

PRODUCT SENTIMENT ANALYSIS FOR AMAZON REVIEWS

Arwa S. M. AlQahtani

Department of Computer Science, Princess Nourah bint Abdulrahman University,
Riyadh, Saudi Arabia

ABSTRACT

Recently, Ecommerce has Witnessed Rapid Development. As A Result, Online Purchasing has grown, and that has led to Growth in Online Customer Reviews of Products. The Implied Opinions in Customer Reviews Have a Massive Influence on Customer's Decision Purchasing, Since the Customer's Opinion About the Product is Influenced by Other Consumers' Recommendations or Complaints. This Research Provides an Analysis of the Amazon Reviews Dataset and Studies Sentiment Classification with Different Machine Learning Approaches. First, the Reviews were Transformed into Vector Representation using different Techniques, I.E., Bag-Of-Words, Tf-Idf, and Glove. Then, we Trained Various Machine Learning Algorithms, I.E., Logistic Regression, Random Forest, Naïve Bayes, Bidirectional Long-Short Term Memory, and Bert. After That, We Evaluated the Models using Accuracy, F1-Score, Precision, Recall, and Cross-Entropy Loss Function. Then, We Analyzed The Best Performance Model in Order to Investigate Its Sentiment Classification. The Experiment was Conducted on Multiclass Classifications, Then we Selected the Best Performing Model And Re-Trained It on the Binary Classification.

KEYWORDS

Amazon, Data Analytics, Analysis, Product Sentiment, Ecommerce

1. INTRODUCTION

Nowadays, the world is becoming digitalized. eCommerce is taking ascendancy in this digitalized world through the availability of products within reach of customers. Furthermore, the eCommerce website allows the people to convey what they think and feel. In fact, people are increasingly relying on the experiences of other customers. Our opinion and purchasing decision-making are affected by the experience of others and their feedback about products. We always ask others about their opinion to get the benefit from their experience; hence, the importance of reviews has grown. However, it is almost impossible for customers to read all such reviews; therefore, sentiment analysis represents an essential role in analyzing them. This research proposes a sentiment analysis to predict the polarity of Amazon mobile phone dataset reviews using supervised machine learning algorithms. Sentiment analysis helps a customer to make their purchasing decision based on the experience of others. Further, it will help companies to improve their products by knowing customers' opinions and needs [1].

1.1. Problem Statement

Customer reviews or ratings aim to define the attitude of the writer towards the product. It may be positive, negative, or neutral. Some people give a product four or five stars and express their final satisfaction with it, and others give a product one or two stars and express their final dissatisfaction with it. This does not present any difficulty in sentiment analysis. However, other people give three

stars, although obviously expressing their final satisfaction with it. This leads to confusing other customers, as well as companies, who want to know their actual opinion. Consequently, customers and companies face difficulty with respect to analyzing reviews and understanding consumer satisfaction. So, the three-star rating doesn't actually represent a neutral sentiment, because in practice people who assign a 3-star rating to a product or service don't necessarily mean that they're absolutely balanced in their opinion between positive and negative. Based on this argument, this research proposes a sentiment analysis to predict the polarity of Amazon mobile phone dataset reviews. We will leave the 3-star rating as is and consider it to represent a neutral sentiment. This is done with the purpose of increasing the challenge and difficulty of this study and to measure the efficiency of state-of-the-art NLP models, like BERT in solving difficult classification problems. Furthermore, four machine-learning models with different feature extraction approaches will be used in this research: Logistic Regression, Naïve Bayes, Random Forest, and Bi-LSTM. Then, we analyse the best performance model in order to investigate its sentiment classification. At the end of the study, we will take the best performing model and re-train it on the dataset with the neutral class removed, effectively recasting the problem as a binary-classification problem. We'd like to measure how much this recasting of the problem will affect model performance.

This paper is structured as follows: Section 2 provides information about related research in the fields of sentiment analysis on data collected from various sources. Section 3 provides the related works. Section 4 and Section 5 explain both methodology and Data collection & Analysis respectively. Section 6 discusses Results and Discussion.

2. BACKGROUND

Natural Language Processing (NLP) is a subfield of Artificial Intelligence and linguistics which was introduced in the 1950s. It is devoted to exploring the understanding and manipulation of natural language with the help of computers. Sentiment analysis is one of the applications of NLP, which involves computer-based study of people's opinions, attitudes, and emotions toward different entities. Recently, it has become an active research area in NLP due to the rapidly growing volume of reviews. Since online blogging or social networking sites are preferred by most of the people to express their views on specific products, services, or organizations, it has become challenging for individuals or organizations to efficiently process the wealth of information included in the corpus of available reviews. Thus, sentiment analysis techniques have grown and can automatically extract and summarize opinions embedded in a huge collection of product reviews [2] [3]. Researchers have classified opinion mining into the following three levels:

- Document-level: At this level, the entire document has to be analysed at the same time. It uses both supervised and unsupervised classification algorithms to classify a particular product review with an overall positive or negative opinion [2].
- Sentence-level: At this level, the document has to be divided into sentences in order to apply subjectivity classification. Further, the opinion of an opinionated sentence is classified as positive, negative or neutral [2].
- Aspect-level: At this level, the analysis focuses on feature terms of entities to provide detailed opinions or sentiments about aspects of those entities [2]

3. RELATED WORKS

Opinion mining at the document level plays a pertinent role in identifying customer interests and preferences regarding specific products. Recently, research interest turned strongly towards

opinion mining. Some of the ongoing studies related to this field are discussed in this chapter concerning their datasets and pre-processing, methodology, and evaluation metrics.

