Sentiment Analysis using NLP and Machine Learning Techniques on Social Media Data

2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) | 978-1-6654-3789-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICACITE53722.2022.9823708

M.Kavitha
Department of CSE
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Chennai, India
mkavi277@gmail.com

Basetty Mallikarjuna,
Associate Professor, School of Computing Science and Engineering
Galgotias Univerisy, Greater Noida,
Uttar Pradesh, India 201 301
basetty.mallikarjuna@galgotiasuniversity.edu.in

R.Srinivasan
Department of CSE
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Chennai, India
rsrinivasan@veltech.edu.in

Bharat Bhushan Naib,
Associate Professor, School of Computing Science and Engineering
Galgotias Univerisy, Greater Noida,
Uttar Pradesh, 201 301,
bharat.bhushan@galgotiasuniversity.edu.in

R.Kavitha
Department of CSE
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology
Chennai, India
rkavitha1984@gmail.com

Abstract - **Internet usage has made social media an integral part of our everyday lives. With the aid of Natural Language Tool Kit (NLTK), sentiment analysis refers to the process of identifying and analyzing a piece of writing in order to determine whether its sentiment, opinions, views, and emotions are positive, negative, or neutral towards a specific issue, item, etc. People today depend on social media to stay connected. Users are allowed to put their ideologies on Twitter, a widely used communication site. People can write short messages and leave comments. An organization can analyze Twitter sentiments to find out how its image is discussed by individuals. With numerous applications for different spaces, there are numerous methods of sentiment analysis. The two main strategies for analyzing opinions are knowledge base and machine learning. In this study, the Twitter data was collected from tweets that were tagged in voting systems. Text mining was used to pre-process Tweets. Then, using the inverse document frequency and term frequency, a vector space model was constructed, and then sentiment analysis was carried out with Random Forest Classifier, Decision Tree Classifier, and Logistic Regression algorithms. Experiments are discussed and conclusions are drawn.**

Keyword: Sentiment Analysis, Machine Learning, Regression, Tweets, Random Forest, Logistic Regression

I. INTRODUCTION

By using the Internet, people are able to communicate instantaneously instead of using telegraphs and letters. Communication environment today permits people to be in touch with anyone with the social media, a new application software that emerged with the smartphone technology. Both individuals and institutions are affected. Shared places, such as movie theaters, stores and cafes, or expressions of positive or negative opinions about them affect everyone and society as a whole. Nowadays, social media is the primary way to communicate. Social media gathers the people and they can communicate their emotions, ideas, personal thoughts, events taken place with others. Due to this, social media has become a useful source of information, which companies use to promote their products as well as researchers to study people's feelings. The success of social media has been enabled by people constantly expressing their opinions about social, economic, health and brand issues, as well as products and brands. Texts are analyzed for sentiment expressions as part of the sentiment analysis studies. Depending on their positive or negative content, people are evaluating their texts. Business can use sentiment analysis to

introduce new products, new movies, etc. based on a preliminary study. Various classification algorithms are used to estimate sentiment on data categorized as true, false or unbiased during the process. Text mining methods are used to preprocess the text before classification. A satisfaction survey is a way to measure consumers' perceptions of products, services, and organizations. According to numerous researchers, the quality of products or services and the happiness of customers are the most important aspects of business performance. Businesses need to carefully evaluate what their customers need and want in order to maintain their competitive advantage. Furthermore, they must be able to satisfy their customers so they will do business with them. Based on the findings, the airline industry has struggled to provide outstanding services for diverse customer groups. There are many unstructured data types in social networks and other platforms. The process of obtaining customer opinions and making decisions based on them is difficult. An analysis of sentiments is a powerful method for determining people's opinions. In Sentiment Analysis, the aim is to determine whether textual data are positive, negative, or neutral in nature. Decision makers can track changes in public or customer sentiment about entities, activities, products, technologies, and services using sentiment analysis tools. Through Sentiment Analysis, businesses can easily improve their products and services, and political parties and social organizations can produce high-quality work. Sentiment Analysis facilitates understanding broad public opinions within a short period of time. Social media platforms generally provide data for sentiment analysis, which is stored as files called datasets. However, analyzing sentiment gets more difficult as datasets become imbalanced, large, multi-classed, etc. This paper analyzes Twitter data. The data is pre-processed and vectorized using NLP techniques. Thereafter, Machine Learning algorithms were

used to classify textual data polarity. In conclusion, we compared the applied Machine Learning algorithms with NLP techniques in order to determine which approach was most effective.

