IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis of Twitter Data Using Machine Learning Algorithm

Kunnal Jaluthria
Department of Applied Mathematics
Delhi Technological University
Delhi, India

Sumedha Seniaray
Department of Applied Mathematics
Delhi Technological University
Delhi, India

***Abstract*—Sentiment analysis is a technique used for mining data, viewing, evaluating, or analyzing a sentence to predict its emotion using Natural Language Processing. It is done to gain an impartial understanding of the writer's viewpoint towards any particular topic. In this paper, sentiment analysis of political leader's timeline tweets which they have tweeted during the election campaign and also include to find out sentiment of the tweets in which these political leaders are tagged. This paper proposes a quicker method of sentiment analysis for a broad population based on the use of multiple classifiers to categorise data opinions as positive, negative, or neutral. Valence Aware Dictionary and Sentiment Reasoner were used for faster and accurate labelling of tweets. Support Vector Machine obtained the highest accuracy out of all the remaining classifiers.**

Keywords*— *Sentimental analysis, Twitter, Valence Aware Dictionary and Sentiment Reasoner, Machine Learning, Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machine.

I. INTRODUCTION

Sentiment Analysis is a technique for determining how a company's consumers or any group of people respond to new products or policies. The main aim of Natural Language Processing (NLP) is to understand and construct a natural language by using the required tools and technique. Sentiment Analysis can also be used to analyse people's feeling about a movie, product, song or other subjects as well as to distinguish between positive, negative, neutral reviews. It can be used to make better forecasts and suggestions in areas like the stock market, e-commerce websites, song recommendations, and so on. Figure 1 below shows the three classes: positive, negative and neutral.

Fig.1. Sentiment Analysis Classes

Users on Twitter can express their thoughts in the form of tweets of up to 280 characters. Twitter is a very popular micro-blogging platform where millions of people share their view in form of tweets. People often use an emoticon, slangs etc. As a result, it is common knowledge that Twitter language is unstructured. Sentiment analysis is used to extract relevant meaning from tweets, and the effect is expressed in terms of the total number of positive, negative, and neutral tweets. A variety of classifiers are used in getting the result and the most accurate one is taken into consideration [1].

The Problem Statement: In this modern era of communication everyone connected to the internet. Many analysts agree that social media has a major impact on election results. In the current scenario, a large portion of India's population uses social media to share their views and opinions on government policies and other social issues. This paper aims to get the sentiment on tweets of political leaders which they tweeted during the election period. This paper also contributes to finding out the sentiment of the tweets in which these leaders are tagged. We then train our model after combining all the data set and this will help to predict the sentiment and what is the accuracy by using machine learning algorithms.

The rest of the paper is organized as follows: Section II presents the relevant background work on sentiment analysis. Section III explains the proposed architecture. Section IV describes different performance metrics which is used for evaluation. Section V describes how the proposed approach is implemented to determine the sentiments associated with different candidates. Finally, Section VI is based on conclusions.

II. RELATED WORK

Several articles on sentiment analysis have been published over the years. This survey presents some of the methods available in the literature, based on the works available:

Analyzing Government Scheme Awareness Using Swachh Bharat Tweets [2]. This model, which employs Twitter as a database, decides whether a new government initiative would have a positive or negative impact on society. In order to achieve higher accuracy, a large number of tweets must be accessed.

The geolocation feature was used in [3] sentiment analysis of the Indian government's demonetization of 500 and 1000 rupee banknotes to make a country and statewide analysis of

reasons for people's dissatisfaction with this government policy.

In this paper [4], a highly suitable model have discussed how to use Twitter data to predict upcoming Hollywood and Bollywood films. They had accomplished this task using a classifier and features such as SVM and Naive Bayes. Both are used for high accuracy, but Naive Bayes outperforms SVM in terms of precision while SVM outperforms Naive Bayes in terms of recall.

In Large-Scale Sentiment Analysis for News and Blogs by the author [5], The Lydia text analysis method was used to evaluate the people's sentiments. The Lydia Text Analysis System is still in its early stages of growth. Instead of Lydia, a better lexicon dictionary that can recognize emoticons, slang terms, and offer a separate ranking for small and capital letters can be used in the following model.

Semantics is introduced as an additional feature to the training set for sentiment analysis in this paper [6]. The semantic features were fed into the Naive Bayes (NB) model training using the interpolation method. The Naive Bayes method, on the other hand, has several drawbacks, including data scarcity and zero frequency. As a result, other classifiers can be used to improve accuracy.