Many studies have been made on data collected from various sources, such as tweets on Twitter, product reviews, customer comments, etc. The authors in [4] concentrate on mining reviews of the Amazon website for three products, i.e., Apple iPhone 5S, Samsung J7, and Redmi Note 3, while other studies also used Amazon reviews for various purposes. In [5], they retrieved 21,500 Amazon reviews using Amazon API in English, then randomly selected 3,000 reviews for the experiment. In [1] and [6], studies were conducted over a dataset of the mobile phone category from Amazon comprising over 400,000 consumer reviews. The work in [7] used a set of 300 reviews of electronic devices from Amazon. In [8], the researchers conducted the study over Amazon reviews in particular product categories including GPS, books and cameras with about 2000 reviews (1,000 positives and 1,000 negatives) in each dataset. In [9], from Amazon, over 100,000 Chinese reviews on clothing products were extracted. In [10], the researcher examined over 1,000 reviews from Amazon.

Text pre-processing is considered as one of the most crucial steps in the NLP, for improving the quality of the textual data, whereby the data-analysis is conducted through various steps and numerous methods. Some of the popular steps are: stop-word removal, tokenization, stemming, lemmatization and POS (Parts of speech) tagging. Stop words do not add meaning to the text, do not contribute to the analysis and, hence, are deleted during the pre-processing step. Tokenization is the process of separating a sentence into meaningful tokens (phrases, symbols or words) by eliminating punctuation marks. Stemming process involves changing a word into its root form, while lemmatization is grouping the forms of a word into a single canonical form. POS tagging is performed to identify different parts of speech in the text, which is quite crucial for natural language processing.

Stop words were eliminated from all product reviews in [4]. [11] Some more steps were included depending on the specific task in addition to the previous steps. In [5], the authors pre-processed the dataset by removing the strings of letters which were repeated for effect. These words were also replaced by two occurrences, i.e. “coool” became “cool”. In [6], stemming, lowercasing, punctuation removal and white space elimination were performed. In [1], tokenisation, lowercasing, spelling check, and lemmatization were conducted. Also, In [8], tokenization and stemming were carried out in the pre-processing step. The dataset in [11]. hence, different pre-processing steps were included, where each half-width Kana character was translated to full-width Kana character and all full-width digits and alphabetic characters were translated to half-width characters. Moreover, the punctuation was removed. Whereas, in paper [7], after the extraction of product reviews from the database, parts of speech was determined. After that, phrases were split into vectors of sentences and then the sentences were split into vectors of words, with the meaning of each word extracted from SentiWordNet. In [9], they adopted the ICTCLAS 4 system for segmentation of the collected Chinese comment texts into words and tagging them with proper POS tags. In addition, they removed stop words and punctuation. In [12] tokenization and spelling check were conducted.

Sentiment classification methods can be classified into machine learning techniques and a lexicon-based approach. In [4], the Naïve Bayes classifier (NB) outperformed Logistic Regression (LR) and SentiWordNet to classify the review as a positive or negative. The performance of the classification methods was evaluated by using Recall, Precision, and F-measure. In [11], NB further performed at two levels of granularity, i.e., sentence level and review level. TF-IDF was used to calculate the relative frequency of each word. The result of the experiment at the review level was measured by Accuracy, Mean Absolute Error (MAE) and Root Mean Square Error

(RMSE). Study [5], applied three machine learning models: Support Vector Machine (SVM), NB, and Maximum Entropy (ME), to classify the reviews positive, negative, and neutral. They used unigrams and weighted unigrams to train machine learning classifiers. The experimental result was evaluated by accuracy: SVM has resulted in the maximum accuracy of 81%.

In [6], the authors used continuous bag of words (CBOW) and skip-gram methods with four different classification algorithms: NB, SVM, LR, and Random Forest (RF) to classify the consumer reviews. The experimental results show that RF, using CBOW, achieves the most superior accuracy 91%. The authors in [7] studied the performance of different machine learning algorithms, LR, Stochastic Gradient Descent(SGD), NB and Convolutional Neural Networks (CNN), using a range of feature extraction techniques, such as bag-of-words, TF-IDF, Glove, and word2vec. The experimental results showed that CNN, with word2vec as a feature extraction technique, provided the best results, with accuracy of 91%. Furthermore, they applied the Lime technique to give analytical reasons for the reviews being classified as either positive, negative, or neutral.

It is noted that Artificial Neural Networks (ANN) have seldom been considered in comparative studies in the sentiment analysis literature. [8] aims to present an empirical comparison between SVM and ANN regarding document-level sentiment analysis. [7] Presents a semantic approach for a sentiment analysis application, which is based on using the SentiWordNet lexical resource. Further study used a semantic approach; the authors of [9] proposed sentiment classification based on word2vec and SVMperf. The study consists of two parts. First, they used word2vec to cluster similar features to capture the semantic features in the selected domain. Second, they used SVMperf to classify the comment texts. The authors in [10]studied the classification of sentiment using SVM, RF, and a hybrid approach, Random Forest Support Vector Machine (RFSVM). This proposed method outperforms individual algorithms.

4. METHODOLOGY

This section provides an overview of the proposed methodology of sentiment analysis for Amazon mobile phone reviews. Figure.1 depicts the phases of the current work starting with the data collection until evaluating each classification model.

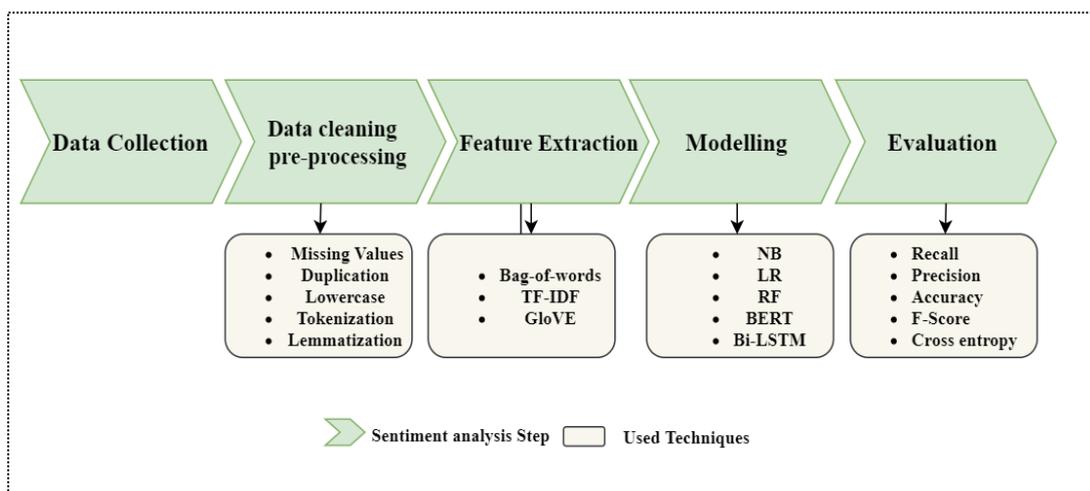


Figure 1. Overall methodology of sentiment analysis for Amazon mobile phone reviews

