Consumer review Analysis using NLP and Data Mining

Md. Nasimuzzaman , Ahmed Nur Merag , Sumya Afroj , MD. Mustakin Alam , Md Humaion Kabir Mehedi , and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{*md.nasimuzzaman, ahmed.nur.merag, sumya.afroj, md.mustakin.alam, humaion.kabir.mehedi* }*@g.bracu.ac.bd*
annajiat@gmail.com

***Abstract*—The popularity and development of the Internet have greatly increased the amount of information that is easily available. For a sizable portion of the population, it has evolved into the primary forum for opinion expression. These viewpoints can be discussed in reviews, online forums, and social media websites like Twitter and Facebook. Sentiment analysis and information extraction are made possible by the ease of access to all of these viewpoints, which is free. Sentiment analysis is concerned with automatically ascertaining, through the use of computer tools, if a reviewer's purpose is positive, negative, or neutral with regard to specific goods, services, and problems. Since they have such a positive impact on every particular organization, these strategies are rapidly gaining favor. As a result, we have researched multiple methods for classification, data preprocessing, and feature extraction in this study and compared the outcomes using the logistic regression, multinomial naive bayes, and other accuracy measures. The results of the simulation demonstrate that, when we have applied to a dataset with features that the model has retrieved, SVM is the most accurate classifier for the given data set.**

***Index Terms*—Consumer review, Data mining, NLP, TF-IDF, SVM, Naive Bayes, logistic regression, multinomial naive bayes, Matplotlib, Seaborn, Itertools, Google Translate.**

I. Introduction

Over the past ten years, internet usage has dramatically expanded, and with it, so has the volume of text data generated. Analysis of this is more important than ever because there are more articles, reviews, and online discussions about a variety of issues. Determining the outlook of consumers toward a certain service is one of text data analysis's most crucial applications. Sentiment analysis is the categorization of opinions about specific products, services, or topics in a text (word, sentence, or a document) to computationally determine the polarity of the author or point of view toward the topic, which can be positive, negative, or neutral [1]. Machine learning (MN) and Data mining knowledge are used to perform sentiment analysis on text data [2]. Word frequency, part-of-speech tagging, and phrase frequency are just a few of the features from sentiment analysis that may be extracted using machine learning [3]. An identified training-supervised machine learning algorithm receives a dataset and is then capable of classifying and learning the polarity (positive, negative, or neutral) using the learned model, analyzing an unknown text [4]. To build a sentiment analysis model that is more accurate, data preparation and feature extraction are necessary. As feature extraction accounts for the majority of a model's accuracy, it is essential. The noise reduction, normalization, tokenization, and vectorization methods are used in data preprocessing [5]. Approaches for feature extraction including SVM, naive bayes and word frequency-inverse document frequency have also been investigated in an effort to improve accuracy. The accuracy of the model can be greatly enhanced by combining these strategies [4] [3]. The suggested project looks at the effects of various preprocessing techniques. On a database of Amazon reviews for mobile phones and feature extraction techniques to determine how they affect the model's precision. Natural language processing, text analysis, text classification techniques, and automatic classification systems have all been used to attain the desired results. This is how the research is organized: The earlier research utilizing opinion mining is discussed in section ii. Architecture modeling and data preprocessing are done in Section iii. Section iv includes the performance analysis, accuracy of implemented models and results.

II. Related Works

A succinct synopsis of previous work on opinion mining is provided in this section. Research and development have been quite active recently pertaining to sentiment analysis and opinion mining. For applications involving sentiment analysis, we carefully analyzed the value of text pre-processing [6].

By using a chopped method, extraneous characteristics were removed. The accuracy of the classifier is significantly increased by suitable text pre-processing, it has been found through experimental findings confirming sentiment analysis with appropriate feature representation and selection. Instead of identifying the text's sentiment before evaluating it, sentiment analysis identifies the text's sentiment first, opinion

979-8-3503-3286-5/23/$31.00 ©2023 IEEE

mining is a method for obtaining and analyzing someone's feelings on something. They examined 54 recent articles and found that there is still an opportunity for advancement in the techniques for sentiment categorization and feature selection [7].

