Application of Machine Learning for Sentiment Analysis of Movies Using IMDB Rating

Sandeep Rathor*[1]
Dept of Computer Engineering & Applications
GLA University, Mathura
sandeep.rahtor@gla.ac.in

Yuvraj Prakash[2]
Dept of Computer Engineering & Applications
GLA University, Mathura
yuvraj.prakash_mt.cs21@gla.ac.in

2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT) | 978-1-6654-8038-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSNT54456.2022.9787663

***Abstract*— In today's environment, the customer behavior related to the product is important. The rapid growth of social media has made it easier to collect information about our products. Billions of customers share information about their products on social media pages. Manually managing large amounts of social media information is difficult. However, we have proposed a framework for analyzing customer sentiment using data mining and machine learning algorithms. The proposed framework consists of several steps to processes the information and try to find the good accuracy of the classification. The proposed framework is used for sentiment analysis of movie using IMDB rating. In this paper, the data preprocessing and feature extraction are done using data mining techniques. The result analysis shows that the proposed framework can be useful for better decision-making.**

Keywords—Sentiment Analysis, Movie Rating, IMDB Rating, Machine Learning, Text Mining.

I. INTRODUCTION

The behavior of your consumers toward your goods determines the success of any firm. Every company strives to please its clients. As a result, it is continually striving to improve. If the consumer enjoys your product, it is a success. Thus, to understand consumer feedback, we must first examine customer behavior. As a result, sentiment analysis is the most important aspect of consumer behavior analysis. Sentiment analysis is a computational approach for recognizing and categorizing opinions based on text and assessing if sentiments toward a product is positive, negative, or neutral.

The purpose of sentiment analysis is to categories each type as a sentiment. Companies may use sentiment analysis to predict product acceptability and find the best strategy to enhance product quality [1]. This study assists businesses in making better decisions in order to improve in the future. Customers can provide feedback in a variety of methods, including social media, blogging, Face-to-face communication and others. There are several real-world applications where we may utilize this technique to analyze the behavior. Some of them are Product sentiment analysis, Brand sentiment analysis, Social-Media sentiment analysis, Customer sentiment analysis, Movie sentiment analysis, and Music sentiment analysis etc [2]. Due to the tremendous development of technology, the maximum amount of activity takes place on the Internet. Social media is one of the platforms where customers share their reactions to a product by posting articles, stories, comments, etc.[3]

The company can verify their product reviews using social media and can analyze customer behavior. Due to the large amount of data and the complexity of the information, it is virtually impossible for decision makers to read all the information pulled from popular social media platforms such as Twitter, Instagram, and Facebook, and from other web sources, like blogs and another online site [4]. Therefore, we need a model that can perform sentiment analysis on a large data set with a shorter computation time. We need to classify customer behavior into positive, negative and neutral using advanced machine learning algorithms. The company can use this recommendation model to understand analysis of their product reviews and can make changes accordingly.

In our proposed framework, we extracted data from Internet Movie Database (IMDB) to analysis the sentiments of the customers, for analyzing the data we used different state of art machine learning techniques, i.e. supervised learning or unsupervised leaning such as Naïve Bayes (NB), Support vector machine (SVM) and Logistical regression (LR) and produces the experimental outcome.

The rest of paper is organized as related work, proposed framework, results and conclusion.

II. RELATED WORK

This section contains a review of previous work done by various researchers in the field of sentiment analysis. So far, a lot of work has been devoted to in this context to inform users on social media to tap into the emotions of individuals towards any topic, product, trend, etc.

Khan et.al., [5], proposed a framework for knowledge acquisition, preprocessing, feature extraction, and using 3 supervised machine learning algorithms to classify customer emotion sentiment. Algorithms used are Support Vector Machines, Decision Trees and Naive Bayes. The proposed framework has also been tested to evaluate the system's performance. In this paper, SVM gives a stronger result than the alternative classification techniques. Accuracy achieved is 90.30%, however, the proposed method may fall into over-fitting situation.

Md Shoeb et al. [6], proposed a dataset testing technique that uses a rapid miner tool to generate a classifier and a text

978-1-6654-8038-3/22/$31.00 ©2022 IEEE
DOI: 10.1109/csnt.2022.36

miner to transform sentences. Three supervised classifications are used. These classifiers are Naive Bayes, Decision Trees, and KNNs. In this paper, decision trees give high results compared to KNN and Naive Bayes. The achieved accuracy is 95.96%, but the proposed approach may not have much impact on huge or complex datasets.

