2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) | 978-1-6654-9602-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/COM-IT-CON54601.2022.9850477

Sentiment Analysis of Twitter Data Using TF-IDF and Machine Learning Techniques

Satyendra Singh
Dept. of Computer Science
Gurukula Kangri (Deemed to be University)
Haridwar, Uttarakhand, India
satyendra.singh701@gmail.com

Krishan Kumar
Dept. of Computer Science
Gurukula Kangri (Deemed to be University)
Haridwar, Uttarakhand, India
krishan.kumar@gkv.ac.in

Brajesh Kumar
Dept. of Computer Science & IT
M.J.P. Rohilkhand University
Bareilly, Uttar Pradesh, India
bkumar@mjpru.ac

***Abstract*—Sentiment analysis technique plays an important role in natural language processing to analyze complex human statements. In the last few years, this technique has become a powerful tool for several social media communication mediums such as WhatsApp, Twitter, Facebook, Instagram, YouTube, LinkedIn, Blog, etc. This paper proposes a machine learning (ML) based method to analyze social media data for sentiment analysis on text data. The presented method is divided into three distinct stages. In the first stage, pre-processing is performed to filter and refine the text data. In the second stage, the feature extraction is performed using the Term Frequency and Inverse Document Frequency (TF-IDF) technique. Moreover, during the third stage, the extracted features are supplied to make predictions for the classifier. The experiments are carried out on a publicly available Twitter dataset for US Airlines. Several ML techniques are utilized for analysis and classification. The results are reported for different evaluation metrics like accuracy, precision, recall, and F1 score. Finally, the support vector machine yielded the most relevant results.**

Keywords—Sentiment Analysis, Natural Language Processing, Machine Learning, Term Frequency and Inverse Document Frequency (TF-IDF), Twitter

I. INTRODUCTION

Sentiment Analysis is a well-known method for extracting subjective information on human opinions. SM provide a powerful medium for people to communicate their opinions and share various information. At present, 70% population of the world is familiar with the social media technology at the global level. Many people post the messages on numerous social media sources [1]. These comments are rich source of information; some provide information explicitly, while many contain implicit information. The sentiment analysis can support organizations to analyze the customer response automatically; understanding people's feelings, and review the customer opinion to take necessary decisions for improving the product and services for customer satisfaction.

Human feeling is a complex multi-level distinction that reflects a few characteristics of human character, atmosphere, and behavior. The statements made by humans are complex; and different people use different words to express the same concept. Natural language processing provides a way to extract hidden or implicit information from comments available on the social media.

Sentiment analysis techniques can process complex statements without human intervention to provide useful information. In recent decade, social networks are commonly used for sentiment detection and analysis of multiple ideas. S. K et al. [2] proposed a novel sentiment analysis that used various sentiment scores of tweet responses to calculate the impact score. They analyzed the user behavior based on their tweets to measure the effect on certain topics. M. K. Sohrabi et al. [3] developed a pre-processing technique to mine the Twitter data. Moreover, a pre-processing technique was combined with ANN and SVM to obtain good accuracy.

K. M. Hasib et al. [4] presented a DCNN based method for sentiment analysis of the Twitter dataset. The model obtained overall 91% accuracy. Furthermore, A. S. Neogi et al. [5] extracted features from the Twitter dataset using TF-IDF approach. The tweets are classified with NB, DT, RF and SVM. The authors reported that RF classifier outperformed the other three classifiers.

B. Gaye et al. [6] introduced a sentiment approach for text datasets using long short-term memory (LSTM), TF-IDF and other traditional ML techniques. They extracted the features with the TF-IDF technique and classified the input using LSTM to obtain best performing results. R. Arulmurugan et al. [7] used cloud machine learning to integrate SVM, ANN, and naïve Bayes methods to classify the tweets. Moreover, the authors applied k-means clustering on the data to handle the outliers.

Twitter is among the most popular microblogging and social networking sites. It is mainly used by performers, diplomats, sports players, corporations, and other different sources throughout the world. In this paper, a sentiment analysis method for text data from Twitter has been proposed. The TF-IDF and machine learning techniques have been successfully used here. The extracted tweet features are expressed in matrix representation based on the TF-IDF technique. Moreover, several classification algorithms are compared and analyzed.

II. PROPOSED METHODOLOGY

A ML based sentiment analysis framework has been proposed in this method to analyze the Twitter data. It has been especially developed for text data only. The step-by-step procedure adopted in the proposed methodology is shown in Fig. 1. From figure, initially the raw data (US airlines sentiment tweets) is preprocessed, where various operations including punctuation, lower case, tokenization, stemming, and removing unimportant words (stop words) or characters are performed. The preprocessed data then used to generate useful features. The features are inputs into the classifier, which makes predictions about the customer opinion.

