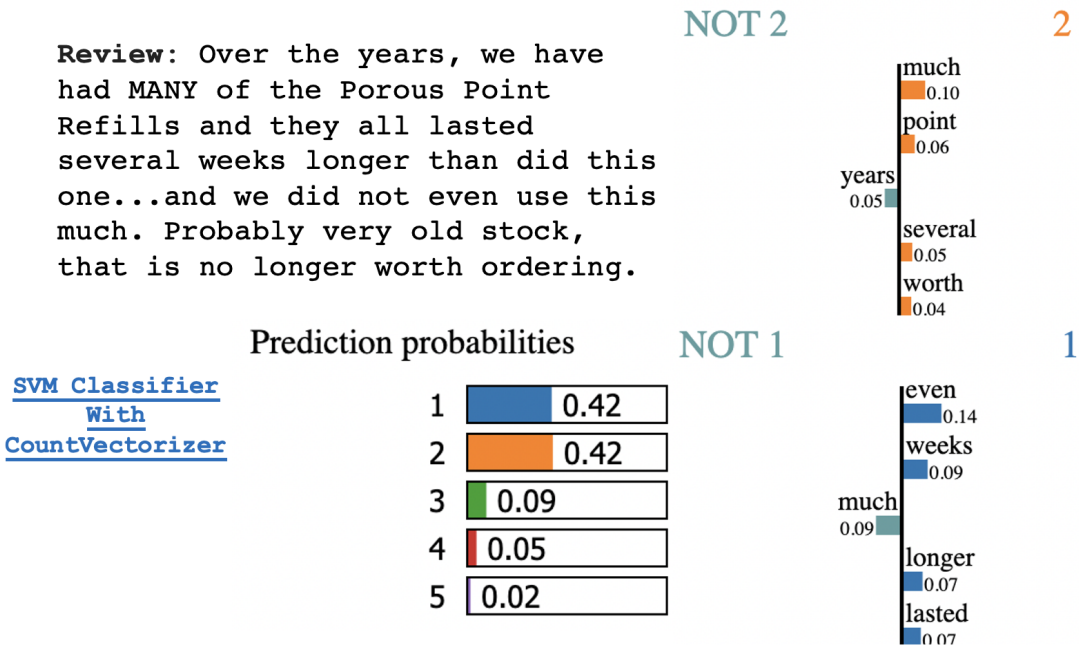


Figure 4.1: Comparing different words for same prediction probabilities of SVM and CountVectorizer using LIME (True label is 2)

Observation 1. LIME can be used to compare how different words impact classification despite having the same prediction probabilities. Figure 4.1 shows an example of a review being predicted using an SVM classifier with CountVectorizer embeddings. The prediction probabilities show how confident the classifier is in classifying each of the classes. As can be seen from the figure, both 1-star and 2-star ratings have the same prediction probability (42%). However, the words used to classify in each other classes are different. The model uses words like *even*, *weeks*, *longer*, and *lasted* to classify for 1-star ratings. On the other hand, the model uses words like *much*, *point*, *several*, and *worth* to classify for 2-star ratings. Despite having the same prediction probabilities for the same text, the words used to classify in each of the classes are different.

Observation 2. Classification models may use different embedding models to make the same predictions. However, the predictions may be based on different features which is referred

to as the Rashomon effect. For instance, different words may be weighed more than others based on the embedding methods. Figure 4.2 shows such an example. A review is being classified (using SVM) as a 2-star rating with both TF-IDF and Word2Vec embeddings. They have been incorrectly rated as the true label is a 1-star rating. The importance of this example is in the words that are used in the prediction probabilities. TF-IDF used with the SVM classifier uses the words *dont*, *light*, and *used* to predict that the rating is 2-star. On the other hand, Word2Vec uses the words *unhappy*, *dont*, and *buy* to predict that the rating is 2-star as well. Different words (other than *dont*) are used and make the same prediction. This example shows that different embedding methods may place an emphasis on different words to make the same predictions. Another important takeaway from this example is to realize how close the classifier is to predicting the correct label (1-star rating). In the TF-IDF example, the model is 30% confident about predicting a 1-star rating while it is 36% confident about predicting a 2-star rating. LIME can also be used to assess how closely a model was confident in predicting the correct label in cases of incorrect prediction-making.