Das et.al., [7], proposed a set of techniques of machine learning with semantic analysis for classifying the sentence and product reviews based on Twitter data. The naïve bayes technique gives a better accuracy result with 88.2%, as compared to SVM and maximum entropy. When the semantic analysis WordNet is followed by the preceding technique, the accuracy increases to 89.9 % from 88.2 %. However, other machine learning techniques have not been used in this paper.

Kher et.al., [8], proposed a framework to analyze the sentiment of the music using the Random k-Label sets and Multi-label techniques which shows similar results, which was classified according to the six emotion classes that are sad-lonely, relaxing-clam, quiet-still, angry-aggressive, amazed-surprised, and happy, and unsupervised A priori association algorithm is used for showing different relations between music feature and emotions. However, accuracy is very limited.

Palak et.al., [9], proposed a method to analyze film reviews based on supervised machine learning techniques. This paper employs the algorithms Naive Bayes, K-Nearest Neighbour, and Random Forest. When compared to other algorithms, the Naive Bayes techniques achieved the highest accuracy of 81.45%. however, the data size used in this paper was very limited.

The use of CNN for sentimental short sentence classification is proposed [10][11]. The author of this research collects text from a variety of sources, including individuals and social networking sites, and performs text categorization on it. The text received from several sources may differ in opinion, be in a different language, or be illegible. As a result, the author proposed a model based on natural language processing and deep neural networks to address this issue. The findings of this study suggest that CNN can achieve the best text categorization results. This work, however, is just concerned with sentiment classification. [12][13]

III. PROPOSED METHODOLOGY

In this section, a framework for sentiment analysis is proposed. It can be utilized to recognize the sentiments in a movie.

As we know that social media data is not completely clear and not properly organized as well as have grammatical errors, however; we have to process the text to obtain customer sentiment in the organized manner. Therefore, to recognize the sentiment of the text, we need to pre-process the text.

All the steps are shown in the figure 1 as:

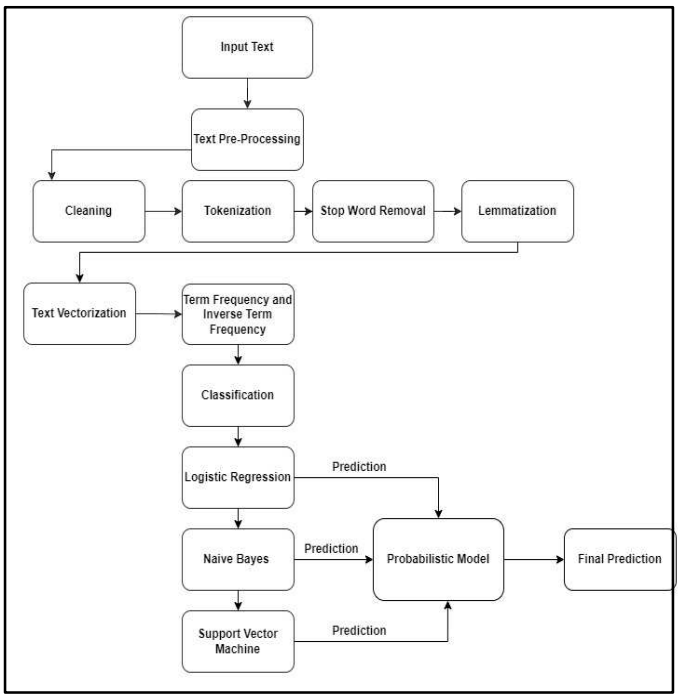

Figure1: Proposed framework for sentiment analysis

The detailed working of all the steps as shown in the figure 1 are:

A. Cleaning

The purpose of cleaning is to remove unwanted symbols, punctuation from the text data.

B. Tokenization

The purpose of tokenization is to divide social media text into meaningful words and convert it into tokens.

C. Stop Words Removal

Removing meaningless word which don't make any senses in the text like "the"," a", "an"," those"," she" etc.

D. Lemmatization

Lemmatization is used to obtain the word into its original form.

After text pre-processing we have to convert the text into numerical from by using text vectorization to check the positive, negative or neutral occurrence in the data. To validate our proposed framework, we calculated the accuracy of the classification as well as precision and recall as:

$$Accuracy = \frac{\sum \text{True Preditions}}{\text{Total Instances}}$$

$$\text{Precision (P)} = TP \div (TP + FP)$$
$$\text{Recall (R)} = TP \div (TP + FN)$$

where

TP: - Positive data with correct classification

TN: - Positive data with incorrect classification

FN: -Negative data with incorrect classification
FP: - Negative data with correct classification

To classify the data, we have to use some supervised or unsupervised algorithms. These algorithms are: -

A. Logistic Regression (LR)

It is a supervised classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means that there can only be two classes. Mathematically P (Y = 1) is a function of x. this technique is used for statically purpose [14]. The function used in this technique is sigmoidal function.