II. LITERATURE REVIEW

[1] Pang et al. used movie database and performed sentiment analysis for movie comments using Maximum Entropy, Naive Bayes (NB) and Support Vector Machines (SVM). SVM provides 82.9% accuracy with unigram dataset, which is the best among the four algorithms. In another study, the TF-IDF was used for emotional analysis of tweets passed during Egyptian presidential election. Elghazaly et al., [2] and the SVM,NB classifier was used. According to their comparison in the study, the accuracy and error rate of the NB method were the highest. Hamoud et al.[3] have analyzed Twitter data for classification of political tweets using Bag of Words (BOW), TF and TF-IDF. SVM and NB are the algorithms used for classification. SVM enabled with BOW provides the best accuracy and F-measure, based on the results. Nikfarjam et al., [4] used Twitter comments of patients to conduct a sentiment analysis of side effects of medication. After comparing the SVM algorithm's performance with the other methods, the researchers found that it performed better by 82.11%. In [5, 11] Yachika Gupta and co-authors suggest various learning models based on tweets collected from an internet database. By analyzing the public sentiment toward various political leaders expressed on Twitter, a case study of election results prediction for the February 2017 elections in Punjab has been used to validate the proposed system. The developed sentiment analysis system is capable of performing real-time analysis of tweets and presenting the results as they occur. The results are presented via a dashboard. Graphical representations are displayed every minute, displaying tweets in real time. To predict the final results, Twitter messages are also saved as CSV files. When compared to actual results, the system produced very encouraging results with very little diversion. Political entities should understand public opinion and, as a consequence, choose their tactics accordingly, according to Jyoti Ramteke et al. [6]. A considerable number of people have viewed sentiment analysis as an enlightening instrument for spotting client preferences and inclinations via online media. To perform Sentiment analysis, Supervised Learning Algorithms such as Naive Bayes and SVM need prepared information. It is equally important that the quality (highlights and logical significance) of the named preparing information is considered as part of the precision of the calculations. The absence of preparation information leads most applications to resort to cross-area assumption checks, which ignore details relevant to the objective information. Generally, grouping of text is less precise when preparation information is lacking. This paper proposes a two-phase system which can be utilized to make an informed preparation from extracted Twitter data without needing to settle on highlights and logical relevance. The approach comprises two phases that can be used to predict the political decision results using an adaptive machine learning model. M. Trupthi et al [7] developed an interactive automatic system that can predict the sentiment of reviews or Twitter comments posted by individuals using Hadoop, which is capable of processing enormous amounts of data. Currently, emotion classification, feature-based classification, and managing negations are the most prevalent problems in this research community. For predicting sentiment polarity, a precise method is employed, which contributes to improving marketing strategies. This paper examines the challenges that arise during Sentiment Analysis. Using Twitter tweets as the source of data, the paper performs real-time sentiment analysis and provides time based analytics to users. Kavya Suppala et al [8] addressed sentiment analysis using Twitter data, namely using naive bayes classifier to categorize tweets. By predicting an unknown tweet's sentiment, we are able to provide precise outcomes and generate an arrangement procedure by using machine learning [9,10]. Sentiment analysis is an important technique that is used to discover the upsides and downside of their products in the marketplace in order to improve their efficiency [11,12]. On the dataset containing tweets, the model developed in this research utilized Natural Language Toolkit (NLTK) [13]. There is an idea of "sack of words" which encompasses both true and false words. The data with the highest value was considered as true or false [14]. The Naive Bayes classifier determines the probability of new information. As a result, we chose a Twitter dataset with efficient features that improves the accuracy and effectiveness of the classifier. In the event more features need to be consolidated in the information base, this model can also be enhanced to any desired level [15].