III. PROPOSED WORK

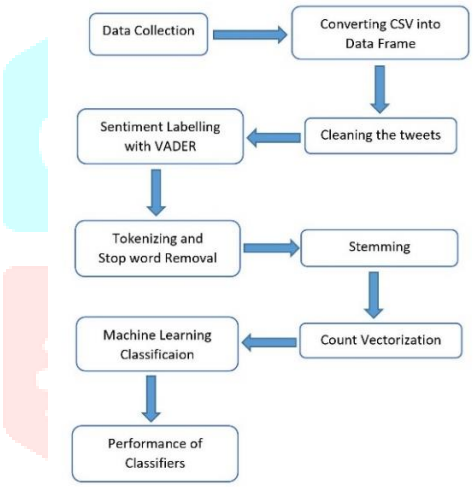

Fig.2. Proposed Methodology

A. Data Collection

The Twitter API is considered as a "Gold mine of Data", so we decided to go with that. In contrast to other social media sites, almost every user's tweets are completely transparent and extractable, resulting in a broad database for analysis, as discussed in [7]. So, to extract Twitter data, one has to make a Twitter developer account. You have to provide some necessary details for creating an application which later will be used for extracting the data. After our application is created, we will get access to some keys such as customer keys, access token key, customer secret key, access secret key. When a user wants to get data, these keys are very important. To retrieve a tweet from Twitter, we created a python script that uses the "Tweepy" python library. Python has an open source library called "Tweepy" that allows it to link to Twitter and collect data from their API, which we will use in our program.

B. Converting CSV into Data frame

Only the related objects are kept from the text file, such as objects containing the entire text of a tweet, screen name, screen id, date and time at which time the tweet is created. The rest of the item will be discarded.

C. Cleaning the tweets

Twitter language is in an unstructured format. It may consist of emoticons, empty spaces, URL's, @tag's, hashtags. We have to pre-process the data first by using the python libraries to retrieve the relevant parts of the data and removing the unwanted ones.

D. Sentiment Labelling with VADER

Sentimental Analysis is a form of statistical analysis that determines whether a piece of text is negative, neutral, or positive and it is performed using one of two methods: Valence-based or Polarity-based. The text is graded as positive or negative in a polarity-based strategy. This ensures that the terms "good" and "superior" would have the same sentiment, i.e. positive. When analysing a piece of text, VADER uses a valence-based approach. VADER analyses text sentiments using lexicons of sentiment-related words. It examines a text and determines if any of the words in the sentence are in its lexicon dictionary. It assigns them a positive, negative, or neutral rating [8]. VADER not only grades sentences based on the words in them, but also the capitalization of the words and the sentence structure. For example, say the sentence "The weather is pleasant today, and I'm in great shape." is considered. In the sentence, "pleasant" and "great" are rated 0.51 and 0.62 respectively. VADER also assigns a score to the sentence depending on the use of exclamation points or emoticons. As a result, it's perfect for social media information. Not only that, but it also considers the usage of modifying terms preceding a sentiment term, such as "extremely," "really," "too," and so on. For example, "just good" reduces the positive intensity of a sentence, while "so good" increases the positive intensity. Another advantage of VADER is that it can accommodate changes in the sentiment of a sentence when it includes the word "but". The sentiment of the sentence before and after the word "but" is taken into account by the rule, but the sentiment of the sentence after "but" is given greater weight than the one before it [9].

E. Stop Words Removal

Stop words are widely used words like 'is', 'an', 'in', 'of' etc. and others that are considered meaningless because they appear regularly in sentences and add no meaningful significance or weight to the sentiment. They unnecessarily increase the size of the data.

F. Stemming

Stemming reduces the words to their stems. Stemming is a raw method for removing the suffix from terms which also involves the elimination of derivational affixes. Stemming algorithms are majorly rule-based ones. For instance, they reduce all the-'consulting' words to 'consult'. It is important to note that the above two cleaning processes (Stop words Removal and Stemming) are not done before (sentiment labelling) because doing so would result in irregularities in sentiment detection [10].

G. Count Vectorization

Count Vectorization is the transformation of any given text into a vector based on the frequency count of occurrence of any word in the text. It makes use of two features:

- min_df that defines the minimum frequency of a word to be used as a feature.
- ngram_range which is a tuple. It defines the minimum and maximum length of the sequence of tokens considered. The n-gram is (1,1) so this finds a sequence of 1 token and the min_df is 4.

