Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model

Hitesh, MSR., Vaibhav, V., Kalki, Y. J. A., Kamtam, S. H. & Kumari, S

Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model

MSR Hitesh, Vedhosi Vaibhav, Y.J Abhishek Kalki, Suraj Harsha Kamtam,

Santoshi Kumari, Narendra Babu C.

Department of Computer Science and Engineering
Ramaiah University of Applied Sciences
Bengaluru, India

Abstract – **Sentiment analysis of social media data consists of attitudes, assessments, and emotions which can be considered a way human think. Understanding and classifying the large collection of documents into positive and negative aspects are a very difficult task. Social networks such as Twitter, Facebook, and Instagram provide a platform in order to gather information about people's sentiments and opinions. Considering the fact that people spend hours daily on social media and share their opinion on various different topics helps us analyze sentiments better. More and more companies are using social media tools to provide various services and interact with customers. Sentiment Analysis (SA) classifies the polarity of given tweets to positive and negative tweets in order to understand the sentiments of the public. This paper aims to perform sentiment analysis of real-time 2019 election twitter data using the feature selection model word2vec and the machine learning algorithm random forest for sentiment classification. Word2vec with Random Forest improves the accuracy of sentiment analysis significantly compared to traditional methods such as BOW and TF-IDF. Word2vec improves the quality of features by considering contextual semantics of words in a text hence improving the accuracy of machine learning and sentiment analysis.**

Index Terms – Sentiment Analysis; Word2Vec; Random Forest; Twitter data analysis; TF-IDF; BOW;

I. INTRODUCTION

Social media nowadays is at a boom in recent times and plays important role in everyone's day to day routine in this modern digital era. Data generated from social media contains real time opinion and sentiments of public in various formats and languages that is huge and unstructured. Twitter provides a microblogging service is a central site where people tend to express their opinions and views on candidates as well as the political parties. Any news or emerging events are almost instantly followed thereby causing a burst in Twitter volume and providing a unique opportunity to find the essence the relationship between electoral events and public sentiment. Twitter provides a platform for the users to deliver, interpret and share 280 characters post which is known as a tweet. Twitter currently has 326M monthly active users and thereby is accessible through SMS, mobile devices and website interface. Moreover, 80% of its current users are active through mobiles. The microblogging service users like Twitter users tend to make spelling mistakes by typing the tweet and the fact that they try to use emoticons for expressing their views and emotions. The natural language processing (NLP) is also playing a major role and can be used to express their opinions.

Having internet access, many social media networking sites are providing enormous information on various topics, at real-time from any-place, any-time, and anywhere. On an average 6000 tweets are generated every second that corresponds to 500 million tweets in a day and 350,000 tweets in a minute. Hence twitter provides huge source of information and data on recent trends, people opinions, sentiment at real time which can be used of data analytical and text analytical research and obtain valuable insights from it.

Identifying sentiments, opinion and emotions from textual information is known as Sentiment analysis and also named as opinion mining. It is one of the major research areas in natural language processing. The main objective of the sentiment analysis is to classify the data into positive or negative polarity in order to identify the sentiments of the public or given data [3]. This analysis is applied in several fields such as Fraud detection, Healthcare, Finance, stock market, selling and purchasing items and several other business organizations to improve client reach and sales, brand building and many more. Real-time sentiment analysis can have a huge impact on various areas like politics, government and organization, elections[2], and businesses as they can quickly act on it and helps them gain benefits by taking necessary actions or decisions [3].

Word2Vec is a set of the unsupervised shallow two-layer neural network model that produces word embedding. Numeric description of word in the form of vector is known as Word embedding. Word2Vec considers contextual semantics of words to produce word embedding i.e., instead of focusing on single word and two or three word it considers the context in which the word is occurring. Similar words with same or relative context are mathematically clustered together into vector space. That further conserves the semantic relationship between words. Hence word embedding produced using Word2Vec, can be used to train machine learning and classification algorithm to improve sentiment calcification accuracy base on context and semantic relation between words.

Random Forest is one of a versatile machine learning algorithm capable of performing both classification and regression tasks. Random Forest is an ensemble learning model, where a few weak models combine to form a powerful model.

In Random Forest, one grows multiple trees as opposed to a single decision tree.

II. BACKGROUND

The explosion in social media data holding valuable, vast, rapidly-emerging unstructured information has created an opportunity to study public opinion and to know the sentiments of the people. Capturing the opinions about social events, company strategies, political movements, marketing campaigns, and product preferences, etc. brings in growing interest from the science and business communities [1] to improve their performances. This has resulted in an emerging field of Opinion mining or Sentiment Analysis.