III. PROPOSED METHODOLOGY

Twitter enables us to share texts, pictures, and videos instantly with 280 characters. Using the Twitter API, 31962 president debate tweets were collected for this study. Python was used to preprocess these data and calculate sentiment scores. Fig. 1 shows the process of Text minin

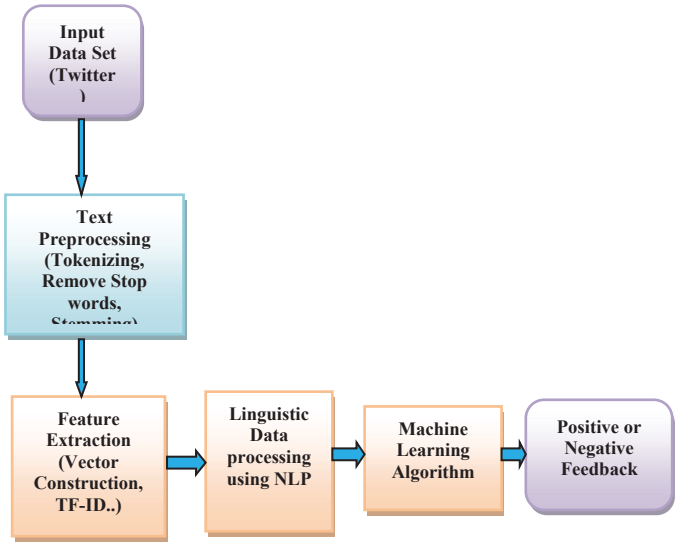

Fig.1. Proposed Methodology Flow Diagram

A. Text Analysis

Prior to analyzing Twitter data, stop words are removed, characters are converted to lowercase, root words are identified and stop words are omitted. These operations were performed using the Python NLTK library. For the feature extraction, the term frequency (TF) and inverse document frequency (IDF) are used. Terms are mined based on their occurrence in the documents. In spite of many terms, the IDF is used to normalize the frequency of terms.

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

Equation 1 shows how term weights are calculated in a document. Using this we can identify the stop words based on the number of words found in multiple documents. This can be accomplished by finding the Inverse Document Frequency, which is shown in (2) [9].

Inverse Document Frequency (i) x Term Frequency (i) is the TF_IDF at term i in document j

$$TF_{(i,j)} = \frac{Term\ i\ frequent\ in\ document\ j}{Total\ words\ in\ document\ j} \quad (1)$$

$$IDF(i) = \log\left(\frac{Total\ documents}{documents\ with\ term\ i}\right) \quad (2)$$

C. Classification

Random Forest Algorithm

Input: Data

a. Random forest takes a set of k records and randomly selects n records from it.
b. For each sample, a decision tree is constructed.
c. An output will be generated by each decision tree.
d. For classification and regression, the final output is determined by majority voting or by averaging

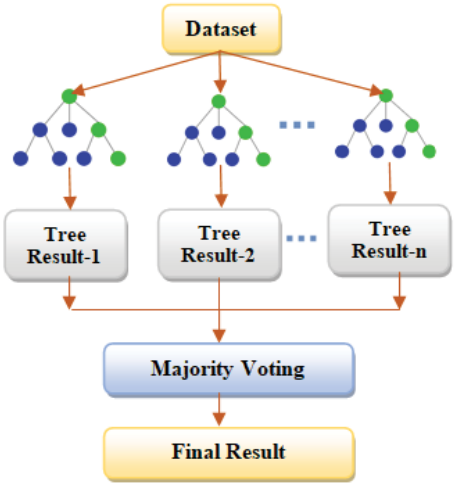

Fig.2. Random Forest Tree

Decision Tree Algorithm

Input : Data

a. Determine the value of T // predicted class
b. Repeat: 1 to M
   i. Using training and test data, calculate d(x,y)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

c. D(x, y) must be sorted descending
d. From the arranged list, fill in the top k results
e. Take the class that is most frequent from this list