To categorize product evaluations using data from Twitter, a number of machine learning techniques that incorporate semantic analysis were used. The Naive Bayes method was found to perform better when paired with an unigram model than when used alone. Furthermore, accuracy increased following the use of WordNet's semantic analysis tool, which was subsequently followed by the earlier method [8] [9].

The issue of a computer being able to predict and comprehend a person's sentiment or contextual opinion on something is known as sentiment analysis. According to his narration, the data and the language employed in it determine the modifications required to enhance the classifier's performance. When transformations are performed, as well as when the least important data is filtered away, the machine learning approach performs more successfully and generalizes better [10].

Using data from Twitter, they conducted a poll on sentiment analysis. Using existing accessible approaches, such as machine learning techniques, unstructured, heterogeneous opinions that are occasionally good, occasionally negative, or neutral. The author got to the conclusion that accuracy was impacted by data cleanliness. Furthermore, the author asserted that using the Bigram model produced better results than using other techniques like Support Vector Machine and Naive Bayes [11].

Another idea that has been presented in detail is sentiment analysis on Twitter data, It is accomplished step-by-step utilizing machine learning techniques [12].

In order to analyze Twitter data, this research recommended using a scalable, rapid, and flexible text framework called Apache Spark. Out of all the algorithms used, a decision tree's accuracy, precision, recall, and F1-Score were all 100 percent [2].

The authors have talked about the difficulties he encountered while undertaking sentiment analysis in relation to the methods and strategies employed [13]. They also employed a machine learning strategy to classify tweets as favorable or negative using Twitter sentiment analysis about electrical equipment like laptops, mobile phones, etc. Several different methods, including Naive Bayes, maximum entropy, ensemble classifiers, and support vector machines, were employed to evaluate the revised feature vector's classification accuracy. They found that the modified feature vector is effective for electrical goods and that all of the classifiers' accuracy was essentially the same [14].

III. Architecture and Modeling

The suggested work offers numerous stages for assessing the opinion using a dataset of product reviews. The fig.1 displays the mind map of how we approached our research work. The subsequent paragraphs outline the numerous steps.

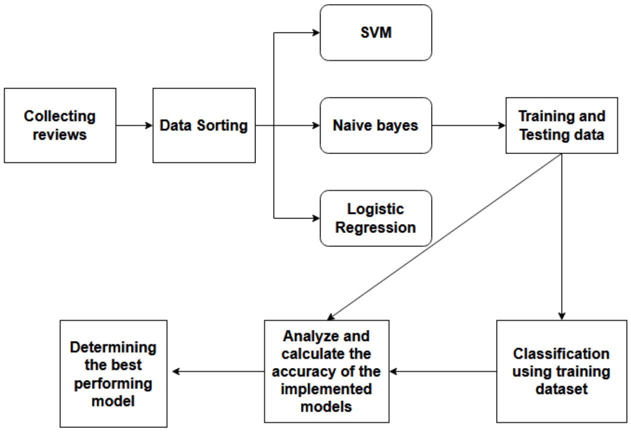
Collecting reviews
Data Sorting
SVM
Naive bayes
Logistic Regression
Training and Testing data
Classification using training dataset
Analyze and calculate the accuracy of the implemented models
Determining the best performing model

Fig. 1: Architectural Schematics

Regarding the approach, firstly we collect the review dataset from kaggle. After that we used the data in models such as svm classifier, naive bayes. Then we train and test the data. Subsequently, the classification is done. Finally we analyzed the results of several implemented models and determine the best performing model and the accuracy.