H. Machine Learning Classification

Classification is the method of constructing a model. Classification falls under the umbrella of supervised learning. A model can be thought of as a mathematical equation that is used to forecast a value by providing it with one or more values. It connects one or more independent variables to one or more dependent variables. The more appropriate the data and the greater the number of dependent variables, the more accurate the model would be. We divided our dataset into training and test datasets in our model by importing sklearn model selection. Sklearn is a Python library that provides data processing functions such as clustering, classification, and model selection. Model selection divides the input data, which can be arrays, lists, or data frames, into random train and test datasets. The training set contains a known output and the model learns from this data to be generalized to other data later on. We have the test dataset (or subset) to test our model's prediction on this subset. We have then used the train test split function to make the split. The test size is 0.2, which means that 20% of the total data is test data, while the remaining 80% is training data. The function's output is stored in the variables X_train, y_train, X_test, and y test. In our dataset, X_train and X_test are the real tweets, and y_train and y_test are the sentiments to which the tweet belongs. Since a classifier can not operate directly on text data, we must first convert it to a vector using the Count Vectorizer function. Thus, X_train and X_test are vectorized and X_train (vector) and y train are fed into the classifiers to train them. Following model training, X_test (vector) is fed into the model, and the model predicts each test data sent as input. Our dataset is divided into three categories: positive, negative, and neutral. As the training data, the following classifier model is trained and fed with known positive, negative, and neutral tweets. After the classifier has been correctly trained, it can be used to detect an unknown tweet and automatically classify it [11]. The Sklearn library is used to import the different classifier models. The following are the different classifier models that we have used:

- Naive Bayes: It is regarded as a generative learning model which is derived from the Bayes Theorem. Different class features are considered independent of one another in this model, regardless of their actual dependence on one another. All of these factors contribute to the probability in their own right. Naive Bayes is useful for analysing large data sets and is easy to implement. Accuracy results that can be obtained are satisfactory given its simple model and the amount of data that can be handled using Naive Bayes as clearly demonstrated in [12].

- Support Vector Machine: It was demonstrated that in SVM, data items are classified based on their position in n-dimensional space with respect to the hyperplane. These observations coordinates are simply converted into support vectors. The support vectors from each class that are the most similar to each other are then chosen. The margin (distance) from all possible hyperplanes is computed for each selected vector. The hyperplane with the greatest margin is the optimal hyperplane that best differentiates the various classes. The classification of new data objects is performed based on the location of these hyperplanes [13].

- Logistic Regression: It is either a classification analytical approach or a predictive learning model. It examines data sets that include one or more dependent variables that influence an outcome. In this case, dichotomous variables, i.e. variables with only two possible outcomes, are used to forecast the result. The best-fitting model that describes the relationship between a set of independent variables and dichotomous characteristic of interest is found using Logistic Regression [14].

- K Nearest Neighbor: KNN is a Lazy Learner classifier since it only uses training data to predict class. The training data is already labelled, so it learns to label new points using similarity measure and distance functions. It identifies the K nearest neighbors based on the distance function. Various distance functions that can be used are Euclidean distance, Manhattan distance, and Minkowski distance [15].

IV. Performance Evaluation of Classifier Model

It is important to know which of the following classification algorithms works better for our dataset, i.e. which one makes the most accurate predictions after training them.

A confusion matrix informs us about true positives, true negatives, false positives and false negatives.

- TP (True Positive): when a case was found to be positive and predicted to be positive.
- TN (True Negative): when a case was negative and predicted to be negative.
- FN (False Negative): when a case was positive but predicted to be negative.
- FP (False Positive): when a case was found to be negative but predicted to be positive.

The classification report is used to assess their prediction quality. Accuracy, Precision, Recall, are the most commonly used performance evaluation metrics.

- Accuracy: It is the number of correct predictions over the total number of instances of data.
- Precision: It is the number of correct positive results over the total number of positive predicted results.
- Recall: It is the number of correct predicted results over the total number of actual positive results.

Statistically accuracy, precision and recall are represented in the below equations.

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

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

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

V. Results

The Twitter data set was used, which was obtained through the Twitter API. Around 8000 tweets were accessed from Twitter for training the classifiers.