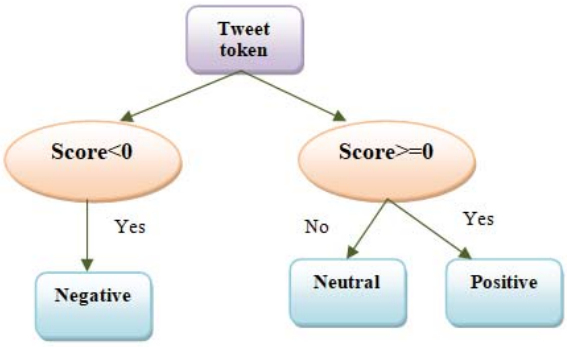

Fig.3. Decision Tree

Logistic Regression

Input: Data

a. Iterate i←1 to n
b. For each data instance di
c. Set the target value for the regression
d. Set the weight of each instance di to P(1|di).(1-P(1|di)

$$z_i \leftarrow \frac{y_i - P(1 \mid d_j)}{[P(1 \mid d_j) \times (1 - P(1 \mid d_j))]}$$

e. F(i) is finalized with class value (zi) and weights (wi)
f. Assign the Class Label

Fig.4. Logistic Regression

IV. EXPERIMENTAL ANALYSIS

The study analyzed 31962 tweets. Sentiment analysis used Random Forest Trees, Decision Trees, and Logistic Regressions to classify the sentiment. Model selection was then used to separate the training and test data sets after text preprocessing and vector space modeling. Figure 5 and 6 shows the distribution of tweet data and Frequently occurring words. Evaluation of the three algorithms was based on their accuracy and F-Score. Table 1 and figure 7 shows the result.

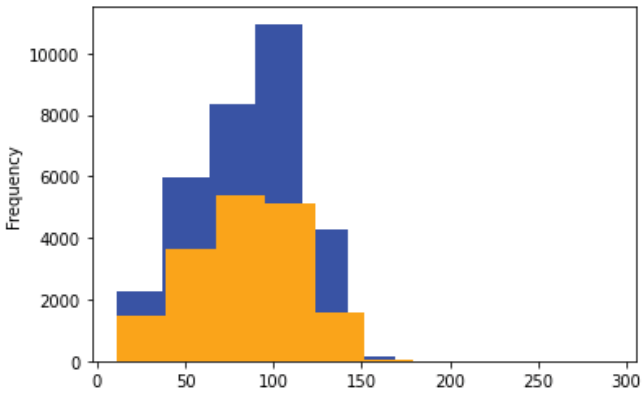

Fig.5. Distribution of Tweet data

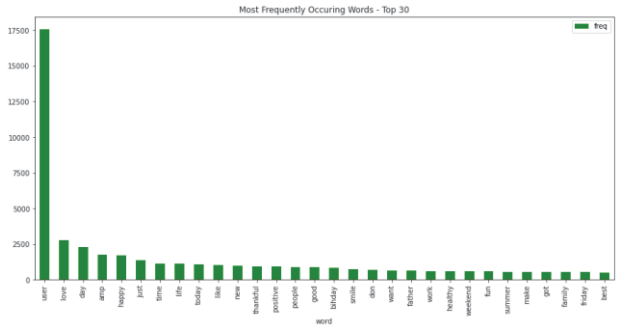

Fig.6. Frequently Occurring words

TABLE I. RESULTS OF ALGORITHMS

Algorithm	Accuracy	F-Score
Random Forest Tree	0.95	0.61
Decision Tree	0.83	0.55
Logistic Regression	0.78	0.59

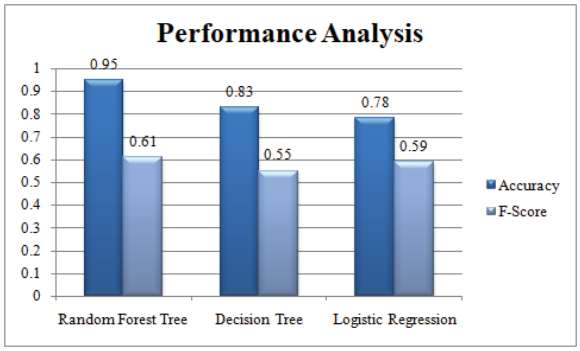

Fig.7. Performance Analysis of Algorithms

Performance is assessed after the dataset is analyzed using classifier models. The random forest tree presented the highest performance. We analyzed the Twitter dataset following the TFIDF word vector process, and the resulting performance values were found to be similar.