Observation 3. Classification models can be compared to assess which words impact the right or wrong classification. Figure 4.3 shows an example of this observation as three classification models are used with TF-IDF embeddings and only two out of the three models correct the correct label (1-star rating). LR and SVM classifiers correctly predict the review to be a 1-star rating, whereas the naive bayes classifier incorrectly predicts the label to be 3-star rating. As it can be seen, the same words (*terrible*, *horrible* and *even*) are used to classify in LR and SVM classifiers. It is also important to note that LR classifier accounts *terrible* as 19% to the overall prediction probability whereas SVM classifier accounts the same word as 21% to the overall prediction probability. This infers that SVM more confidently uses the word, *terrible*, to

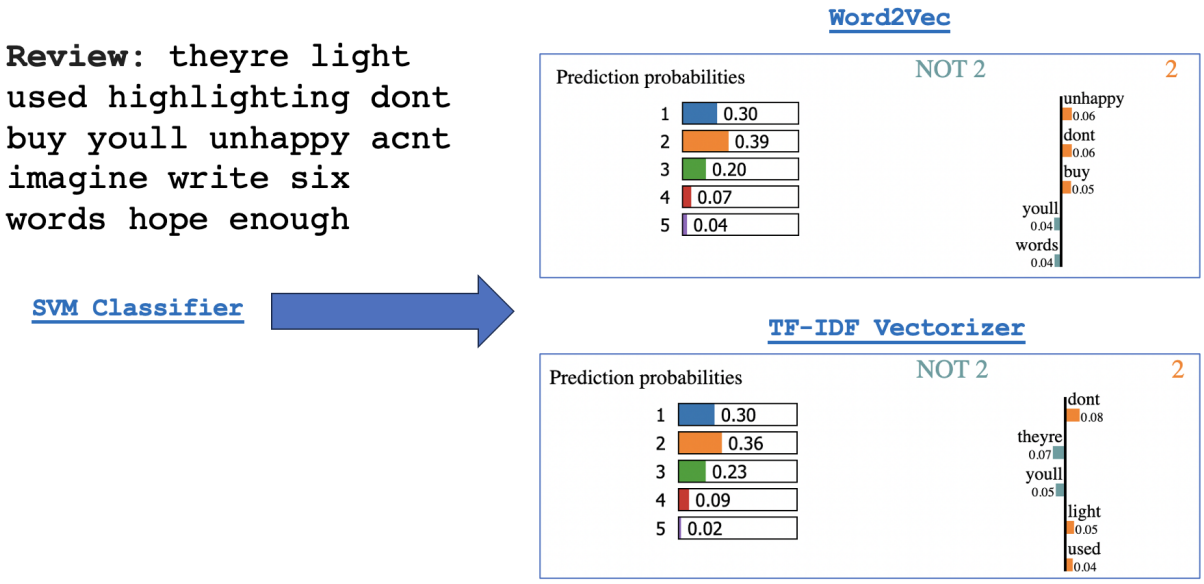


Figure 4.2: Different words being used to make the same predictions for different embedding methods (True label is 1)

predict a review as 1-star rating. On the other hand, the NB classifier places a higher emphasis on *decent*, *older*, *cartoonish*, and *rendering* to predict the review as 3-star rating. It can also be seen that the prediction probabilities in NB classifier are more closer to each other whereas the other classifiers are more confident about predicting as 1-star rating. This shows that NB classifier is not as confident about one particular class compared to other classification models. The model should be placing more importance on words like *terrible* and *horrible* rather than *decent* and *older*. Based on the assessment of words using LIME, the LR and SVM classifiers are better for this specific example review.