4.1. Data Pre-Processing

Text pre-processing is an important step in NLP to develop the textual data quality; Figure 2 depicts all pre-processing steps that were applied to the Amazon mobile phone dataset for this research. The reviews were pre-processed by altering all the letters to lowercase, not mixed capitals and lowercase; for example, "Good," and "GrEat" are converted to "good" and "great". Also, all punctuation and stop words that frequently appears and does not significantly affect meaning, including "-", /, :, ? , the, a" were eliminated. Further, the reviews were tokenized which is the process of separating a sentence into a sequence of words called "tokens." Usually, each token is identifiable or separated from another token by a space character; consequently, the tokenizing process relies on the space character to perform word separations [13] [14]. Then, all tokens were returned to their base or dictionary form by applying lemmatization process.

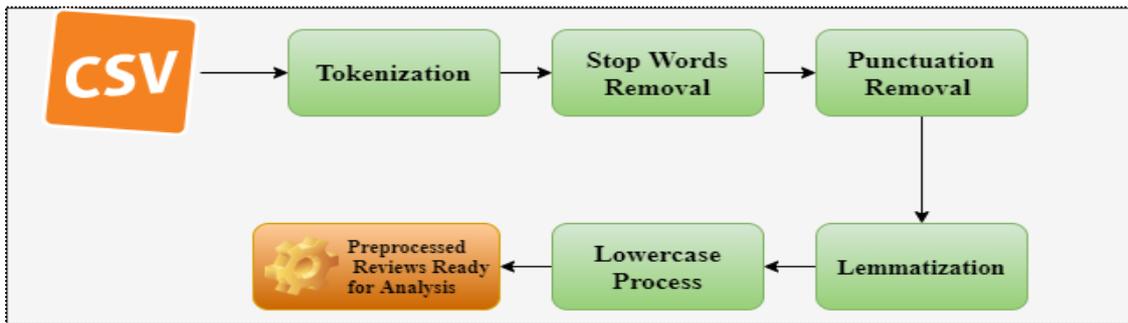


Figure 2. Summary of Pre-processing Steps

Each review in the dataset was labeled to positive, negative, or neutral based on its star rating in the same way performed by [1]. Then, the dataset was split into 60% training, 20% validation, and 20% testing for baseline models. On the other hand, the dataset was split into 90% training, 5% validation, and 5% testing for deep learning. Figure 3 shows an illustration of a review after pre-processing has been carried out.

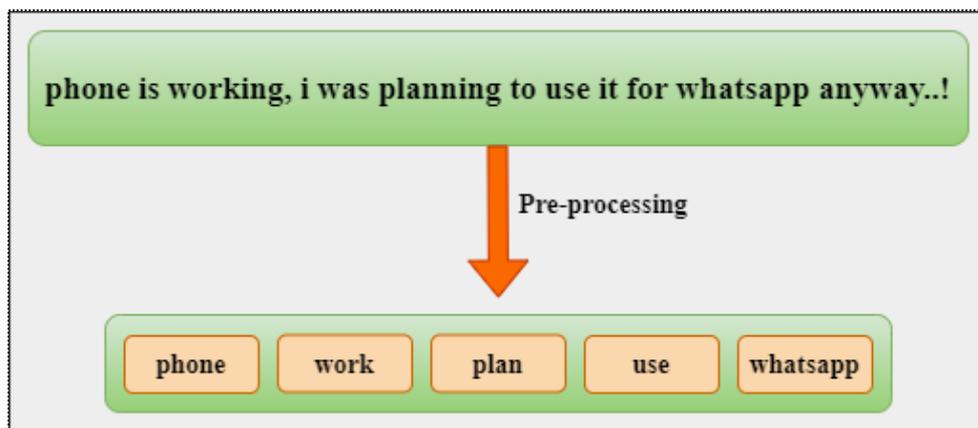


Figure 3. Review after pre-processing steps

4.2. Feature Extraction

Natural Language Processing (NLP) demands computers to interpret human language. Initially, the textual data is converted to a numerical form suitable to fit into the machine learning models. Bag-

of-Words, Term frequency - Inverse document frequency, GloVe are being used in this project. They are discussed in the following sub-sections.

4.3. Classification Models

Classification is a technique that groups data into various categories [21]. It is applied in the field of Sentiment Analysis in order to classify data into binary classification (e.g., “positive” and “negative”) and ternary classification (e.g., “positive,” “negative” and “neutral”) and based on that the sentiment analysis process is completed [22]. Two approaches are mostly used in sentiment classification of customers’ reviews: lexicon-based and machine learning [23], shown in figure 4

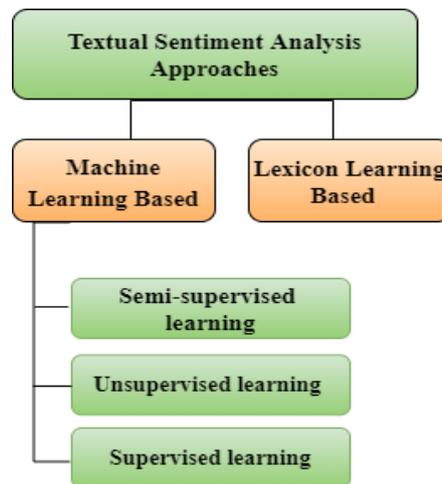


Figure 4. Sentiment analysis approaches

Lexicon-based approaches predict the polarity of the textual reviews based on words that are annotated by polarity or polarity scores [24]. On the other hand, machine learning techniques are divided into: supervised learning, and unsupervised learning. Our project includes supervised machine learning, that is popularly used to create sentiment classification models in the field of sentiment analysis. First, these models build a training set and label the training data by sentiments. Then, a collection of features are taken from the training data and forwarded to a classifier model, such as Naive Bayes (NB), Logistic Regression (LR), Random Forest (RF), and so on. After the training phase with the sentiment labels, the classifier can be used to predict the sentiment orientation of a sample on new data.

