IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Real Time Twitter Sentiment Analysis

Mr. Samarth Shete
Dept. Of Computer Science & Engineering
ATS's, SBGI, Miraj
India

Ms. Lahasya Gowda
Dept. Of Computer Science & Engineering
ATS's, SBGI, Miraj

Ms. Akshada Lattimardi
Dept. Of Computer Science & Engineering
ATS's, SBGI, Miraj
India

Prof. Kartik Nikam
Dept. Of Computer Science & Engineering
ATS's, SBGI, Miraj
India

Ms. Mubashirin Pinitod
Dept. Of Computer Science & Engineering
ATS's, SBGI, Miraj
India

***Abstract*— Sentiment analysis of social media data includes attitudes, evaluations, and emotions, which can be thought of as a way of thinking. It is extremely tough to comprehend and classify the enormous collection of papers into good and bad elements. Twitter, Facebook, and Instagram, for example, give a forum for gathering information on people's feelings and opinions. Social media platforms are increasingly being used by businesses to provide various services and communicate with customers. Attitude Analysis (SA) divides the polarity of a set of tweets into positive and negative categories in order to better understand public sentiment. The goal of this work is to do sentiment analysis on real-time 2019 election twitter data using the feature selection model word2vec and the sentiment classification technique random forest. When compared to traditional methods like BOW and TF-IDF, Word2vec plus Random Forest enhances sentiment analysis accuracy dramatically.**

***Keyword*—** Sentiment Analysis; Word2Vec; Random Forest; Twitter data analysis; TF-IDF; BOW;

INTRODUCTION

In our modern digital era, social media has exploded in popularity, and it now plays a vital role in everyone's daily routine. Social media data is vast and unstructured, containing real-time public opinion and sentiments in a variety of formats and languages. Twitter is a microblogging website that serves as a prime location for people to communicate their opinions and views on politicians and political parties. Any breaking news or breaking events are almost rapidly followed on Twitter, resulting in a spike in volume and a unique opportunity to discover the substance of the relation between election events and popular emotion Twitter has 326 million monthly active users and is accessible via SMS, mobile devices, and the internet interface. Furthermore, mobile devices are used by 80% of its existing consumers. Users of microblogging services, such as Twitter, are prone to spelling errors when typing a tweet, as well as using emoticons to express their thoughts and emotions. Natural language processing (NLP) also plays a significant role and can be utilized to convey one's views. With internet connectivity, many social media networking sites are able to provide massive amounts of information on a variety of topics in real-time from any location, at any time, and from any place. On average, 6000 tweets are sent every second, equating to 500 million tweets every day and 350,000 tweets per minute. As a result, Twitter provides a significant supply of real-time information and data about recent trends, people's opinions, and sentiments, which may be employed in data and text analytical studies to gain useful insights.

Sentiment analysis, often known as opinion mining, is the process of extracting sentiments, opinions, and emotions from textual data. It's one of the most important fields of natural language processing study. The fundamental goal of sentiment analysis is to categorize data into positive or negative polarity in order to determine public or given data sentiments [3]. Real-time sentiment analysis could have important implications for many different areas of the, including politics, government, organizations, [2] elections, and business. Because they can act quickly and help profit by making necessary actions and decisions [3].

Random Forest is a powerful machine learning method that can do classification and regression problems. Random Forest is an ensemble learning model that combines a few weak models to create a powerful one. In Random Forest, instead of a single decision tree, several trees are grown.

BACKGROUND

The emergence of social media data, which contains rich, large, rapidly-emerging unstructured information, has offered an opportunity to research public opinion and learn about people's feelings. The science and business communities [1] are increasingly interested in gathering information about social events, company plans, political movements, marketing activities, and product preferences, among other topics. Opinion mining, also known as sentiment analysis, is a new field that has emerged as a result of this.

On social media, trending topics reveal people's interests, intentions, and, most crucially, recent activity all over the world. Interestingly, the topics that are currently popular on social media may or may not be popular after an hour, and may or may not be popular an hour ago.

A. Literature Survey

Turney et.al. [1] applied a bag-of-words method for sentiment analysis, in which the connections between words were ignored and a sentence was simply treated as a collection of words. To estimate the sentiment of the entire sentence, the sentiment of each individual word was determined separately and then aggregated using aggregation methods.