A. Data Pre-Processing

Our dataset which is reviews left on Amazon by cell phone users collected from kaggle has been trained using SVM and Naive bayes. The dataset must first be categorized before it is trained and tested. To import the dataset, we use Python 3.6's Pandas library. Preparing the text for use by computers in activities like analysis, prediction, etc. are a process known as text preprocessing. In this article, we will focus on removing reviews with short word counts, translating all reviews into English, changing all reviews to lowercase letters, removing stop words, and special characters, and implementing the stemming of words using a variety of libraries. To eliminate the stop-words from the review, we employed the NLTK corpus. For example, stop-words such as 'not' are removed during this step. The dataset has been trained by the SVM classifier to have three feedback sections which are positives, negatives, and neutrals where the feedback has been labeled as follows: negatives as 0, neutrals as 1, and positives as 2. Then we dropped the unrelevant columns to make the dataset more compact. Making root words from the original words is the last stage in pre-processing. A word is referred to as a root word if it does not have a prefix or suffix. Python 3.6, Pandas, google translate, natural language toolkit (NLTK), regular expression (RE), logistic regression, and multinomial naive bayes, matplotlib, spacy, keras, cartropy, seaborn, and itertools libraries and packages are used to train the dataset to determine if a review is favorable or unfavorable.

Firstly, we plotted a figure 2 of the most used words in the reviews left by the consumers in the dataset which are denoted as positive reviews. The figure size for positive, neutral, and negative words used in the reviews is (8,6).

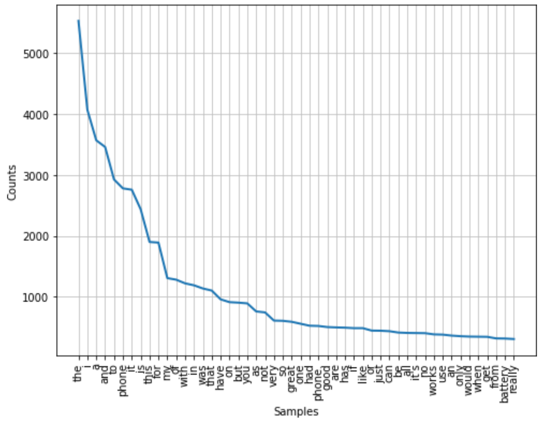

Fig. 2: Most used words in positive reviews

Similarly, we have also generated a plot 3 where the reviews given by the consumers are neither positive nor negative. This plot represents the neutral words left in the reviews.

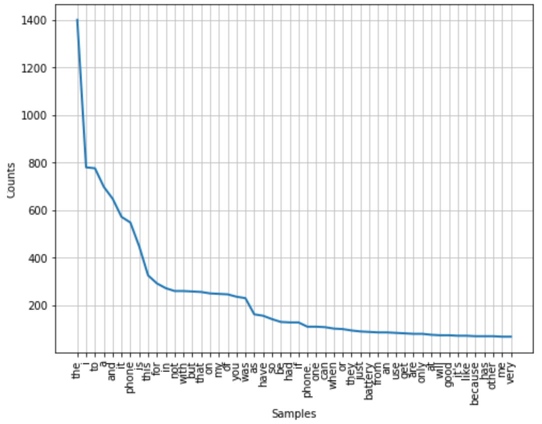

Fig. 3: Most used words in neutral reviews

Likewise, this plot 5 represents the negative comments containing the most used words left by the consumers in the dataset.

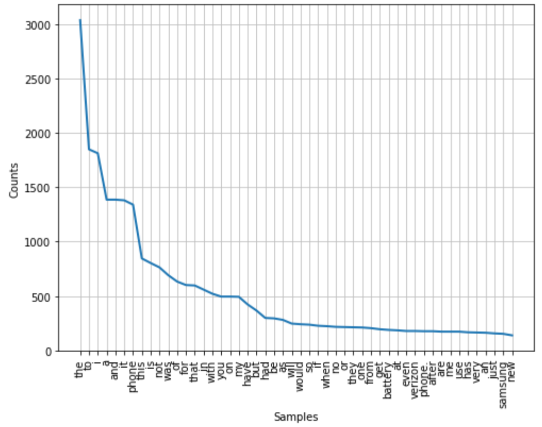

Fig. 4: Most used words in negative reviews

B. Term Frequency-Inverse Document Frequency (TF-IDF)

Combining the idf with the tf results in the Term Frequency-Inverse Document Frequency. Idf is a method for determining a word's information density and how common or uncommon it is across all texts. The result is the logarithm. The ultimate result, represented by the equation, is the TF-IDF 1.

$$TFIDF(t, d) = TF(t, d) * IDF(t, d) \tag{1}$$

C. Support Vector Machine (SVM)

An example of a machine used as a teaching method for classification, regression, and outlier detection is called a Support Vector Machine (SVM). By locating the hyperplane that optimizes the margin between the classes, SVMs determine the optimum boundary between two classes in a dataset. The line known as the hyperplane separates the two classes in a two-dimensional space, and a hyperplane also separates the two classes in a higher-dimensional space. The margin is the separation between each class's nearest data points and the hyperplane.