4.4 Summary

In this chapter, LIME is discussed and experiments are conducted with the Amazon datasets to further understand the different classification and embedding methods. Many observations were made based on the data. The analysis demonstrated that different words can have different

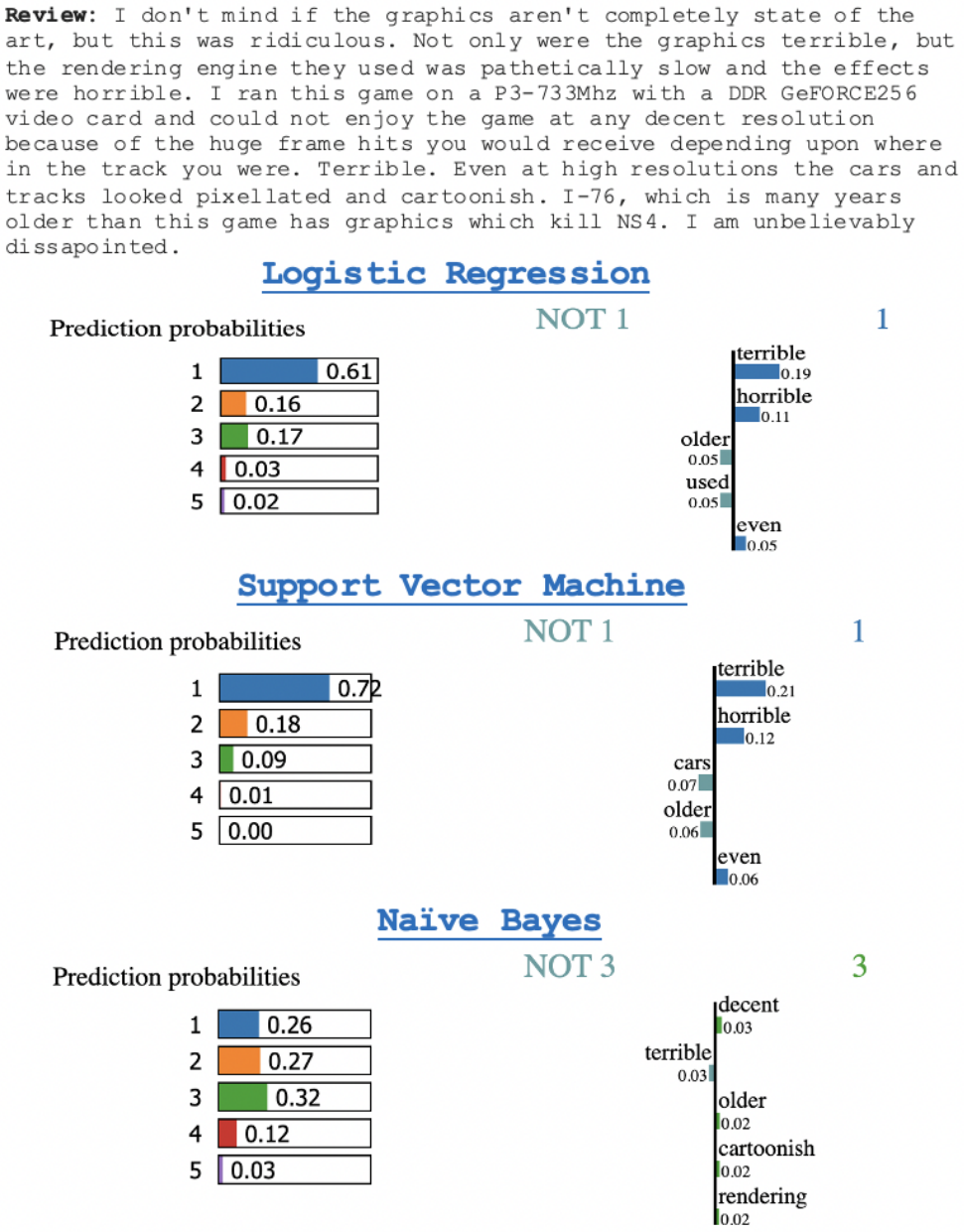


Figure 4.3: Different classification models and corresponding words can be tested via LIME (True label is 1)