4.4. Performance Evaluation Parameters

The performance of the classification methods can be found by using Accuracy, F-Score, Cross-entropy, Recall, and Precision. These parameters are helpful to evaluate the performance of supervised machine learning algorithms, based on the element from a matrix known as the confusion matrix or contingency table [32]. A confusion matrix is typically used for allowing visualization of the performance of an algorithm. From the classification viewpoint, terms such as ‘True Positive (TP)’, ‘False Positive (FP)’, ‘True Negative (TN)’, ‘False Negative (FP)’ are used to compare labels of classes in this matrix, as shown in Table 1. True Positive represents positive reviews that were classified as positive by the classifier, whereas False Positive is predicted as negative but is actually classified as positive. Conversely, True Negative represents negative reviews that were classified as negative by the classifier, whereas False Negative is predicted as

positive actually classified as negative. According to the data of the confusion matrix, precision, recall, f-measure, and accuracy are used for evaluating the performance of classifiers.

Table 1. Contingency Table

		Predicted class	
		Positive	Negative
Actual class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

- **Precision**

This is defined as the ratio of the number of reviews correctly classified as positive to the total number of reviews that are truly positively classified.

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

- **Recall**

This is defined as the ratio of the number of reviews correctly classified as positive to the total number of reviews that are classified positively.

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

- **Accuracy**

This is the ratio of the reviews that are correctly classified to the total number of reviews.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

- **F-score**

This is a combined measure for precision and recall.

$$F - score = 2 \frac{(Precision*Recall)}{(Precision+Recall)} \tag{10}$$

- **Cross-entropy**

Cross-entropy or log loss is used further to measure the performance of the classification models. The output of log loss is a probability value between 0 and 1.

5. DATA COLLECTION & ANALYSIS

Recently, the availability of customer reviews has increased for a wide range of products and services. Customer reviews typically have two components: star ratings and review text [33]. In this research, we will use customer reviews and ratings in order to classify the review as Positive, Negative, or Neutral.

5.1. Data Collection and Description

An Amazon dataset extracted via Prompt Cloud is considered for sentiment analysis. The dataset concerns unlocked mobile phones and it was acquired in December 2016. It is a publicly available dataset from Kaggle.com. The Amazon reviews dataset consists of beyond 400,000 consumer reviews in the category of mobile phone. Particularly, it covers 413,840 reviews and 6 features, classified as follows: i) Information about mobile phone (Brand Name, Product Name, Price, Rating). ii) Information about reviews (Reviews and Review Votes). Undoubtedly, data cleansing the process in which we go through all of the data within a data frame and either remove or update any incomplete, incorrect or duplicated information is considered as data cleaning. This step is valuable to maximize data accuracy, since data quality is of central importance in order to obtain the desired result with high efficiency and accuracy. As shown in Table 2, the Amazon mobile phone dataset comprises missing values in some features; some steps have been taken to clean the data.

Table 2. Amazon mobile phone data description

Features	Description	Data type	Missing value
Brand Name	<i>Name of the manufacturing company, e.g., Apple, LG</i>	Object	65,171
Product Name	The name given by a company to each model of mobile phone.	Object	0
Price	The cost of the mobile phone	Float	5,933
Rating	Star rating [1-5]	Integer	0
Reviews	Customer opinion toward the mobile phone.	Object	62
Review Votes	The number of customers who voted in the reviews.	Float	12,296

First, records with a null value (62) in “Reviews” were dropped. Second, all null values in “Review Votes” (12,296) were substituted with Zero. Since they are genuinely reviews receiving no votes, i.e., they received zero votes, zero is the most appropriate value. Third, all null values in “Price” were substituted with 144.71; this is the median value of the price (12,296 null values). Additionally, duplication in the data frame was handled by deleting all duplicated records (64,079 records). Concerning the “Brand Name”, brand names of mobile accessories were deleted. Moreover, we removed some brand names that contained additional quotations. Many brand names have spelling mistakes and not standardized names; for these, we normalized the brand name by substituting them with the correct names and unifying them. To illustrate this step, “Lenovo Manufacturer” was replaced by “Lenovo” and “Samsung” was replaced by “Samsung” can be considered as an instance. After this step, the total number of brand name changed from 382 to 282.

5.2. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a way of visualizing and interpreting information that is hidden in rows and column formats. Furthermore, it helps us to maximize insight into the dataset by utilizing a number of charts. We examined the rating distribution by star rating and number of

reviews of Amazon mobile phone. We can see from figure that the five-star count is the most frequent, and two-star is the least. On the other hand, figure 5 shows that 52.4% of consumers gave 5 stars, 15.1% gave 4 stars, 7.96% gave 3 stars, 6.21% gave 2 stars, and 18.4% gave 1 star.