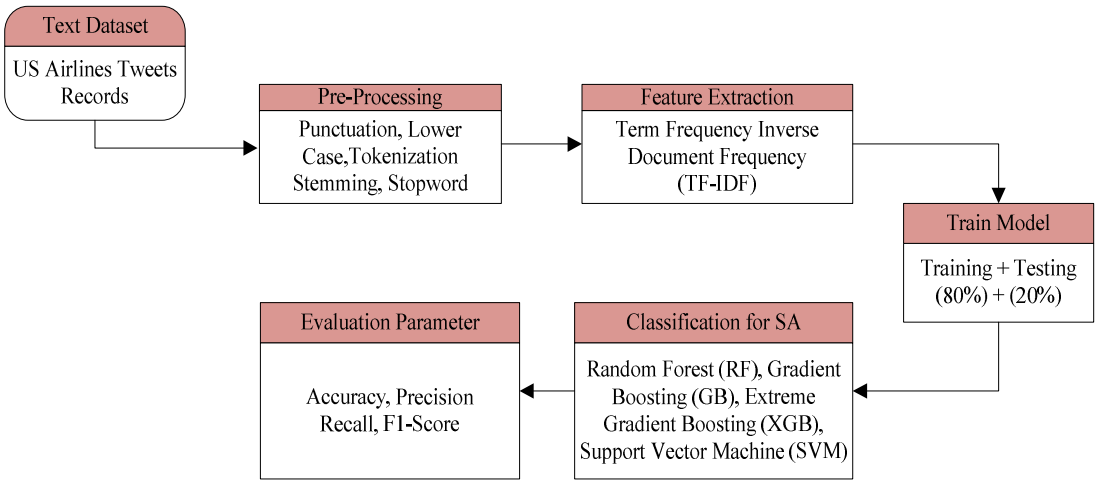

Fig. 1. Proposed Methodology

A. *Pre-processing*

Data pre-processing is performed to refine the raw data (US airlines sentiment tweets) to make it suitable for further processing. It is important to make data clean and in proper format for the better performance of the machine learning algorithms. The main parameters exist for preprocessing step, data cleaning, normalization, and elimination of duplicate values. All preprocessing techniques applied for US airlines dataset. These techniques are used to normalize text data. At this stage, the unnecessary words and characters such as punctuation marks are eliminated for example: %, !, $, @, dot,?, / etc. In addition, all uppercase letters are converted to lowercase as both make no difference to the sentiment. The raw text is divided into smaller chunks with the help of tokens. The process of tokenization consists of dividing each paragraph into single units of word and sentence to interpret the text. Stemming is a kind of normalization technique, used to reduce computational load in natural language processing. It reduces the inflected words to their root words example: compute, computing, computes, computed. Many words in text tweets such as I, will, is, for, a, an, the, etc. have no sense of phrases and therefore are unnecessary for sentiment analysis. These words are called stop words [3]. All these words are deleted and the sentiment analysis is performed on words concerned with the English terms.

B. *Feature Extraction*

When extracting data features, an immense problem arises in sentiment analysis. Due to its clarity and efficacy, the TF-IDF models are regularly used for so many natural language processing techniques. The TF-IDF technique assists a sentiment analysis by analyzing sentiment text corpora. This method is a mathematical-statistical approach, which is determined the significance a word to the corpus in a text document [8]. TF-IDF method obtains the dominant features based on the word frequency of each document from text data. It is the multiplication of two sections, namely TF and IDF. It calculates the values of TF and IDF using the following formula [9].

$$TF = \frac{\text{No. of times terms appears in a document}}{\text{No. of terms in a document}}$$

$$IDF = \log \frac{\text{Total No. of documents}}{\text{Total No. of documents in which term appears}}$$

C. *Classification Method*

The features extracted from the raw data (US airlines sentiment tweets) are provided as input to the classifier that classifies the input into given categories. There are various ML algorithms for example random forest (RF), support vector machine (SVM), gradient boosting (GB), and extreme gradient boosting (XGBoost), etc. are widely used for sentiment analysis. RF is a versatile and direct method which does not involve tuning of many hyper parameters. It is a kind of ensemble classifier and uses multiple decision trees [10]. GB is another ensemble learning based classifier which reduces the risk of overfitting. It aims to minimize the expected value of a loss function. Like gradient boosting, XGBoost is also uses decision tree as base classifier. It uses feature interaction constraints for limiting the input attributes, which is beneficial to reduce loss and increase accuracy [11]. SVM is a supervised method that is utilized for both regression and classification. This algorithm is more flexible and easier to implement than other classification algorithms. SVM algorithm purposes to locate a hyper-plane using different classes data points in the n-dimensional feature space [12]. It can solve linear and non-linear problems accurately with low computing power.