TABLE I. TWITTER DATA ANALYZED

Name of Twitter	No. of Tweets
Narendra Modi Timeline	901
Rahul Gandhi Timeline	203
Public contain string @narendramodi	3499
Public contain string @rahulgandhi	3499

A. Narendra Modi Timeline Tweets

We have extracted tweets from Narendra Modi Timeline @narendramodi. All tweets are between 1st January 2021 to 5th April 2021.

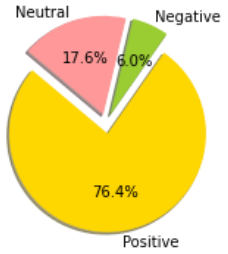

Fig.3. Narendra Modi Timeline tweets sentiment in %

Figure 3 above shows that the majority of tweets or retweets have a positive sentiment and 17.6% are neutral, with just 6.0% having a negative sentiment.

B. Rahul Gandhi Timeline Tweets

We have extracted tweets from Rahul Gandhi Timeline @rahulgandhi. All tweets are between 1st January 2021 to 5th April 2021.

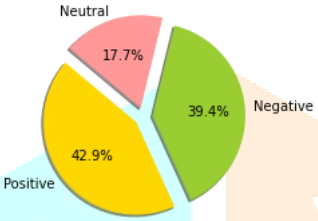

Fig.4. Rahul Gandhi Timeline tweets sentiment in %

Following a review of Rahul Gandhi's Twitter timeline, we discovered that he tweets much more negative sentiment tweets than Narendra Modi. Statistics show that out of 203 tweets and retweets, 42.9% are positive, 17.7% are neutral, and 39.4% are negative as shown in Figure 4.

C. Public containing string @narendramodi

In this portion of the analysis, we are going to analyse those tweets in which people tag Narendra Modi. This part of the analysis will help us to understand what are the people sentiment when they are tagging these leaders. For this part of the analysis we are using a data set that has 3499 entries.

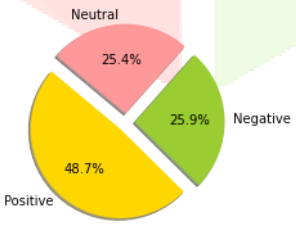

Fig.5. Sentiment of @narendramodi used by people

Figure 5 illustrates the sentiment of the tweets in which @narendramodi were tagged. This analysis shows that for Narendra Modi, 48.7% of tweets have positive sentiment and around 25.9% of negative as well as 25.4% of neutral sentiment.

D. Public containing string @rahulgandhi

In this portion of the analysis, we are going to analyze those tweets in which people tag Rahul Gandhi. For this part of analysis we are using a data set that has 3499 entries.

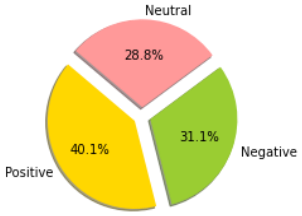

Fig.6. Sentiment of @rahulgandhi used by people

Figure 6 is for @RahulGandhi tagged tweets. And it shows that 31.1% of tweets about Rahul Gandhi are negative, 40.1% are positive, and 28.8% are neutral.

E. Performance Evaluation

TABLE II. ACCURACY OF CLASSIFIERS

Classifiers	Accuracy
Naive Bayes	66.78
Logistic Regression	77.91
Support Vector Machine	78.09
K Nearest Neighbor	41.76

As shown in Table 2, the accuracy of the Naive Bayes algorithm is 66.78%, the accuracy of Logistic Regression is 77.91%, the accuracy of Support Vector Machine is 78.09%. We made our final prediction utilizing SVM since the accuracy of the algorithm is higher than the rest.

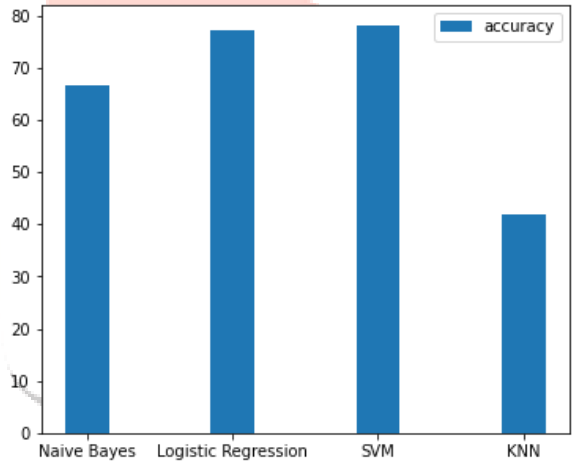

Fig.7. Performance of the Classifiers

TABLE III. PRECISION RECALL OF CLASSIFIERS

Classifiers	Precision	Recall
Naive Bayes	0.66	0.67
Logistic Regression	0.78	0.78
Support Vector Machine	0.79	0.78
K Nearest Neighbor	0.58	0.42