Adaption to change quickly is very crucial in this ever-changing world for taking any actions or decisions faster. Faster the data is available, faster decision or an action should be taken and, in some cases, it has even necessary to prevent a loss of lives and save the lives. Such as, analysis of twitter data is used for real-time event detections such as earthquakes in Japan, attacks in Paris, tsunami alerts, etc. It has also played an important role in US presidential elections in the year 2016 based on public opinions [4].

Trending topics on social media expose people's interests, their intentions and most importantly recent activities throughout the world. Interestingly, currently trending topic on social media may not be trending after one hour and might not be trending an hour ago.

A. Literature Survey

According to Efthymios Kouloumpis, Theresa Wilson [4], methods and methodologies used in this paper are N-grams and Lexicon features: PoS. In paper [4], they have examined the use of linguistic features for identifying the sentiment from the tweets and evaluated the use of existing lexicons as well as the features that capture information about the creative and informal language used in twitter. Having taken a supervised approach to solve the problem, there is always leverage for the existing hashtags in the tweets for building training data and concluded that that part-of-speech features may not be suitable for sentiment analysis in the twitter domain. However existing sentiment lexicon features areuseful in conjunction with twitter providing microblogging features.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, Zhenyu Yang in paper [5], they have projected a novel technique that combines the sentiment analysis and context-based topic modeling to analyze the opinions from the people on the collected microblogging site data. The data they have used is Australian Federal Election 2010 twitter data. The results show that the topic modeling can reveal unseen topics and communities of tweets by grouping twitter data into clusters, and further topic modeling on a group of tweets also uncover an underlying pattern deeply hidden in that collection of tweets.

Real time Tweeter data is analysis performed in papers [11] [13] to analyze the sentiment of the people towards two national political parties of India. In this paper lexicon based approach is for sentiment calcification where eight different kinds of emotions are represented such as happy, sad, joy.

Neethu M.S, Rajasree R [6], they have analyzed the twitter posts about electronic products like mobiles, laptops, etc using Machine Learning approach. They have also presented a new feature vector for classifying the tweets as positive, negative and extract people's opinions about products. Classification accuracy of vector is being tested using different algorithms such as SVM, Maximum Entropy, Ensemble Classifiers and Naive Bayes. Among all these classifiers Naïve Bayes has better precision however slightly lower accuracy and recall. But almost similar performance for all the classifiers is obseverd.

Geetika Gautam, Divakar Yadav, in paper [7] worked on classification of customer reviews based on opinion mining. Analyzed the unstructured tweets to mine opinion in the form negative or positive, or neutral. Machine learning algorithms used in paper [7] are Naive Bayes, Maximum entropy, SVM and Semantic analysis (WordNet). The accuracy results for different machine learning algorithms are 88.2% for Naïve Bayes, 83.8% for Maximum entropy, 85.5% for Support Vector Machine (SVM) and 89.9% for Semantic Analysis (WordNet).

André L. F. Alves, Cláudio de S. Baptista, Anderson A. Firmino, Maxwell G. de Oliveira, Anselmo C. de Paiva, paper [8] where two sentiment analysis techniques: Naive-Bayes vs SVM classifiers are applied on a case study and compared. The data considered are tweets in Portuguese language during the 2013 FIFA Confederations Cup. The achieved results indicated SVM technique outdid the Naive-Bayes one, concerning performance issues. The results obtained by the Naive-Bayes sentiment classifier presented an accuracy of 72.7% and F-measure of 0.791 and SVM sentiments classifier results shows an accuracy of 80.0% and F-Measure of 0.873 and for the detection of sentiment polarity.

Andi Rexha, Mark Kroll, Mauro Dragoni, Roman Kern, paper [9] they have presented a paper where automatic prediction of polarity is done in a tweet using Word2Vec, and the Semeval 2016 Task #4 dataset is used for the evaluation. In the classification setting, they use only the semantic information obtained from a Word2Vec model trained on the tweets and they do not have any polarity information but to evaluate the feature representation the well-known classification algorithms such as the Naive Bayes and Support Vector Machine. The approach achieves F1-measures of up to ~54 % for the negative class and ~90 % for the positive class without using polarity information about single words.