In the method shown in Fig. 1, all the techniques discussed in this section are used at appropriate stage. The preprocessing includes removal of unnecessary words, punctuation, lower case tokenization, and stemming. Feature extraction is done using TF-IDF algorithm. At classification stage various techniques including RF, GB, XGBoost and

SVM are used individually. The results of all individual classifiers are compared and analyzed.

III. RESULTS AND ANALYSIS

The proposed method is evaluated on a publicly available benchmark Twitter dataset collected for US airlines. It contains 14,640 records and 15 attributes. The Tweets messages distributed in three broad categories: negative, neutral, and positive. The negative class contains the highest number of 9178 tweets accounting for 62.69%. The positive class contains 2363 tweets accounting for 16.14%, and the remaining 3099 tweets accounting for 21.17% lie in neutral class. The fraction of each class is illustrated in Fig. 2 with the help of pie chart.

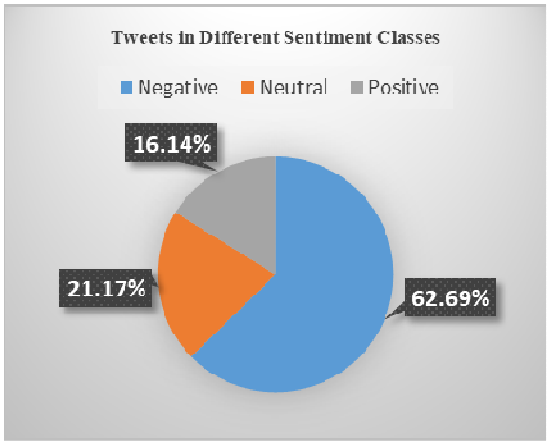

Fig. 2. Tweets in Different Sentiment Classes

TABLE I. DISTRIBUTION CLASS OF SENTIMENT TWEETS

Sentiment Class	Negative		Neutral		Positive	
No. of Tweets	Training	Test	Training	Test	Training	Test
	7308	1870	2485	614	1919	444

The dataset is divided into two sections: training and testing sets. The 80-20 rule are utilized as training set (80%) and remaining 20% samples form the test set as given in Table I. The experimental results are analyzed in terms of various evaluation metrics for example precision, recall, F1 score and accuracy. The calculation equations [4] as shown below.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}} \quad \text{(I)}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive}+\text{False Negative}} \quad \text{(II)}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision}+\text{Recall}} \quad \text{(III)}$$

$$\text{Accuracy} = \frac{\text{True Positive}+\text{True Negative}}{\text{True Positive}+\text{True Negative}-\text{False Positive}+\text{False Negative}} \quad \text{(IV)}$$

The results are reported on these evaluation parameters as depicted in Table II. From Table II, its show the SVM classifier achieves the highest accuracy of 83.74%. It is also shown in Fig. 3 in graphical form. The precision of all the classifiers is good. SVM works more precisely as observed from the table. The Recall is greater than 0.5, which is desirable. The F1-Score is also best for the SVM classifier. The class wise results are reported in Table III. It observed that the SVM performs better for Negative class. For Positive and Neutral classes RF gives better results than other. The performance of the SVM classifier is better than among all used classifiers in our experiment.

TABLE II. ACCURACY OF ADOPOTED METHODOLOGY FOR TF-IDF TECHNIQUE

Classification Techniques	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	77.90	75.36	96.36	84.57
GB	80.46	80.46	91.65	85.69
XGBoost	81.55	81.90	91.28	86.33
SVM	83.74	84.00	92.08	87.85