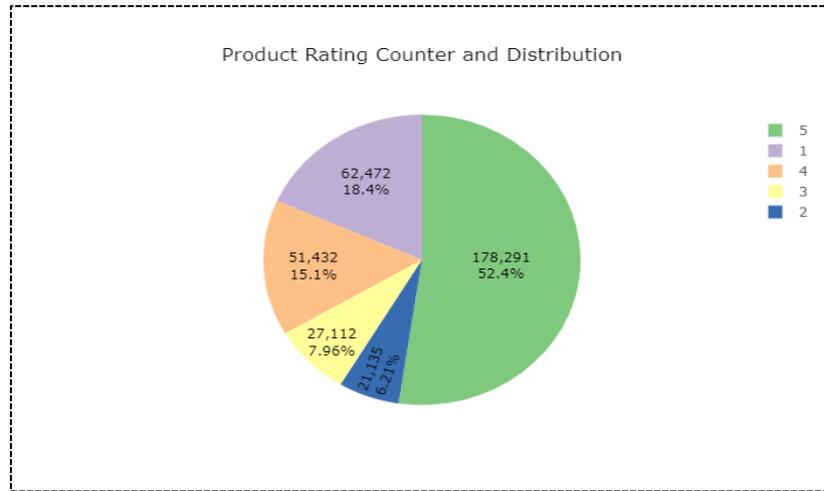


Figure 5. Rating counter over the reviews

Then, the polarity of the Amazon mobile phone dataset is studied and represented by a pie chart. Review sentiment is a new column that includes three values. Reviews that are rated 4 or 5 will be considered as “Positive” while 1 or 2 are considered “Negative” and 3 as “Neutral.” Figure 6 indicates 67.5% of reviews are “Positive,” 24.6% are “Negative,” and 7.97% are “Neutral.” Further, we studied the number of reviews and brand name. It indicates that Samsung obtained the maximum number of reviews (58,395), while BLU and Apple received 51,780 and 51,077 reviews respectively.

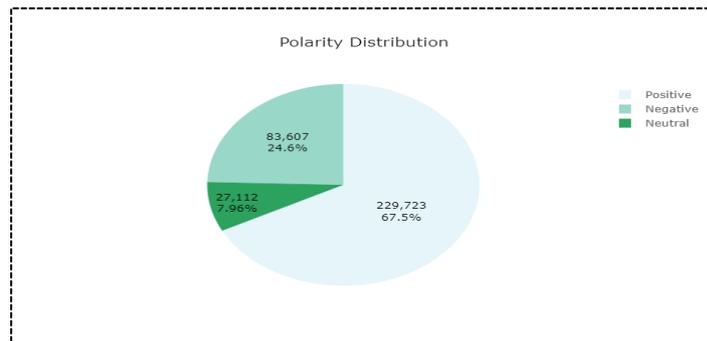


Figure 6. Polarity distribution of Amazon mobile phone dataset.

6. RESULTS AND DISCUSSION.

Python and its notebook, Jupyter, was used in this project alongside other supporting libraries to accomplish data cleansing, visualization, pre-processing, and machine learning modelling. Additionally, because of the limited resources of my personal laptop, Google Colab was used for faster implementation. The current section presents the results of the proposed models: LR, RF, NB, BERT, and Bi-RNN. To evaluate the models studied here, different evaluation metrics were used including Cross Entropy Loss, Accuracy, Precision, Recall, F-score: all of them were

described in section 4.4. First, we will present the result for each model with the validation dataset and different feature extraction techniques: BOW with N-gram, TF-IDF, and Glove. Then, we will represent the final result for each classification model with the test dataset. The experiments were first conducted for a multiclass classification applying all the proposed models. Second, they were conducted for a binary classification with the model that had the highest accuracy in validation.

6.1. Experiment 1: Multiclass classification

In the current study, Amazon mobile phone reviews were classified into positive, negative, and neutral according star rating, since one- and two-star ratings are considered as negative, while four and five stars are considered as positive, and three-star ratings considered as neutral. Accordingly, all proposed models were applied with all feature extraction methods.

The LR model applied with BOW, TF-IDF, and GloVe Table 3 shows the results of the LR with each of these approaches. The LR with BOW (Bigram) achieved high accuracy, 85.5% and 0.59 cross entropy loss. Further, the recall of the LR with BOW (Bigram) are superior to other feature extraction methods.

Table 3. The Result of Logistic Regression Model

	Cross Loss	Accuracy	Precision	Recall	F1-score
LR –BOW	0.717	0.813	0.838	0.813	0.824
LR – BOW- Bi	0.594	0.855	0.865	0.855	0.859
LR – BOW-TRI	0.637	0.832	0.835	0.832	0.833
LR – TF-IDF	0.495	0.831	0.877	0.831	0.849
LR – Glove	0.641	0.778	0.836	0.778	0.799

Furthermore, Table 4 presents the evaluation results of the RF model with different feature extraction approaches. The RF with GloVe exhibits 90% higher accuracy and 0.390 cross-entropy loss as compared to other approaches. The apparent reason behind getting this result is that GloVe is a word embedding method that aims at building a low dimensional vector, which in turns makes the sparsity in GloVe go down in comparison with other feature extraction approaches.

Table 4. The Result of Random Forest Model

	Cross Loss	Accuracy	Precision	Recall	F1-score
RF –BOW	0.638	0.842	0.865	0.842	0.849
RF – BOW- Bi	0.637	0.842	0.866	0.842	0.849
RF – BO-TRI	0.637	0.840	0.864	0.840	0.847
RF – TF-IDF	0.581	0.861	0.875	0.861	0.865
RF – Glove	0.390	0.900	0.902	0.900	0.898

Table 5 presents the evaluation results of the NB classifier with all proposed feature extraction approaches in this project except GloVe. GloVe is a word vector representation that puts similar words together, while the NB theory assumes that features are independent. The NB classifier with BOW (Trigram) exhibits higher accuracy, but also higher cross-entropy loss as compared to the other approaches, with 78% and 1.18, respectively.

Table 5. The Result of Naïve Bayes Classifier

	Cross Loss	Accuracy	Precision	Recall	F1-score
NB – BOW	3.038	0.703	0.742	0.703	0.714
NB – BOW- Bi	2.329	0.759	0.782	0.759	0.764
NB – BOW-TRI	1.176	0.784	0.786	0.784	0.780
NB – TF-IDF	2.986	0.707	0.745	0.707	0.718
NB – Glove	-	-	-	-	-

Table 6 presents the result of BERT. It achieved accuracy of 94% and cross entropy loss of 0.189, which means the performance of the model is quite good. Further, the model has a high precision and recall.