D. Naive Bayes

A popular machine learning technique in nlp for tasks that classify texts like sentiment evaluation, spam detection, and subject categorization is the Naive Bayes classifier. First, we instruct the model with a collection of labeled training data, which consists of text documents and their associated categories, to be able to apply the Naive Bayes classifier for nlp. The model determines the likelihood that each word will appear in each category during training. The model then determines the likelihood that a fresh text document will fall into each category when classifying a document, we do it based on how frequently certain words appear in it. The document is then placed in the category with the highest likelihood.

E. Theme Creation

Bi-gram frequencies allow us to identify motifs that provide context for the product. This can be accomplished manually by choosing the themes that will contribute the most insight to the investigation. Using bi-gram models involves extracting two-word combinations from text data in order to identify topics of discussion. A bi-gram model can be used to identify the most frequently occurring combinations of words in a review, which can then be used to create themes or topics that are discussed in the review. This can help to provide more insight into the sentiment of a review and the topics that the reviewer has discussed.

F. Isolation of Training and Test Data

Pre-processing, which is the most difficult step, is followed by the training and testing process. A limited quantity of data is trained to recognize the pattern in the data.

G. Classification

We first decompose the dataset into its components and then we provide training data. Several classification methods have been used, including the NB classifier, the svm classifier, etc. With the help of a number of questions and criteria, A conditional probability model acts as the foundation for the Naive Bayes technique. If the vector form of this data set is to be classified as x=(X1,..., Xn) with n independent features, Equation 2 provides the probability. Feature independence is a presupposition of the Naive Bayes classifier.

$$P\left(X_1, ...X_n\right) = P\left(C_k\right) \prod_{i}^{n} = P\left(X_i|C_k\right) \tag{2}$$

The k'th class name is represented here by the letter . In order to classify the input data, the Support Vector Machine classifier creates a hyperplane connecting the x points in the equation's set 3.

$$\vec{w}.\vec{x} - b = 0 \tag{3}$$

Here, is the hyper-normal plane's vector. The performance of each classification method varies depending on the characteristics of the data set, and each has advantages and disadvantages. Several algorithms are used to evaluate each Natural Language Processing technique's accuracy performance.

IV. Results and Discussion

The concepts and models which are implemented here are text preprocessing, SVM classifier, logistic regression, multinomial Naive Bayes along with python 3.6, pandas, google translate, natural language toolkit, regular expressions, matplotlib, seaborn and itertools. A unique approach has been taken in our research where we incorporate all these methods to analyze consumer reviews where we analyse the accuracy of models which we implemented. Based on the comparison of the accuracies of the implemented models, we determine if the review has a positive, neutral, or negative attitude. According to our findings, The outcomes of our research are more satisfactory than the previous works because all the available methods and concepts of natural language processing and data mining are used efficiently.

TABLE I: Classification Report of SVM

SL num	precision	recall	f1-score	support
0	00.93%	00.90%	00.92%	200
1	00.89%	00.63%	00.74%	81
2	00.92%	00.98%	00.95%	469
accuracy			00.92%	750
macro average	00.92%	00.84%	00.87%	750
weighted average	00.92%	00.92%	00.92%	750

The calculated accuracy of the SVM classifier model is 92.93% which is the highest accuracy I, calculated so far. This accuracy shows the superiority of the svm classifier model comparing to other models implemented here.