In paper [12] Qufei Chen and Marina Sokolova applied two popular word embedding model word2vec and doc2vec to perform sentiment analysis of clinical discharge summaries using unsupervised method.
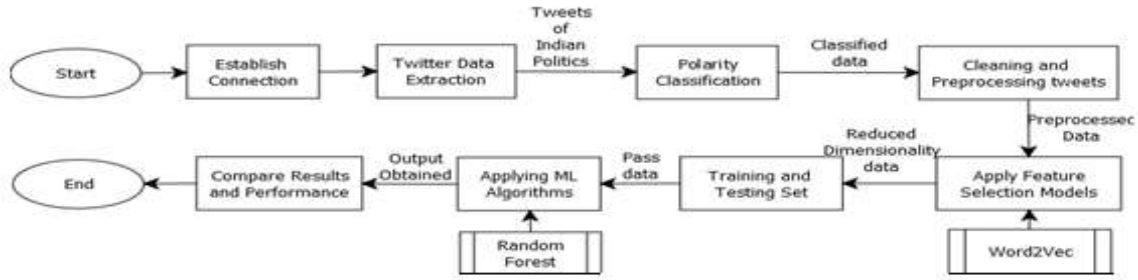
Fig. 1 Flowchart of the Proposed System

III. PROPOSED SYSTEM

The proposed system for the whole process starting from collecting tweets until getting the desired output can be seen in the flowchart which can be seen in Figure 1 below. Each of the following steps is explained in the following:

A. Establishing Connection
1. Getting Twitter API keys

Initially, a user needs to create an account of the Twitter developer to obtain credentials such as API key, API secret, Access token and Access token Secret required to access the Twitter API. The Steps involved are as follows:

Step 1: User needs to visit https://developer.twitter.com/ to create an account if he/she doesn't have a Developer Account.

Step 2: User needs to visit the following URL: https://developer.twitter.com/en/apps and login with Twitter developer credentials created in Step 1.

Step 3: User needs to create an Application by selecting an option "Create an App" by filling the form with the required Application details and select the "Create" button.

Step 4: Then in the next page user needs to click on the "Keys and Access Tokens" tab, then under the Consumer API keys section the details of API Key and API secret are displayed which user can copy for use in the program for accessing the Application for collecting the tweets.

Step 5: In the same "Keys and Access Token" tab the user needs to visit the Access token and Access token secret section then click and select the "Create" button to generate the Access token and Access token secret. The user can copy them for accessing the Twitter application for collecting the tweets.

B. Twitter Data Extraction
- Successful connection establishment – User needs to obtain the Twitter Developer Account and then can obtain the

required credentials which are API Key, API secret key, Access token and the Access token secret in order to access the Twitter API.
- Installing Twitter library – By using the required libraries the user can connect to Twitter API and then download the tweets directly from the Twitter through the Twitter API. There are multiple libraries available and supported by most of the programming languages.
- Establishing a connection on Programming Language - In order to extract tweets, one needs to establish a secure connection between the programming language and Twitter. One will be directed to Twitter's authorization screen. Click on Authorize App and note the PIN generated. Go back to the programming language and enter the PIN. Note, this only needs to be done once. Thereby, can successfully access Twitter API and extract tweets.
- Extracted tweets using the searchTwitter function and collect tweets in English without Retweets on the term Indian Elections.

C. Polarity Classification
Classification of polarity is one of the key role of sentiment analysis. Opinion mining and sentiment analysis have attracted increasing attention in natural language processing and data science research in recent years. Most of the previous sentiment analysis approaches mainly focused on subjective part of the text such that considering word sentiment instead of context in which word is present. [10] it is essential to classify polarity based on context in which word is present and semantic relationship between the words.

D. Cleaning and Preprocessing data
Data preprocessing process can be seen in Figure 2 which is a flowchart that shows the flow of the data in order for the data to preprocess and the steps followed are as shown below to clean the raw tweets in data.
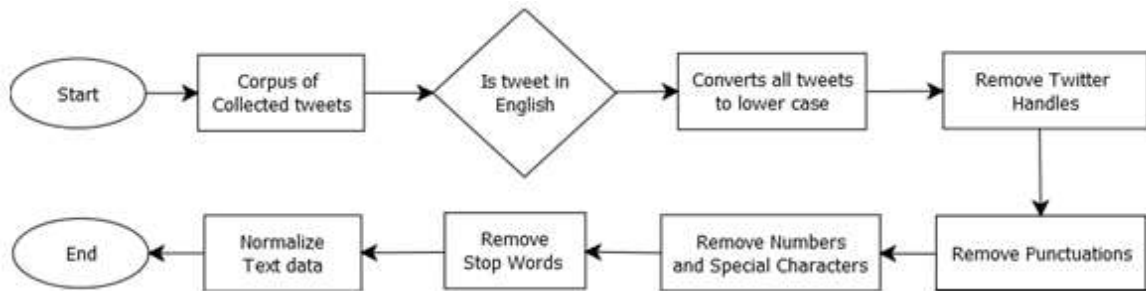