Table 6. The Result of BERT

	Cross Loss	Accuracy	Precision	Recall	F1-score
BERT	0.189	0.947	0.946	0.947	0.946

Table 7 presents the result of Bi-LSTM with two different embeddings: fine-tuned GloVe embedding and jointly trained embedding. In both approaches, the model achieved accuracy of 93% and cross entropy loss of 0.189. It is good result this is because the structure of Bi-RNN enables the networks to trained in both time directions simultaneously (backward and forward). This structure helps to incorporate information from both sides of a sequence, further enhancing the performance of the model.

Table 7. The Result of Bi-LSTM

	Cross Loss	Accuracy	Precision	Recall	F1 score
Bi-LSTM (Jointly-trained embedding)	0.234	0.933	0.929	0.933	0.929
Bi-LSTM (fine-tuned GloVe embedding)	0.240	0.929	0.925	0.929	0.925

6.2. Final Evaluation

Table 8 indicates the final results for all classification models. BERT achieved the highest accuracy of 94.7%. Further, Bi-directional short-term memory resulted 93% accuracy. On the other hand, the baseline model of RF with GloVe outperformed the LR and NB with an accuracy of 90%. Figure 7 shows that the BERT model has the highest accuracy compared to other models.

Table 8. Final Result for all Classification Models

	Cross Loss	Accuracy	Precision	Recall	F1-score
LR– BOW-Bi	0.596	0.853	0.863	0.853	0.857
RF-GloVe	0.403	0.899	0.901	0.899	0.897
NB-BOW-TRI	1.159	0.784	0.785	0.784	0.779
BERT	0.189	0.947	0.964	0.947	0.946
Bi-LSTM (Jointly-trained embedding)	0.234	0.933	0.929	0.933	0.929

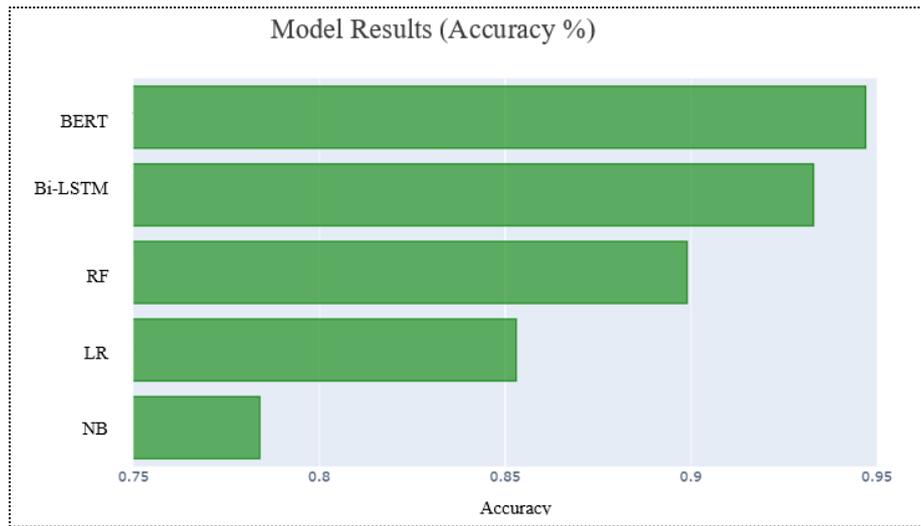


Figure. 7. Final results for multiclass classification

The confusion matrix was further used to describe the performance of classification models. Figure 10 exhibits that the BERT model outperformed other classification models. On the other hand, it is observed from the figure that the BERT model's success in classifying the neutral class achieved only 65%; this was expected, since a review with three-star rating does not mean the customer was absolutely balanced in their opinion between positive and negative. On the other hand, it achieved 97% in classifying positive reviews and 94% in classifying negative reviews.

6.3. Experiment 2: Binary classification

After applying all proposed models with a variety of feature extraction in multiclass classification, the results of our experiment are shown in Table 8. In this experiment, Amazon mobile phone reviews were classified into positive, negative, based on the star rating, since one- and two-star ratings are considered as negative, while four and five stars are considered as positive. Accordingly, the binary classification will apply with the same feature extraction approaches that were achieved with the final results of multiclass classification.

Table 9 represent the results of binary classification for all models on test dataset. BERT achieved an excellent result with the heights accuracy of 98%. Further, Bi-LSTM gave results of 97% accuracy. On the other hand, the baseline model of RF with GloVe outperformed the LR and NB with an accuracy of 90%. Figure 8 shows that the BERT model in binary classification has the highest accuracy compared to other models. After recasting the problem as a binary-classification problem, we observed that the performance of all models has improved. Thus, the neutral class has an effect on the models performance.

Table 9. The Final Results of Binary Classification on Validation Dataset

	Cross Loss	Accuracy	Precision	Recall	F1-score
LR- BOW-Bi	0.270	0.931	0.931	0.931	0.931
RF-GloVe	0.209	0.942	0.943	0.942	0.942
NB-BOW-TRI	0.452	0.880	0.878	0.880	0.874
BERT	0.064	0.983	0.986	0.986	0.986
Bi-LSTM (Jointly-trained embedding)	0.088	0.977	0.977	0.977	0.977

Table 10. The final results of binary classification on test dataset

	Cross Loss	Accuracy	Precision	Recall	F1-score
LR- BOW-Bi	0.275	0.927	0.928	0.927	0.928
RF-GloVe	0.212	0.940	0.940	0.940	0.940
NB-BOW-TRI	0.469	0.878	0.876	0.878	0.872
BERT	0.071	0.984	0.984	0.984	0.984
Bi-LSTM (Jointly-trained embedding)	0.092	0.974	0.974	0.974	0.974