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

Where, X is the independent variable and a, b are the parameter of model. When the value of X is zero then the value of a yields P.

B. Support Vector Machine (SVM)

SVM are a classification methodology that is based on the notion of structural risk-minimization. In order to determine the decision function, SVM identifies the best hyper plane that partitions the data with no inaccuracy if the training data were linearly separable. [15].

C. Naïve Bayes (NB)

Naïve Bayes is an algorithm based on the Bayesian theorem. It is not a single algorithm but a family of algorithms that all share the same principle, that each pair of properties is classified independently of the other.

IV. RESULTS

To analyze the sentiments of customers, the proposed model requires some comments, post or blogs related to the products for that purpose we took reviews of customers on the social media from the data set IMDB. This data set is taken from Kaggle which have 50000 movies review of IMDB (Internet Movie Database). In 50000 entities 25000 are positive tweets and 25000 are negative tweets, which has shown in graph and the word cloud of positive and negative words.

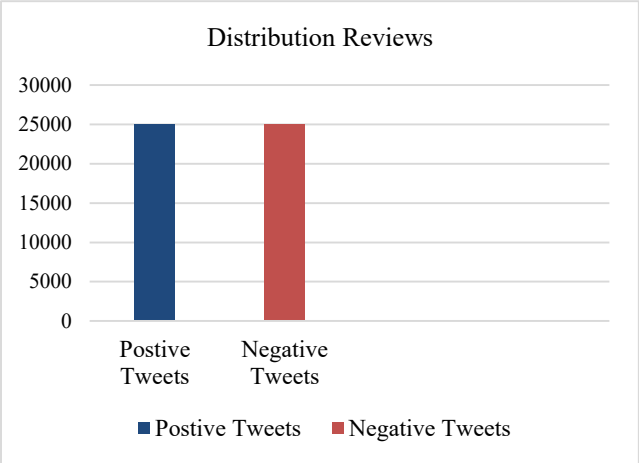

Figure 2 Distribution of Reviews of IMDB data set.

In the proposed model, after preprocessing and text vectorization, we applied state of art machine learning classification techniques. The received accuracy of Logistical regression, Naïve Bayes and SVM classifications are 75%,75% and 58% respectively. When we increased the epoch then this accuracy is changed. Therefore, we need a model which should be static, if we increase or decrease the epoch size. So, for the same purpose we used probabilistic model which can produces the best accuracy.

Performance Measurement

To measure the performance of the proposed model, we also calculated the measured parameters [16] like True positive rate, False positive rate and F-score, etc.

A. False Positive rate (FPR)

It's the likelihood of a false alarm being triggered: a positive result being given while the genuine result is negative. FPR also named as miss rate. The false positive rate can be calculated as:

$$FPR = \frac{FP}{FP + TN}$$

Where the number of false positives denoted as FP and number of true negatives denoted as TN.

B. True Postive Rate (TPR)

The TPR is the chance that a true positive would tested positive. TPR also named as sensitivity. The true positive rate can be calculated as:

$$TPR = \frac{TP}{TP+FN}$$

Where the number of false negative denoted as FN and number of true positives denoted as TP.

C. F-Score

The F_score is the harmonic average of the P (precision) and R (recall) depending on the first three decisions. F-Score can be calculated as:

$$P = \frac{TP}{TP+FP}, \quad R = \frac{TP}{TP+FN}, \quad F = \frac{2PR}{P+R}$$

The measure parameters are calculated for Naïve Bayes, Linear regression, SVM classifier and our proposed model. The calculated values of precision, recall, F-1 score and accuracy is shown in the figure 3.