Po-Wei Liang et al. [7] collected data from Twitter using the Twitter API. Opinion-related tweets were filtered out. For polarity recognition, the Unigram Naive Bayes model was created. They also used the Mutual Information and Chi Square feature extraction methods to eliminate unnecessary features.

In their paper [5], Geetika Gautam and Divakar Yadav looked at how to classify customer reviews using opinion mining. Analyzed unstructured tweets to extract negative, positive, or neutral sentiment. Naive Bayes, Maximum Entropy, SVM, and Semantic Analysis are the machine learning techniques employed in the paper [5]. (WordNet). The accuracy of different machine learning methods is 88.2% for Nave Bayes, 83.8 percent for Maximum Entropy, 85.5 percent for Support Vector Machine (SVM), and 89.9% for Semantic Analysis (WordNet)

In articles [6] [7], real-time Tweeter data was used to examine public attitudes toward two national political parties in India. In this work, a lexicon-based strategy for sentiment calcification is used, with eight different types of emotions such as glad, sad, and joy represented.

III. PROPOSED MODEL

In the previous topic, we looked at the different methods used by the author, from random forests to neural networks. We have now looked at the existing model used by the existing authors. The support vector machine is a machine learning algorithm. It uses a random forest so you can compare the result with other existing results as they are no longer in use.

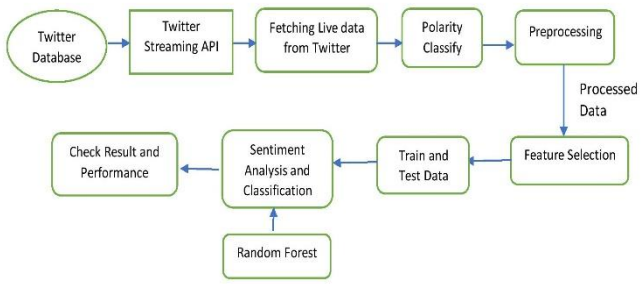

Fig. 1 Flowchart of Proposed Model

The steps for the proposed system that we are implementing in this project are listed below, along with some explanations for each step.

A. Obtaining Twitter API keys.

To access the Twitter API, a user must first create a Twitter developer account in order to get credentials such as the API key, API secret, Access token, and Access token Secret. The following are the steps involved:

Step 1: If the user does not already have a Developer Account, he or she must go to https://developer.twitter.com/ and set up one.

Step 2: Go to https://developer.twitter.com/en/apps and log in with the Twitter developer credentials.

Step 3: The user must build an application by selecting the option "Create an App," filling out the form with the necessary application information, and clicking the "Create" button.

Step 4: Log in with the Twitter developer credentials you created in Step 1 at https://developer.twitter.com/en/apps.

Step53: The user must create an app by selecting "Create an App," filling out the form with the required app information, and hitting the "Create" button.

B. Twitter Data Extraction

• Successful connection formation – In order to access the Twitter API, the user must first create a Twitter Developer Account and then receive the appropriate credentials, which include the API Key, API Secret Key, Access token, and Access token secret.

• By Installing the Twitter library – The user can connect to the Twitter API and then retrieve tweets directly from Twitter using the Twitter API by installing the appropriate libraries.

There are numerous libraries available, and most programming languages support them.

• Creating a secure connection between the programming language and Twitter - In order to extract tweets, you must first create a secure connection between the programming language and Twitter. The user will be taken to Twitter's authorization page. Click on the authorized application and note the PIN that appears. Return to the programming language and type in your PIN. It's important to note that this only needs to be done once.

As a result, you'll be able to successfully access Twitter API and extract tweets.

C. Classification of Polarity

One of the most important functions of sentiment analysis is polarity classification. In recent years, sentiment analysis and opinion mining have gotten a lot of interest in natural language processing and data science research. Most earlier sentiment analysis approaches concentrated on the subjective aspects of the text, such as considering word sentiment rather than the context in which the word appears.

D. Removing Usernames, stop-words, punctuations, symbols

To use the dataset in the right format, we will preprocess it by eliminating user names, stop-words, punctuations, symbols, and so on. We will also use methods like stemming, lemmatization, and tokenization, which are NLTK techniques. E. Features extraction methods These methods (Bag of Words and TFIDF) are used to extract features from the text and split the dataset into training and test datasets.

F. Modeling-Random Forest Classifier

After preprocessing the data in the right manner, we will build the model using Random Forest Classifier in this

phase. A random forest classifier is defined as multiple decision trees acting as multiple machines learning algorithms, achieving better prediction performance than using a single decision tree algorithm acting as a predictor.