Fig. 2 Flowchart of the Data Cleaning and Preprocessing

Step 1: Removal of Twitter Handles involves removing the "@user", it is a masked term by Twitter to prevent any privacy issues these need to be removed as they don't give any relevant information about the nature of the Tweet.

Step 2: Removal of Numbers, Punctuations and Special Characters since they don't play any significant role in differentiating the types of Tweets.

Step 3: The Small words and the Stop words used in the tweets do not add significant value for the analysis. Words such as "His", "All". So, these kinds of words need to be removed from the Tweet Data.

Step 4: It involves normalizing the Text data. For example, reducing terms like loves, loving, and lovable to their base word, i.e., 'love' is often used in the same context. Normalizing the text helps in reducing the total number of unique words in the tweets data without losing any important/relevant information.

The detailed explanation for the above steps is as follows:

1. Removing Twitter Handles (@user)

User needs to create a column to store the cleaned and processed tweets. To remove the Twitter handles a regular expression pattern is passed which is "@[]*" so any word starting with "@" is removed.

2. Removing Punctuations, Numbers, and Special Characters

User needs to replace the everything else except the characters and hashtags with spaces. The regular expression "[^a-zA-Z#]" means anything except alphabets and '#' thereby removing the punctuations,

numbers and special characters.

3. Removing Small and Stop Words

User needs to be cautious while selecting the length of the words to be removed. So, according to politics, the length chosen is 3 so all the words having length less than it will be removed. So, the terms such as "Ha", "Oh" are having no use as they give no relevant information in the tweets they need to be removed.

4. Text Normalization

Here using Stemming function one can normalize the tweets. But before that tokenization of the tweets needs to be done. Tokens in terms of NLP can be defined as the individual words/terms, and Tokenization can be defined as process which involves splitting a string of text into tokens.

E. Applying Feature Selection Model
● Word2Vec Feature Selection Model

Word embedding is a way used for depicting the words in the form of vectors. The aim of word embedding is to redefine the high- dimensional word characteristics into low-dimensional feature vectors by preserving the corpus contextual similarity.

The advantages of using word embedding's over Bag of words (BOW) or TF-IDF are:

1. Dimensionality reduction - a significant reduction in the no. of features required to build a model.
2. It captures the meanings of the words, semantic relationships and the different types of contexts the words are utilized in sentences.

● Word2Vec Embedding's

Word2Vec algorithm is a combination of 2 Techniques namely continuous bag of words (CBOW) and Skip-gram model. Both the techniques are shallow neural networks used for mapping a word/words to a target variable which can also be a word or a set of words. Also, these techniques learn weights of words which are represented in the form of word vector representations.

There are two types of Word2Vec model designs are used:

1. Continuous Bag-of-Words model (CBOW)
2. Skip-Gram model

CBOW tends to predict the likelihood of a word given context it is being utilized in and the context can be a single adjacent word or a group of surrounding words. The Skip-gram model works in a reverse manner, it tries to predict the context for a given word.

The Skip-gram model is being used and has the following advantages:

Advantage 1: Skip-gram captures 2 semantics for a word that is it will have 2 vector representations of 'Apple'. One can be representing the fruit and the other representing the technology company.

Advantage 2: Generally, the Skip-gram with negative sub-sampling outperforms Continuous bag of words.

F. Training and Testing Set

Training a Word2Vec model on our data is very important in order to obtain vector representations for all the unique words present in our corpus. There is one more option of using pre-trained word vectors instead of training our own model. However, in this paper, we will be training our own vectors since the size of the pre-trained word vectors is generally huge.

One can see from the training data that the Word2Vec model does a good job of finding the most similar words for a given word. But how is it able to do so? That's because it has learned vectors for every unique word in our data and it uses cosine similarity to find out the most similar vectors.