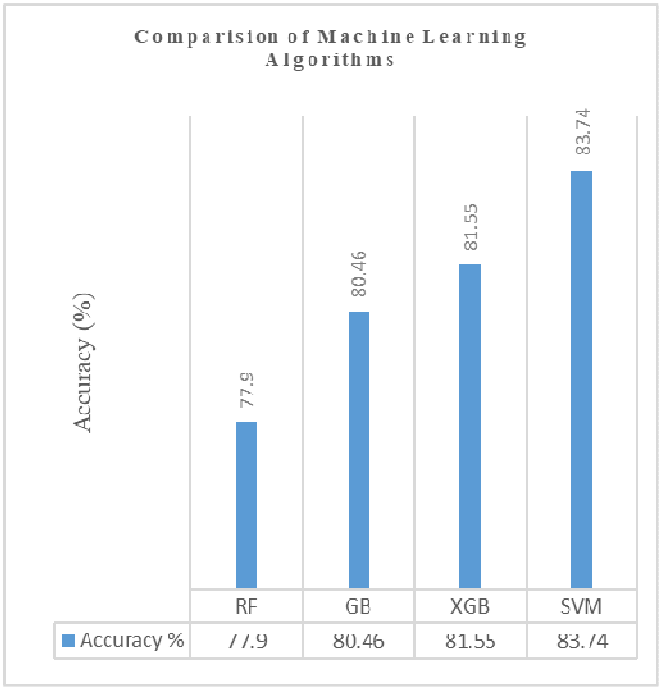

Fig. 3. Comparision of Machine Learning Algorithms

IV. CONCLUSION AND FUTURE DIRECTION

This paper described the ML based method for analyzing the sentiment tweets of Twitter dataset. The feature extraction and classification, two major components of the proposed method, used TF-IDF technique for feature extraction to select effective features. Various classification

methods have been successfully tested as classifiers. The US airline sentiment twitter dataset has been used for evaluating the method. Finally, it is found that SVM gives the best results with 83.74% accuracy, 84% precision, and better F1-score of 87.85%. The future work includes testing other feature extraction techniques.

TABLE III. CLASSIFICATION RESULTS FOR DIFFERENT SENTIMENT CLASS WITH MACHINE LEARNING TECHNIQUES

Classification Techniques	Accuracy (%)	Negative Class			Neutral Class			Positive Class		
		Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
RF	77.90	0.76	0.96	0.85	0.69	0.39	0.50	0.84	0.38	0.52
GB	80.46	0.80	0.92	0.86	0.65	0.49	0.56	0.78	0.59	0.67
XGBoost	81.55	0.82	0.91	0.86	0.63	0.51	0.57	0.78	0.61	0.68
SVM	83.74	0.84	0.92	0.88	0.67	0.57	0.62	0.79	0.63	0.70

REFERENCES

[1] R. B. Koyel Chakraborty, Siddhartha Bhattacharyya, "A survey of sentiment analysis on social media," *IEEE Trans. Comput. Soc. Syst.*, vol. 4, no. 1, pp. 1–11, 2020, doi: 10.1109/TCSS.2019.2956957.

[2] S. K. and A.Reda, "Emotion and sentiment analysis from Twitter text," *J. Comput. Sci. Elsevier*, vol. 36, p. 101003, 2019, doi: doi.org/10.1016/j.jocs.2019.05.009.

[3] M. K. Sohrabi and F. Hemmatian, "An efficient preprocessing method for supervised sentiment analysis by converting sentences to numerical vectors: a twitter case study," *Multimed. Tools Appl.*, 2019, doi: 10.1007/s11042-019-7586-4.

[4] K. M. Hasib, M. A. Habib, N. A. Towhid, and M. I. H. Showrov, "A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service," in *2021 IEEE International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, 2021, no. July, pp. 450–455, doi: 10.1109/ICICT4SD50815.2021.9396879.

[5] A. S. Neogi, K. A. Garg, R. K. Mishra, and Y. K. Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100019, 2021, doi: 10.1016/j.jjimei.2021.100019.

[6] B. Gaye, D. Zhang, and A. Wulamu, "A tweet sentiment classification approach using a hybrid stacked ensemble technique," *Inf.*, vol. 12, no. 9, 2021, doi: 10.3390/info12090374.

[7] R. Arulmurugan, K. R. Sabarmathi, and H. Anandakumar, "Classification of sentence level sentiment analysis using cloud machine learning techniques," *Cluster Comput.*, vol. 22, no. S1, pp. 1199–1209, Jan. 2019, doi: 10.1007/s10586-017-1200-1.

[8] 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)*, 2017, vol. 2018-Septe, pp. 80–83, doi: 10.1109/DCABES.2017.24.

[9] B. Ray, A. Garain, and R. Sarkar, "An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews," *Appl. Soft Comput.*, vol. 98, no. xxxx, p. 106935, Jan. 2021, doi: 10.1016/j.asoc.2020.106935.

[10] Ankit and N. Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis," in *Procedia Computer Science*, 2018, vol. 132, no. Iccids, pp. 937–946, doi: 10.1016/j.procs.2018.05.109.

[11] E. K. Ampomah, Z. Qin, and G. Nyame, "Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement," in *Information*, 2020, vol. 11, no. 6, p. 332, doi: 10.3390/info11060332.

[12] S. Kumar and M. Zymbler, "A machine learning approach to analyze customer satisfaction from airline tweets," *J. Big Data*, vol. 6, no. 1, pp. 1–16, 2019, doi: 10.1186/s40537-019-0224-1.