V. CONCLUSION

Using Twitter datasets and fields, we analyzed a variety of sentiment classification technologies. Only textual information is used for our discussion and reviews, which are drawn from social networking sites. A process for preprocessing, and further classification of social data is retrieved by machine learning techniques. The study concludes that Random Forest Tree algorithm outperforms with 95% accuracy. Several challenging problems in the field of sentiment analysis exits remain to be solved, and there is a great deal of future scope for research in the filled with sentiment analysis. Only a few researchers focus on user's replies to tweets retrieved from social networking sites like Twitter for predicting their behavior. Based on a few machine learning techniques as a reference model, we will develop an algorithm for predicting human or social networking users' future behavior.

REFERENCES

[1] Pang, B., Lee, L. and Vaithyanathan, S., Thumbs up? Sentiment classification using machine learning techniques. Published in EMNLP, 2002.

[2] Elghazaly, T., Mahmoud, A. and Hefny, H.A., March. Political sentiment analysis using twitter data. In Proceedings of the International Conference on Internet of things and Cloud Computing (pp. 1-5), 2016.

[3] Al Hamoud, A., Alwehaibi, A., Roy, K. and Bikdash, M., Classifying political tweets using Naïve Bayes and support vector machines. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (pp. 736-744). Springer, Cham, 2018.

[4] Nikfarjam, A., Sarker, A., O'connor, K., Ginn, R. and Gonzalez, G., Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association, 22(3), pp.671-681, 2015.

[5] Gupta, Y. and Kumar, P., Real-Time Sentiment Analysis of Tweets: A Case Study of Punjab Elections. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-12). 2015.

[6] Ramteke, J., Shah, S., Godhia, D. and Shaikh, A., Election result prediction using Twitter sentiment analysis. In Proceedings of IEEE International conference on inventive computation technologies (ICICT), Vol. 1, pp. 1-5, 2016.

[7] Trupthi, M., Pabboju, S. and Narasimha, G., Sentiment analysis on twitter using streaming API. In Proceedings of IEEE 7th International Advance Computing Conference (IACC) (pp. 915-919). 2017.

[8] Suppala, K. and Rao, N., Sentiment analysis using naïve bayes classifier. International Journal Innovation Technology Exploration Engineering, Vol.8(8), pp.264-269, 2019.

[9] Sjögren, R., Stridh, K., Skotare, T., and Trygg, J., Multivariate patent analysis–Using chemometrics to analyze collections of chemical and pharmaceutical patents. Journal of Chemometrics, Vol.34(1), 2020.

[10] Mallikarjuna, Basetty, D. J. Anusha, and Munish Sabharwal. "An Effective Video Surveillance System by using CNN for COVID-19." Handbook of Research on Advances in Data Analytics and Complex Communication Networks. IGI Global, 2022. 88-102.

[11] Mallikarjuna, B., Addanke, S., & Anusha, D. J. (2022). An Improved Deep Learning Algorithm for Diabetes Prediction. In Handbook of Research on Advances in Data Analytics and Complex Communication Networks (pp. 103-119). IGI Global.

[12] Mallikarjuna, B., Addanke, S., & Sabharwal, M. (2022). An Improved Model for House Price/Land Price Prediction using Deep Learning. In Handbook of Research on Advances in Data Analytics and Complex Communication Networks (pp. 76-87). IGI Global.

[13] Mallikarjuna, B., Shrivastava, G., & Sharma, M. (2021). Blockchain technology: A DNN token ‐ based approach in healthcare and COVID ‐ 19 to generate extracted data. Expert Systems, e12778.

[14] Mallikarjuna, Basetty, et al. "Development of Efficient E-Health Records Using IoT and Blockchain Technology." ICC 2021-IEEE International Conference on Communications. IEEE, 2021.

[15] Mallikarjuna, Basetty, et al. "The effective SVM-based binary prediction of ground water table." Evolutionary Intelligence (2020): 1-9.