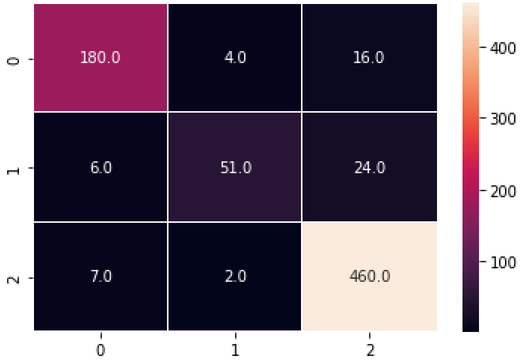

Fig. 5: Confusion matrix of SVM

The different outcomes of the prediction and findings are arranged in a table in a confusion matrix 5, which helps in visualizing the classification task's results. All of the predicted and actual values from a classifier are shown in a table. In order to help you comprehend, we've created a confusion matrix with the heat map in this case.

TABLE II: Classification Report of Logistic regression

SL num	precision	recall	f1-score	support
0	00.85%	00.41%	00.55%	110
1	00.00%	00.00%	00.00%	38
2	00.70%	00.99%	00.82%	227
accuracy			00.72%	375
macro average	00.51%	00.47%	00.46%	375
weighted average	00.67%	00.72%	00.66%	375

The above table highlights the accuracies we calculated using logistic regression. From this table II, we can get an overview of the performance of logistic regression which has an accuracy of 71.73%.

TABLE III: Classification Report of Multinomial Naive Bayes

SL num	precision	recall	f1-score	support
0	01.00%	00.01%	00.02%	110
1	00.00%	00.00%	00.00%	38
2	00.61%	01.00%	00.76%	227
accuracy			00.61%	375
macro average	00.54%	00.34%	00.26%	375
weighted average	00.66%	00.61%	00.46%	375

The accuracy of multinomial naive bayes is shown below. From this table III we calculated the accuracy of multinomial NB which is 60.80%.

V. Conclusion

The performance of several algorithms and NLP techniques on a dataset of Amazon cell phone reviews is compared in this study, and the model with the best performance and accuracy is the support vector machine model. The SVM classifiers have the maximum accuracy in this case which is 92.93% as we've seen. However, there is always a cost associated with logistic regression and multinomial naive bayes when determining the model's accuracy. As a result, SVM achieves the greatest accuracy, it is therefore the best classification algorithm for the dataset in question. Different kernel functions and settings can be tested in order to further enhance SVM performance. In this research, we are looking for information from product reviews to determine what the main drawbacks and benefits are. To accomplish the same, we have used data mining and natural language processing concepts. Python 3.6 was used in this instance. Natural language processing (NLP) in consumer review analysis is a powerful technique for assessing customer sentiment and opinions about goods, services, and brands. However, it has several drawbacks, such as variation in language, contextual understanding, lack of ability to extract non-textual data, and limited generalizability. To ensure that the analysis is precise and thorough, it is important to apply a range of methods and procedures.

REFERENCES

[1] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," *Artificial intelligence review*, vol. 52, no. 3, pp. 1495–1545, 2019.

[2] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, pp. 628–632.

[3] Y. Yang, "Research and realization of internet public opinion analysis based on improved tf-idf algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*. IEEE, 2017, pp. 80–83.

[4] ——, "Research and realization of internet public opinion analysis based on improved tf-idf algorithm," in *2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES)*. IEEE, 2017, pp. 80–83.

[5] V. Singh and B. Saini, "An effective tokenization algorithm for information retrieval systems," 2014.

[6] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia computer science*, vol. 17, pp. 26–32, 2013.

[7] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey. ain shams eng. j. 5 (4), 1093–1113 (2014)."

[8] G. Gautam and D. Yadav, "Sentiment analysis of twitter data using machine learning approaches and semantic analysis," in *2014 Seventh international conference on contemporary computing (IC3)*. IEEE, 2014, pp. 437–442.

[9] R. Ravi, "Ravi k., ravi v," *A survey on opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[10] Ò. R. Llombart, "Using machine learning techniques for sentiment analysis," 2017.

[11] V. Kharde, P. Sonawane, *et al.*, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.

[12] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016, pp. 628–632.

[13] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, vol. 30, no. 4, pp. 330–338, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1018363916300071

[14] A. Addiga and S. Bagui, "Sentiment analysis on twitter data using term frequency-inverse document frequency," *Journal of Computer and Communications*, vol. 10, no. 8, pp. 117–128, 2022.