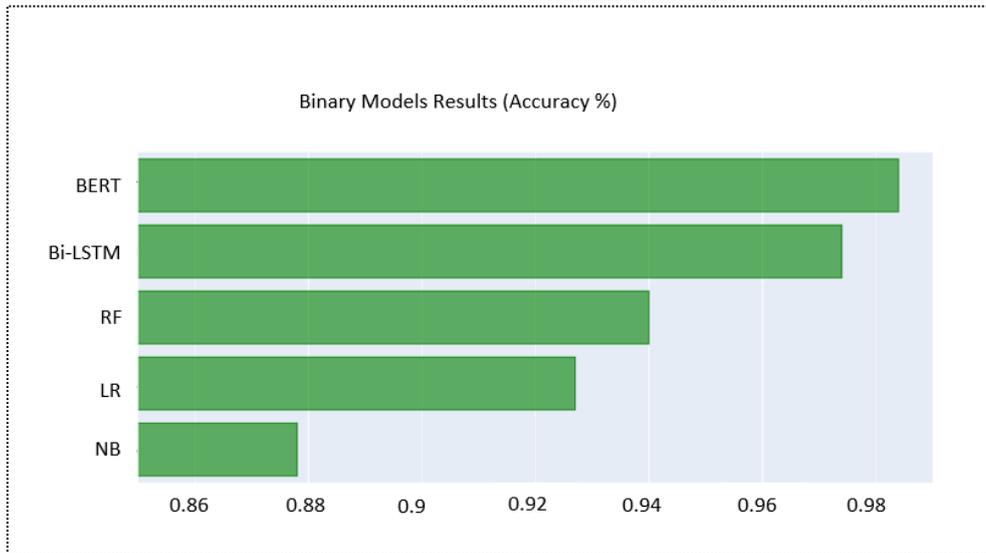


Figure 8. Final Results for Binary Classification

6.4. Analyze the Performance of BERT model

The BERT model has the highest accuracy of 94%. Thus, we want to analyse its performance in order to investigate its sentiment classification. Currently, there has been a focus on model interpretability and inspecting misclassifications. We need to know the reasons that cause models to give the wrong classification in some reviews while identifying correctly in others. Therefore, we looked for the misclassified reviews with the highest loss values. We interpreted them to explain the BERT model’s decision boundary in the same way as our understanding. By looking at figure 9, we observed that the BERT model classified the review as positive with a probability of 0.99. While the actual label of the review was neutral. This is an example of when the BERT model is correct, but the label is wrong. The model recognizes and highlights the positive word 'love' with green colour and its variant intensity relative to other positive words. The customer gave the product a three-star rating, although obviously expressing his final satisfaction with it. That s why we said in the beginning that the three-star ratings do not always represent a neutral sentiment.



Figure 9. An example of customer misclassifying the review.

Further, by looking at figure 10 it is observed that the customer expresses his opinion with dissatisfaction toward the mobile phone. In fact, the model classifies the review correctly as negative while the actual class was wrongly given as positive.



Figure 10. An example of wrong actual label of the review

6.5. Discussion

This section discusses the results of our models and compares them with some of the popular models that are used for the same task, i.e., sentiment analysis on Amazon mobile phone reviews. Some models that have been applied in this project, i.e., BERT and Bi-LSTM, have not previously been used with the Amazon mobile reviews dataset; thus, the comparison between existing studies will not cover all sides. In study [6], the authors applied LR, NB, SVM, and RF in a binary classification task. They obtained the best accuracy of 90% with RF and word2vec as feature extraction, while, in our experiment, RF with GloVe obtained 94%. Further, the authors in [7] applied four models: Logistic Regression, Naïve Bayes, Stochastic Gradient Descent, and Convolutional Neural Network (CNN) in multiclass classification. The CNN model with word2vec for feature extraction showed the best accuracy of 92% while our result with BERT model achieved 94% on multiclass classification.

6.6. Conclusion

Sentiment analysis is a necessarily and commonly used approach to extracting knowledge from text data in eCommerce websites. E-commerce portals are generating a massive amount of text data daily in the form of suggestions, feedback, tweets, and comments. Besides, the opinion of the people is implied by reviews, ratings, and emoticons. Extracting information about a product from the review will help a customer to exploring more about the product and help them in decision-making. In this study, multiclass and binary classification for Amazon mobile phone using supervised machine learning algorithms: Logistic Regression, Naïve Bayes, Random Forest along with different feature extraction approaches is studied Further, this project applied Bidirectional Long-Short Memory (Bi-LSTM) with GloVe embedding and joint-learned embedding. Moreover, Bidirectional Encoder Representations from Transformers (BERT) model was also applied. BERT model has achieved an excellent result in multiclass classification and binary classification, with accuracy of 94% and 98%, respectively. On the other hand, Bi-LSTM with joint-learned embedding also provides a very good result, with accuracy of 93% for multiclass classification and 97% for binary classification. Random Forest with word embedding (GloVe) outperforms other baseline models, LR and NB, with accuracy of 90% for multiclass classification and 94% for binary classification.

6.7. Limitations and Suggestions for Future Work

This project shows the execution of a variety of machine learning models LR, NB, RF, and Bi-LSTM, with different feature extraction approaches for a text classification task. Further, we used the pre-trained BERT model and finetuned it for the sentiment analysis task on the Amazon mobile

phone reviews dataset. For future work, we are planning to use word2vec for feature extraction with our models and to detect fake reviews. Besides, varied classifiers could be used other than those already mentioned, e.g., Gated Recurrent Unit (GRU), and Support Vector Machine. Moreover, the dataset is in the domain of Amazon mobile reviews; it can be extended to the analysis of Amazon reviews in general. The limitation of this study is use of Google Colab to make the implementation faster due to the limited resources of my personal laptop.