contributions to ratings in different classification models with similar quantitative performance. In addition, different classification models can be compared regardless of if their predictions are correct or not. It is important to clearly understand why and how ML models make certain predictions, especially if they are being used in application-based systems and/or applied research.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study demonstrated the use of recent ML and NLP methods to study the relationship between product reviews and ratings. Bidirectional Encoder Representations from Transformers (BERT) based sentiment analysis was employed for reviewing the tone of product reviews. Based on the results, BERT outperformed the traditional sentiment analysis method VADER. The relationship between product reviews and ratings was evaluated via ML classification models. Different classification models and embedding methods were assessed. In addition, it was found that using BERT sentiment scores as an additional input mostly yields better accuracy. Lastly, local interpretable model-agnostic explanations (LIME) were adapted to explain the impact of certain words in the review to their corresponding rating. Specific observations were recorded to gain a better understanding of the models.

The results help us gain a better understanding of the relationship between reviews and ratings. Though, it is also important to understand that reviews and ratings are human-interpretable. Certain ratings may be good for one customer, but not good enough for others. There are many factors that can affect NLP models to not work as expected. Such factors may include

misspellings in the reviews which cause certain words to not be correctly embedded, and reviews being too vague and/or unclear. For instance, *"very good"* may be worth a 3-star rating for one but could be a 5-star rating for another customer. It is difficult to clearly distinguish what rating a review deserves. However, the vast amount of data available makes it possible to learn about such relationships and work towards building an algorithm that may automatically recommend a rating based on the written reviews.

5.2 Future Work

The rapid development in machine learning and natural language processing fields has significantly improved the quality of research in multiple domains. Techniques such as sentiment analysis and local explanations have allowed researchers to precisely learn about the back end of machine learning models, allowing them to make better decisions. For instance, BERT is not only used in sentiment analysis. It can be fine-tuned for multiple purposes depending on the aim of the task.

While the relationship between product reviews and ratings was explored, a good future direction is to expand and evaluate and compare the impact of reviews versus the impact of ratings. though customers may tend to look at both reviews and ratings, one may be more helpful in making decisions than the other. Another good future direction would be the implementation of algorithm-based recommendation systems which provide recommendations to users about what ratings and/or reviews they may want to give a product based on given information. As mentioned previously, this is a very tedious task as such factors are human-interpretable. As product reviews are very dependent on the users' feedback, a study can also be done to rank

reviews and ratings based on how helpful or genuine they are. For instance, reviews that are not descriptive or contextual may not rank highly, whereas reviews which provide detailed feedback about a product may be more useful to users and weigh more.

Bibliography

[1] Thomsen, J. F. How Sentiment Analysis Can Add Value To Numeric Ratings. (2017). URL: https://www.linkedin.com/pulse/how-sentiment-analysis-can-add-value-numeric-ratings-review-thomsen/. Last accessed: 7/20/2023.

[2] Thomsen, J. F. Online Review Statistics: The Definitive List (2023 Data). (2023). URL: https://www.https://www.luisazhou.com/blog/online-review-statistics/. June 2023.

[3] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. RecSys, 2013.

[4] Olsson F. NLP: Gaining insights from text reviews. (February 2020). Towards Data Science. URL: https://towardsdatascience.com/nlp-gaining-insights-from-text-reviews-94ef955c58c0. Last accessed: 7/25/2023.

[5] Müller, S., Toborek, V., Beckh, K., Bauckhage, M.J., Welke, P. (2023). An Empirical Evaluation of the Rashomon Effect in Explainable Machine Learning. ArXiv, abs/2306.15786.

[6] Hamman, F., Noorani, E., Mishra, S., Magazzeni, D., Dutta, S. (2023). Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees. arXiv preprint arXiv:2305.11997.

[7] Marx, C., Calmon, F., Ustun, B. (2020, November). Predictive multiplicity in classification. In International Conference on Machine Learning (pp. 6765-6774). PMLR.

[8] Dinov, I.D. (2018). Black Box Machine-Learning Methods: Neural Networks and Support Vector Machines. In: Data Science and Predictive Analytics. Springer, Cham. https://doi.org/10.1007/978-3-319-72347-1_11.