● Preparing Vectors for Tweets

Since our data contains tweets and not just words, one would have to figure out a way to use the word vectors from the Word2Vec model to create vector representation for an entire tweet. There is a simple solution to this problem, that is by simply taking the mean of all the word vectors present in the tweet. The length of the resultant vector will be the same, i.e. 200. Thereby, repeating the same process for all the tweets in our data and obtain their vectors. Now that one has 200 word2vec features for our data.

G. Applying Machine Learning Algorithms
● Random Forest

Random Forest is an algorithm for machine learning that can perform both regression and classification tasks. It's a kind of learning method that combines a few weak models to form a powerful model. In Random Forest, in contrast to a single tree decision, one grows multiple trees.

Working of the Random Forest is as follows where each tree is implanted & developed as follows:

Step 1: data set as training set. Let us consider N number of cases in training set. Random sample of of N cases are taken with replacement. This random samples are considered as training set for growing tree.

Step 2: If D input variables are there, and number of variable d<D are at each node, d random variables are selected out of D. The best split on these d variables is used to split the node.

While growing the forest, value of d is kept constant.

Step 3: Each tree is grown to the largest extent possible and there is no pruning.

Step 4: Aggregating the prediction of n trees used to predict new data. (i.e., majority votes for classification, average for regression).

IV. ANALYSIS AND RESULT

A. Dataset

The dataset has a Train set of 18685 tweets and the test set has 17,197 tweets. In the training data, the data has been classified based on the polarity of tweets into negative and positive then they were assigned 0 and 1 respectively. It has been seen that there are 12890 are positive tweets and 5795 are negative tweets. The tweets used for training have been gathered through twitter developer API using the Hashtags related to Indian politics since 2019 excluding the retweets. The testing data is the unclassified set of tweets used for testing purposes to fit into the Machine Learning model.

B. Evaluation Metrics

F1 score is being used as the evaluation metric. It is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is suitable for uneven class distribution problems.

The important components of F1 score are:

- True Positives (TP) - These are the correctly predicted positive values which means that the value of the actual class is yes and the value of the predicted class is also yes.
- True Negatives (TN) - These are the correctly predicted negative values which mean that the value of the actual class is no and the value of the predicted class is also no.
- False Positives (FP) – When actual class is no and predicted class is yes.
- False Negatives (FN) – When actual class is yes but predicted class in no.

$$Precision = TP/TP+FP$$

$$Recall = TP/TP+FN$$

$$F1\ Score = 2(Recall\ Precision) / (Recall + Precision)$$

C. Experimentation Results

1. Construction of Word Cloud

A word cloud represents importance of word in the document. Where size and color of the word represents how frequently a particular word is being used in the document as shown in Fig. 3 and Fig. 4.

Fig. 3 Word Cloud of Positive Tweets for 2019 Indian Elections

Fig. 4 Word Cloud of Negative Tweets 2019 Indian Elections

The words that people used more frequently are lok sabha, elect, Indian, BJP, Congress, namo, rahul, gandhi, and modi. Fig 3 and fig 4 represents people opinion and sentiment are related to 2019 elections on various leading topics like terrorism, arm force, loksabha elections, and vote. Discussion on parties like BJP and Congress and political leaders like Narendra Modi and Rahul Gandhi.

2. Bar Plot Analysis

The Training set of 18685 tweets and the test set of 17,197 tweets used for training random forest and analyzing the sentiment based on polarity classification onto positive and negative tweets for the 2019 Indian elections. Fig 5 and Fig 6 presents the bar plot representation of positive and negative classification of tweets on various topics related to Indian Loksabha Elections 2019. The horizontal axis represents ten different Hashtag count and the vertical axis represents the total sentiment counts.
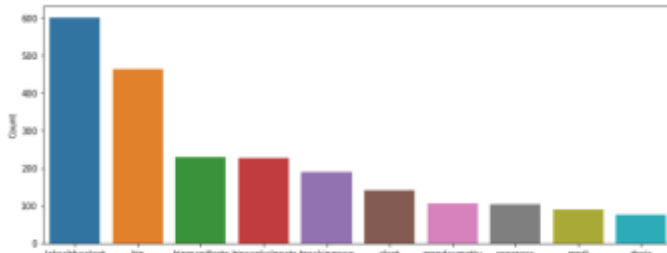
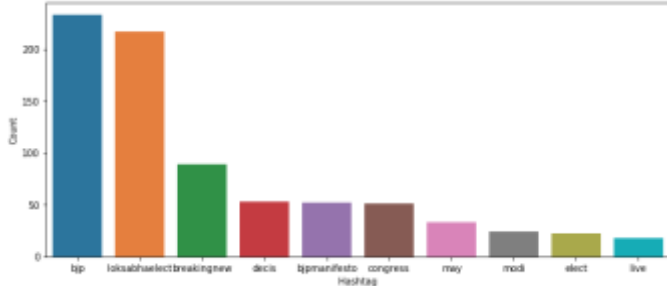*Fig. 5 Bar Plot for Positive Polarity Tweets of 2019 Indian Election*