REFERENCES

- [1] S. A. a. A. N. S. Aljuhani, "A Comparison of Sentiment Analysis Methods on Amazon Reviews of Mobile Phones," *International Journal of Advanced Computer Science and Applications*, vol. 10, 2019.
- [2] L. a. L. B. Zhang, "Aspect and entity extraction for opinion mining," in Zhang, Lei and Liu, Bing, Berlin, Heidelberg, Springer, 2014, pp. 1--40.
- [3] Y.-C. a. K. C.-H. a. C. C.-H. Chang, "Social media analytics: Extracting and visualizing Hilton hotel ratings and reviews from TripAdvisor," *International Journal of Information Management*, vol. 48, pp. 263--279, 2019.
- [4] K. S. a. D. J. a. M. J. Kumar, "Opinion mining and sentiment analysis on online customer review," in *IEEE International Conference on Computational Intelligence and Computing Research (ICIC)*, Chennai, 2016.
- [5] A. S. a. A. A. a. D. P. Rathor, "Comparative study of machine learning approaches for Amazon reviews," *Procedia computer science*, vol. 132, pp. 1552--1561, 2018.
- [6] B. a. S. S. Bansal, "Sentiment classification of online consumer reviews using word vector representations," *Procedia computer science*, vol. 132, pp. 1147--1153, 2018.
- [7] A. a. S. V. a. M. B. ernian, "Sentiment analysis from product reviews using SentiWordNet as lexical resource," in *2015 7th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Bucharest, 2015.
- [8] J. F. V. W. P. G. N. Rodrigo Moraes, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Systems with Applications*, vol. 40, pp. 621--633, 2013.
- [9] D. a. X. H. a. S. Z. a. X. Y. Zhang, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Systems with Applications*, vol. 42, pp. 857--1863, 2015.
- [10] Y. a. L. M. a. E. K. K. E. Al Amrani, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Computer Science*, pp. 511-520, 2018.
- [11] Y. a. K. V. Saito, "Classifying User Reviews at Sentence and Review Levels Utilizing Naïve Bayes," in *21st International Conference on Advanced Communication Technology (ICACT)*, PyeongChang Kwangwoon_Do, Korea (South), 2019.
- [12] .. X. C. T. S. M. W. N. J. Sobia Wassan, "Amazon Product Sentiment Analysis using Machine," *Revista Argentina de Clínica Psicológica*, pp. 695-703, 2021.
- [13] Bahrawi, "Sentiment Analysis Using Random Forest Algorithm-Online Social Media Based.," *JOURNAL OF INFORMATION TECHNOLOGY AND ITS UTILIZATION*, vol. 2, pp. 29-33, 2019.
- [14] M. a. S. R. Fikri, "A Comparative Study of Sentiment Analysis using SVM and SentiWordNet," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, pp. 902-909, 2019.
- [15] N. Tamara and Milievi, "Comparing sentiment analysis and document representation methods of Amazon reviews," *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 000283--000286, 2018.
- [16] K. a. M. W. a. C. W. Ogada, "N-gram Based Text Categorization Method for Improved Data Mining," *Journal of Information Engineering and Applications*, vol. 5, pp. 35--43, 2015.
- [17] R. a. P. B. a. S. S. Al-Rfou, "Polyglot: Distributed word representations for multilingual nlp," *arXiv preprint arXiv:1307.1662*, 2013.
- [18] Y. a. A. G. a. J. P. a. K. T. Sharma, "Vector representation of words for sentiment analysis using GloVe," in *2017 international conference on intelligent communication and computational techniques (icct)*, Jaipur, 2017.
- [19] R. S. C. D. M. Jeffrey Pennington, "GloVe: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014.
- [20] A. H. C. H. H. B. G. Marwa Naili, "Comparative study of word embedding methods in topic segmentation," *Procedia Computer Science*, vol. 112, pp. 340-349, 2017.

- [21] H. a. K. A. Sinha, "A Detailed Survey and Comparative Study of sentiment analysis algorithms," in 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS), Mathura, India, 2016.
- [22] M. a. O. T. Bouazizi, "A pattern-based approach for multi-class sentiment analysis in Twitter," in 2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 2016.
- [23] V. M. N. Harpreet Kaur, "A survey of sentiment analysis techniques," in 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2017.
- [24] Z. a. F. Y. a. J. B. a. L. T. a. L. W. Li, "{A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools and Applications*, vol. 78, pp. 6939--6967, 2019.
- [25] A. K. A. a. A. B. A. Hassan, "Reviews Sentiment analysis for collaborative recommender system," *Kurdistan journal of applied research*, vol. 2, pp. 87--91, 2017.
- [26] V. M. a. V. J. a. B. P. Pradhan, "A survey on Sentiment Analysis Algorithms for opinion mining," *International Journal of Computer Applications*, vol. 133, pp. 7--11, 2016.
- [27] H. a. B. S. a. S. G. Parmar, "Sentiment mining of movie reviews using Random Forest with Tuned Hyperparameters," in *International Conference on Information Science.*, 2014, Kerala.
- [28] J. a. C. M.-W. a. L. K. a. T. K. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [29] Y. a. L. M. a. L. L. a. F. Z. a. W. F.-X. a. W. J. Yu, "Automatic ICD code assignment of Chinese clinical notes based on multilayer attention BiRNN," *Journal of biomedical informatics*, vol. 91, pp. 103-114, 2019.
- [30] Y. a. S. X. a. H. C. a. Z. J. Yu, "A review of recurrent neural networks: LSTM cells and network architectures," *Neural computation*, vol. 31, pp. 1235--1270, 2019.
- [31] C. a. S. C. a. L. Z. a. L. F. Zhou, "A C-LSTM Neural Network for Text Classification," *arXiv preprint arXiv:1511.08630*, 2015.
- [32] A. a. A. A. a. R. S. K. Tripathy, "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117--126, 2016.
- [33] S. M. Mudambi, D. Schuff and Z. Zhang, "Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings," in 2014 47th Hawaii International Conference on System Sciences, Waikoloa, 2014.