[9] Gokce, E. Sentiment Analysis on Amazon Reviews. (May 2020) Towards Data Science. URL: https://towardsdatascience.com/sentiment-analysis-on-amazon-reviews-45cd169447ac. Last accessed: 7/25/2023.

[10] McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. In Proceedings of the 7th ACM conference on Recommender systems (RecSys '13). Association for Computing Machinery, New York, NY, USA, 165–172. https://doi.org/10.1145/2507157.2507163.

[11] Fang, X. and Zhan, J. Sentiment analysis using product review data. (2015). Journal of Big Data 2, 5. https://doi.org/10.1186/s40537-015-0015-2.

[12] Salinca, A. Business Reviews Classification Using Sentiment Analysis. (2015). 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, pp. 247-250, doi: 10.1109/SYNASC.2015.46.

[13] Hu, Z., Hu, J., Ding, W., and Zheng, X. Review Sentiment Analysis Based on Deep Learning. (2015). IEEE 12th International Conference on e-Business Engineering, Beijing, China, pp. 87-94, doi: 10.1109/ICEBE.2015.24.

[14] Gan, Q., Ferns, B. H., Yu, Y., and Jin, L. A text mining and multidimensional sentiment analysis of online restaurant reviews. (2017). Journal of Quality Assurance in Hospitality & Tourism 18.4: 465-492.

[15] Thet, T. T., Na, J.-C., and Khoo, C. S. Aspect-based sentiment analysis of movie reviews on discussion boards. (2010). Journal of information science 36.6: 823-848.

[16] Baid, P., Gupta, A., and Chaplot, N. Sentiment analysis of movie reviews using machine learning techniques. (2017). International Journal of Computer Applications 179.7: 45-49.

[17] Zhou, Y. and Yang, S. Roles of Review Numerical and Textual Characteristics on Review Helpfulness Across Three Different Types of Reviews. (2019). In IEEE Access, vol. 7, pp. 27769-27780, https://doi: 10.1109/ACCESS.2019.2901472.

[18] Tsao, H.-Y., Chen, M.-Y., Campbell, C., and Sands, S. (2020). Estimating numerical scale ratings from text-based service reviews. Journal of Service Management, Vol. 31 No. 2, pp. 187-202. https://doi.org/10.1108/JOSM-06-2019-0167.

[19] Devi, L., Bai, V., Ramasubbareddy, S., and Govinda, K. Sentiment analysis on movie reviews. (2020). Emerging Research in Data Engineering Systems and Computer Communications: Proceedings of CCODE 2019. Springer Singapore.

[20] Zhang, Y., Sun, J., Meng, L., and Liu, Y. (2020). Sentiment Analysis of E-commerce Text Reviews Based on Sentiment Dictionary. IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, pp. 1346-1350, doi: 10.1109/ICAICA50127.2020.9182441.

[21] Murthy, G. S. N., Allu, S. R., Andhavarapu, B., Bagadi, M., and Belusonti, M. (2020). Text based sentiment analysis using LSTM. Int. J. Eng. Res. Tech. Res, 9(05).

[22] Singh, J., Singh, G., and Singh, R. 2016. A review of sentiment analysis techniques for opinionated web text. CSI transactions on ICT 4: 241-247.

[23] Abirami, A. M., and Gayathri, V. A survey on sentiment analysis methods and approach. (2017). Eighth International Conference on Advanced Computing (ICoAC). IEEE.

[24] Pancy, N. and Verma, R. A review on sentiment analysis and emotion detection from text. (2021). Social Network Analysis and Mining 11.1: 81.

[25] Nanli, Z., Ping, Z., Weiguo, L. I. , and Meng, C. (2012). Sentiment analysis: A literature review. In International Symposium on Management of Technology (ISMOT) (pp. 572-576). IEEE.

[26] Jain, P. K., Pamula, R., and Srivastava. G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. Computer Science Review, Volume 41, 100413, https://doi.org/10.1016/j.cosrev.2021.100413.

[27] Umer, M., Ashraf, I., Mehmood, A., Ullah, S., and Choi, G. S. (2020). Predicting numeric ratings for Google apps using text features and ensemble learning. ETRI Journal, Volume 43, Issue 1, Pages 95-108.