Fig. 6 Bar Plot for Negative Polarity tweets of 2019 Indian Election

Finally data visualization plays important role is summarizing the results in attractive format that is easy to understand and explain. In table 1 word2vec feature selection model is compared with two tradition method Bag-of-Word and TF-IDF for selecting important feature from tweets and training them to random forest machine learning model for sentiment classification. However word2vec gives highest accuracy of 86.87% compare to BOW and TF-IDF.

Model	Vector-Space		
	Bag-of-Words	**TF-IDF**	**Word2Vec**
Random Forest	83.49%	84.49%	86.87%

Table 1 Comparing Accuracies of different Models

Accuracy of the Word2vec model is more as it uses word embedding for the understanding context of tweets. This shows the importance of word embedding in solving Natural language processing (NLP) problems.

V. CONCLUSION

Sentiment analysis (also known as opinion mining) is one of the many applications of Natural Language Processing. A set of methods and techniques used for extracting subjective information from text or speech, such as opinions or attitudes. In simple terms, it involves classifying a piece of text as positive and negative.

This paper introduces the combination of the Word2Vec and Random forest model for performing sentiment analysis of 2019 election twitter data. Initially, created a twitter developer API and using the details of the API such as Consumer Key (API Key), Consumer Secret (API Secret), Access Token, Access Token Secret extracted tweets using the searchTwitter function and collected 18000 tweets in English without

Retweets on the term Indian Elections and classified data into positive and negative tweets using the inbuilt function. Then performed data cleaning and data pre-processing on the classified data. Later, applied the Word2Vec feature selection model for extracting features from cleaned tweets. Finally, used these feature sets to build a model for sentiment analysis using the Random Forest machine learning algorithm. Our approach achieves 83.4% for Bag-of-Words, 84.4% for TF-IDF and 86.8% for Word2Vec. This shows the importance of word embedding in dealing with NLP problems, especially in sentiment analysis. Further, Parts-of-Speech tagging, word2vec with various combinations of machine learning algorithms can be analyzed in future work to improve the accuracy of sentiment analysis for large scale real-time social media data on various platforms.

REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R., 2011. Sentiment analysis of twitter data. In Proceedings of the Workshop on Language in Social Media (LSM 2011) (pp. 30-38).

[2] Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

[3] Liu, B. (2015) "Preface," in Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press, pp. xi-xiv. doi: 10.1017/CBO9781139084789.001.

[4] Kouloumpis, E., Wilson, T. and Moore, J., 2011, July. Twitter sentiment analysis: The good the bad and the omg!. In Fifth International AAAI conference on weblogs and social media.

[5] Zhou, X., Tao, X., Rahman, M. and Zhang, J. (2017). Coupling topic modelling in opinion mining for social media analysis. Proceedings of the International Conference on Web Intelligence - WI '17.

[6] Neethu, M.S. and Rajasree, R., 2013, July. Sentiment analysis in twitter using machine learning techniques. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[7] 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). IEEE.

[8] Firmino Alves, A., Baptista, C., Firmino, A., Oliveira, M. and Paiva, A. (2014). A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets. Proceedings of the 20th Brazilian Symposium on Multimedia and the Web - WebMedia '14.

[9] Rexha, A., Kröll, M., Dragoni, M. and Kern, R. (2016). Polarity Classification for Target Phrases in Tweets: A Word2Vec Approach. The Semantic Web, pp.217-223.

[10] Noferesti, S. and Shamsfard, M., 2015. Using Linked Data for polarity classification of patients' experiences. Journal of biomedical informatics, 57, pp.6-19.

[11] 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.

[12] Chen, Q. and Sokolova, M., 2018. Word2Vec and Doc2Vec in Unsupervised Sentiment Analysis of Clinical Discharge Summaries. arXiv preprint arXiv:1805.00352.

[13] Haripriya, A. and Kumari, S., 2017, July. Real time analysis of top trending event on Twitter: Lexicon based approach. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-4). IEEE.