As shown in Table 2 and Table 3, the accuracy of SVM is 78.09%. Its weighted average recall is 0.78 while precision is 0.78. That is, the number of false positives and false negatives predicted by the classifier model is nearly equal. On the other hand, KNN has an accuracy of 41.76%. It has high precision and a low recall, which means that the classifier's accuracy of positive predictions or the ability of a classifier not to mark an instance positive that is actually negative is fine, but its ability to identify all positive instances is poor. As a result, KNN will have more false negatives than false positives. Hence, SVM is the most preferred model.

VI. Conclusion

The Support Vector Machine was found to be more reliable than the other classifiers used. Support Vector Machine has a faster prediction time than other classifiers because it is an Eager Learner. More tweets can be accessed to improve the accuracy of the following classifiers. The model's accuracy has currently been tested by accessing around 8000 tweets, but more could be accessed in the future. So, by analysing the sentiment of those tweets, we have determined which leader is more popular than the other. This sentiment analysis results suggest that Narendra Modi has a better positive opinion in public as compared to Rahul Gandhi. Since VADER already has a lexicon of slang terms and emoticons, it is easier to use it for review of any social media database. Automated Sentiment Analysis is one of the quickest approaches an agency or government can take for a deeper understanding of the overall feelings of any broad population against policies, actions, or reforms that it wishes to introduce or pass in the near future.

References

[1] Sourav Das, Anup Kumar Kolya. "Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm" International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2017.

[2] Pooja Dhede, Samruddhi Hagone, Gaurav Gaikwad, Gaurav Chaudhari, "Analyzing Awareness of Government Scheme using Swachh Bharat Tweets", VJER-Vishwakarma Journal of Engineering Research, 2019.

[3] P. Singh, K. Singh Kahlon, R. Singh Sawhney, "Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government", 2018.

[4] Amolik, Akshay, Niketan, Jivane, Bhandari, Mahavir & Venkatesan, "Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques", 2016.

[5] N. Godbole, M. Srinivasaiah and S. Skiena, "Large-Scale Sentiment Analysis for News and Blogs", 2007.

[6] Saif, Hassan & Alani, Harith, "Semantic Sentiment Analysis of Twitter", 2012.

[7] Neethu, M.S., Rajasree, "Sentiment analysis in twitter using machine learning techniques", International Conference on Computing, Communications and Networking Technologies, 2013

[8] Suresh, Annamalai & Bharathi,"Sentiment Classification using Decision Tree Based Feature Selection". International Journal of Control Theory and Applications, 2016.

[9] V.K. Chauhan, Dr. Amita Goel & A. Bansal, "Twitter Sentiment Analysis Using Vader", International Journal of Advance Research, Ideas and Innovations in Technology, 2018.

[10] A. Alsaeedi, M. K. Khan "A Study on Sentiment Analysis Techniques of Twitter Data" in (IJACSA) International Journal of Advanced Computer Science and Applications, 2019.

[11] Dr. A.Senthilrajan, P.B Matharasi, "Sentiment Analysis of Twitter Data using Naïve Bayes with Unigram Approach", International Journal of Scientific and Research Publications, 2017.

[12] K. Suppala, N. Rao, "Sentiment Analysis Using Naive Bayes Classifier", International Journal of Innovative Technology and Exploring Engineer- ing (IJITEE), 2019.

[13] Mullen, Tony & Collier, Nigel, "Sentiment Analysis using Support Vector Machines with Diverse Information Sources" (2004).

[14] A. Tyagi, N. Sharma, "Sentiment Analysis using Logistic Regression and Effective Word Score Heuristic" International Journal of Engineering & Technology, 2018

[15] Dey, Lopamudra, Sanjay, Biswas, & Bose, "Sentiment Analysis of Review Datasets Using Na¨ıve Bayes and K-NN Classifier", International Journal of Information Engineering and Electronic Business", 2004