[28] Lee, G. T., Kim, C. O., and Song, M. (2021). Semisupervised sentiment analysis method for online text reviews. Journal of Information Science, 47(3), 387–403. https://doi.org/10.1177/0165551520910032.

[29] El-Din Mohamed Hussein, D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences 30.4: 330-338

[30] Alantari, H. J., Currim, I. S., Deng, Y., and Singh, S. (2022). An empirical comparison of machine learning methods for text-based sentiment analysis of online consumer reviews. 2022. International Journal of Research in Marketing, Volume 39, Issue 1, Pages 1-19.

[31] Almansou, A., Alotaibi, R., and Alharbi, H. (2022). Text-rating review discrepancy (TRRD): an integrative review and implications for research. Futur Bus J 8, 3. https://doi.org/10.1186/s43093-022-00114-y.

[32] Mutinda, J., Mwangi, W. and Okeyo, G. (2023). Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network. Appl. Sci., 13, 1445. https://doi.org/10.3390/app13031445.

[33] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778.

[34] Bramhall, S., Horn, H., Tieu, M., and Lohia, N. (2020). Qlime-a quadratic local interpretable model-agnostic explanation approach. SMU Data Science Review, 3(1), 4.

[35] Zafar, M. R., and Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. Machine Learning and Knowledge Extraction, 3(3), 525-541.

[36] Zhao,X., Huang, W., Huang, X., Robu, V., and Flynn, D. (2021, December). Baylime: Bayesian local interpretable model-agnostic explanations. In Uncertainty in artificial intelligence (pp. 887-896). PMLR.

[37] Zhong, Y., Bhattacharya, A., and Bukkapatnam, S. (2023). EBLIME: Enhanced Bayesian Local Interpretable Model-agnostic Explanations. arXiv preprint arXiv:2305.00213.

[38] Zafar, M. R., and Khan, N. (2019). DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv preprint arXiv:1906.10263.

[39] Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., and Costa da Silva, E. (2019). Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors, 19(13), 2969.

[40] Kumarakulasinghe, N. B. , Blomberg, T., Liu, J. , Leao, A. S., and Papapetrou, P. (2020, July). Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 7-12). IEEE.

[41] Graziani, M., Palatnik de Sousa, I., Vellasco, M. M., Costa da Silva, E., Müller, H. and Andrearczyk, V. (2021). Sharpening local interpretable model-agnostic explanations for histopathology: improved understandability and reliability. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, Proceedings, Part III 24 (pp. 540-549). Springer International Publishing.

[42] Schlegel, U., Vo, D. L., Keim, D. A., and Seebacher, D. (2021, September). Ts-mule: Local interpretable model-agnostic explanations for time series forecast models. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 5-14). Cham: Springer International Publishing.

[43] Farhood, H., Saberi, M., Najafi, M. (2021, October). Improving object recognition in crime scenes via local interpretable model-agnostic explanations. In 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW) (pp. 90-94). IEEE.

[44] Wang, B., Pei, W., Xue, B., and Zhang, M. (2022). Explaining Deep Convolutional Neural Networks for Image Classification by Evolving Local Interpretable Model-agnostic Explanations. arXiv preprint arXiv:2211.15143.

[45] Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., Trajanov, D. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. IEEE Access, 8, 131662-131682.

[46] Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics.

[47] Altrabsheh, N., Gaber, M. M., Cocea, M. (2013). SA-E: sentiment analysis for education. In International Conference on Intelligent Decision Technologies (Vol. 255, pp. 353-362).

[48] Hutto, C., Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

[49] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing and Management, vol. 39, no. 1, pp. 45–65, 2003.

[50] Mikolov, Tomas Chen, Kai Corrado, G.s Dean, Jeffrey. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR. 2013.

[51] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. Machine Learning, 20, 273-297. http://dx.doi.org/10.1007/BF00994018

[52] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[53] Bayes, Mr; Price, Mr (1763). "An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S". Philosophical Transactions of the Royal Society of London. 53: 370–418. doi:10.1098/rstl.1763.0053.

[54] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.