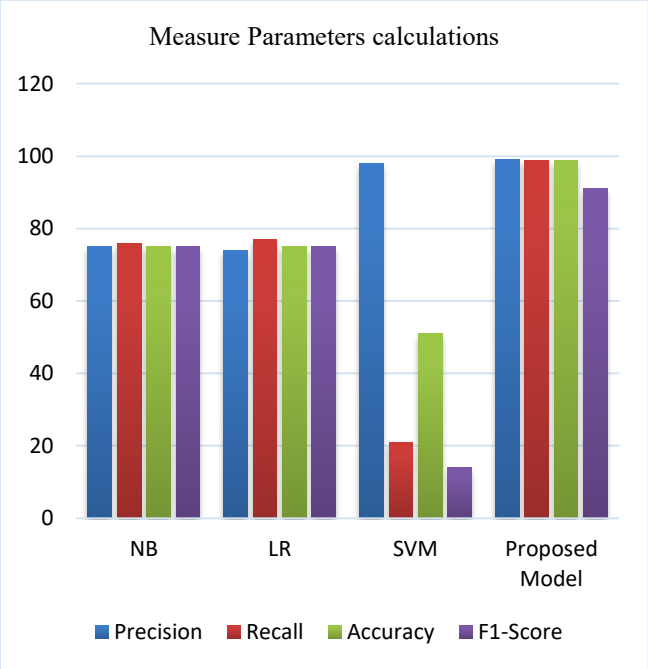

Figure 3 Measure parameters calculation.

IV. CONCLUSION

In this paper, we discussed sentiment analysis of movie using IMDB data set. The demand for movies and web series is increasing day by day. Customers post related product reviews on their social networking sites. We can easily collect information and analyze customer feedback. In this paper, we have proposed a framework for analyzing customer sentiment, using probabilistic model. Dataset used in the paper regarding movie ratings and reviews are taken from IMDB dataset. However, the results show that Naive Bayes, Logistical Regression have 75% accuracy and Support Vector Machine has only 58% accuracy while our proposed model has accuracy of 98.7%.

REFERENCES

[1] Gui, L., Jia, L., Zhou, J., Xu, R., & He, Y. (2020). Multi-Task Learning with Mutual Learning for Joint Sentiment Classification and Topic Detection. IEEE Transactions on Knowledge and Data Engineering.

[2] Gautam, G., & Yadav, D. (2014, August). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In 2014 Seventh international conference on contemporary computing (IC3) (pp. 437-442). IEEE.

[3] Gómez, L. M., & Cáceres, M. N. (2017, June). Applying data mining for sentiment analysis in music. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 198-205). Springer, Cham.

[4] Luo, L. X. (2019). Network text sentiment analysis method combining LDA text representation and GRU-CNN. Personal and Ubiquitous Computing, 23(3-4), 405-412. (2020).

[5] Khan, D. M., Rao, T. A., & Shahzad, F. (2019). The classification of customers' sentiment using data mining approaches. Global Social Sciences Review, 4, 146-156.

[6] Shoeb, M., & Ahmed, J. (2017). Sentiment analysis and classification of tweets using data mining. International Research Journal of Engineering and Technology (IRJET), 4(12).

[7] Das, O., & Balabantaray, R. C. (2014). Sentiment analysis of movie reviews using POS tags and term frequencies. International Journal of Computer Applications, 96(25).

[8] Kher, D. (2021). Multi-label emotion classification using machine learning and deep learning methods (Doctoral dissertation, Laurentian University of Sudbury).

[9] Rahman, R., Masud, M. A., Mimi, R. J., & Dina, M. N. S. (2021, December). Sentiment Analysis on Bengali Movie Reviews using Multinomial Naïve Bayes. In 2021 24th International Conference on Computer and Information Technology (ICCIT) (pp. 1-6). IEEE.

[10] Pradhan, R., Gangwar, K., & Dubey, I. (2022). PDF Text Sentiment Analysis. In International Conference on Innovative Computing and Communications (pp. 679-690). Springer, Singapore.

[11] Basarslan, M. S., & Kayaalp, F. (2021). Sentiment analysis on social media reviews datasets with deep learning approach. Sakarya University Journal of Computer and Information Sciences, 4(1), 35-49.

[12] A. K. Sharma, S. Chaurasia, and D. K. Srivastava, "Sentimental Short Sentences Classification by Using CNN Deep Learning Model with Fine Tuned Word2Vec," Procedia Comput. Sci., vol. 167, pp. 1139–1147, 2020, doi: 10.1016/j.procs.2020.03.416.

[13] L. L. Jiaxin Ma, Hao Tang, Wei-Long Zheng, "Emotion Recognition using Multimo dal Residual LSTM NetworkTitle," ACM, pp. 176–183, 2019, doi: https://doi.org

[14] Pandya, V., Somthankar, A., Shrivastava, S. S., & Patil, M. (2021, December). Twitter Sentiment Analysis using Machine Learning and Deep Learning Techniques. In 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4) (pp. 1-5). IEEE.

[15] Dabade, M. S. (2021). Sentiment Analysis Of Twitter Data By Using Deep Learning And Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(6), 962-970.

[16] Rathor, S., & Jadon, R. S. (2019). The art of domain classification and recognition for text conversation using support vector classifier. International Journal of Arts and Technology, 11(3), 309-324.