The classes at the class prediction levels form the individual votes. We get the most votes in the results of generating O / P model predictions for random forests. The

Random Forest's operation is as follows: Each tree is implanted and developed as follows:

Step 1: create a training set from the data. Let's say there are N examples in the training set. With replacement, a random sample of N cases is taken. This set of random samples is used as a training set for tree growth.

Step 2: If D input variables are present, and each node has d<D variables, d random variables are chosen from D. To split the node, the best split on these d variables is utilized. The value of d is kept constant while the forest grows.

Step 3: Each tree is allowed to grow to its full potential without being pruned.

Step 4: Combining the predictions of the n trees that were used to forecast fresh data. (For classification, the majority vote is required; for regression, the average vote is required).

IV. ANALYSIS AND RESULT

A. Dataset

There are 18685 tweets in the Train set and 17,197 tweets in the Test set in this dataset. The training data was divided into two categories depending on the polarity of tweets: negative and positive, with 0 and 1 assigned to each. There are 12890 favorable tweets and 5795 negative tweets, according to the data.

Since 2019, the tweets used for training have been collected using the Twitter developer API with Hashtags linked to Indian politics, ignoring retweets. The unclassified set of tweets used for testing purposes to fit into the Machine Learning model is referred to as the testing data.

B. Evaluation

We will utilize the accuracy score and f1 score after fitting the model with Random Forest Classifier and testing for prediction.

C. Result

1. Generating Word Cloud

The importance of each word in the document is represented by a word cloud. We will build word cloud for Positive and Negative Tweets. As demonstrated in Fig. 3 and Fig. 4, the size and color of the word signify how frequently a certain term is used in the document.

Fig. 2 Word Cloud of Positive Tweets for Narendra Modi

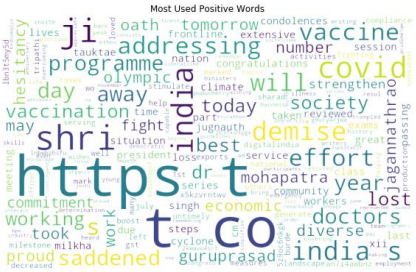

Fig. 3 Word Cloud of Negative Tweets for Narendra Modi

2. Pie chart Analysis

Fig 6 presents Pie Chart representation of positive and negative classification of tweets on various topics related to Narendra Modi.

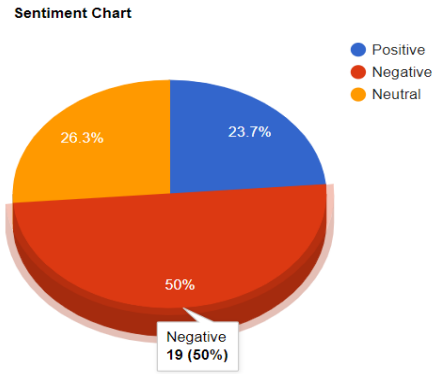

Fig. 4 Analysis of Tweets for Narendra Modi.

V. CONCLUSION

The project's implementation finds that using the Random Forest Algorithm to create and test the model, the f1 score is 0.66354. This demonstrates that Random Forest's Hindi Tweets dataset has a f1 score of 0.66354 and an accuracy score of 90.24. As a result of these findings, we can deduce that: This score of 0.6635 is somewhat similar to the results obtained by existing authors who assigned in the literature search.

If we look at the results, we can see that the authors mostly used SVM (Support Vector Machine) and mostly Neural Networks, which prompted us to change our method to only use the Random Forest Classifier method and then use the Hindi Language Dataset. After we got the score, we compared it to the results of previous authors, and we found that our score was mostly similar to theirs, with results ranging from 70% to 80%.

REFERENCES

[1] Alexander Pak, Patrick Paroubek. 2010, Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

[2] Wang, H., Can, D., Kazem Zadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle.

[3] Liu, B. (2015) “Preface,” in Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University.

[4] Alec Go, Richa Bhayani, Lei Huang. Twitter Sentiment Classification using Distant Supervision.

[5] Gautam, G. and 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).

[6] Kuamri, S. and Babu, C.N., 2017, July. Real time analysis of social media data to understand people emotions towards national parties. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[7] Po-Wei Liang, Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE 14th International Conference on Mobile Data Management, Milan, Italy.