Election Result Prediction Using Twitter Sentiments Analysis

Payal Khurana Batra, Aditi Saxena, Shruti and Chaitanya Goel
Dept. of Computer Science and Engineering
Jaypee Institute of Information Technology, Noida, India
payal.f12@gmail.com, aditisa24@gmail.com, shreya27jan@gmail.com, chaitanya.goel31@gmail.com

2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC) | 978-1-7281-7132-6/20/$31.00 ©2020 IEEE | DOI: 10.1109/PDGC50313.2020.9315789

***Abstract*—Democracy is the form of government where people have the right to choose their leaders. In our country, India, elections are held periodically to choose their political leader. Nowadays, people like to know beforehand that who is going to win the elections. They predict results based on news, their discussions and online platforms. Online sites such as Facebook, Twitter and WhatsApp have become a trend these days. Anyone with internet connection has access to these social sites. The views, news and discussions available on these platforms help in predicting the results beforehand. These are the platforms to share ones views' on an online junction. Twitter is one such platform which becomes popular during any event in our country. Anything that is news becomes a trend on twitter. People share their views, hatred, campaigns towards any political party, figure or leader. It is very useful to predict electoral results before the exit poll results. Therefore, this work analyzes tweets collected from twitter and predicts election results by performing sentimental analysis on them.**

***Index Terms*—Sentimental Analysis, Tweets, Twitter, supervised learning, Predictive Analysis**

I. Introduction

Sentimental Analysis is a method to teach a machine to extract emotion from a given text. A text can be anything, a simple review, a social statement, tweets or messages. A large amount of social data which is of high-value and variety has been accumulated on digital platforms. This big social data could be computationally processed and analyzed to learn people associations and tendencies towards any subject [18]. The views of people on social platforms when processed and analyzed can be used to predict results that are useful in making business or any other life decisions.

Sentimental analysis is a process to analyze the tweets to give them a polarity score usually from -1 to 1. -1 is considered as negative and 1 as positive with 0 being neutral [11]. Tweets regarding elections can be analyzed and sentimental analysis can be performed on them to calculate their polarity and further predict election results.

People, in large number have been using social media for the past few years. Social Media has become a powerful tool to share one's views. It has provided strong mediums such as Facebook, Twitter and Google+ to share opinions, reviews, and ratings [4]. The thoughts, opinions, and sentiments of general public play an important factor in deciding election results. Presently, telephone sampling of small group of people is randomly carried out to evaluate presidential approval polls. This technique is both time consuming and costly. Therefore, an automated way of measuring the election polls from easily available social data would be extremely useful in reducing the required costs and time [2]. It is also seen that with advancement of technology, online platforms are much reliable and cost-effective to predict results. Lately, it has been noticed that conventional polls may fail to make a precise and an accurate prediction. Therefore, scientists and researchers have turned their interest in inspecting and analyzing web data, such as blogs or activity of social networks' users' as a different way to predict election outcomes, with a hope of more accuracy. Furthermore, conventional techniques of polls are too expensive, while online information is easily accessible and can be retrieved without expending any money [5]. According to [3] the public timeline carrying the tweets of all social media users around the world is a vast realtime information flow with more than one million messages per hour. The primary purpose behind micro blogging was to provide personal status updates. However, nowadays, posts on micro blogging sites include every possible topic, like political news, product information and many more. These posts can have different formats such as short sentences, direct word to other users or links to different websites. The political news becomes prominent and famed discussion for many social media users during the time of election war between political parties.

In proposed approach, an inbuilt API called as Valence Aware Dictionary and sEntiment Reasoner (VADER) [12] is used for labelling a part of the dataset and that labeled subset is further used for training five machine learning models namely Logistic Regression, Decision Tree, XGBoost, Naive-Bayes and LinearSVC [20]. The two feature extraction techniques namely Bag-of-words (BOW) [13] and term frequency–inverse document frequency (tf-idf) [14] are used with each model. The analysis of five models with two different feature extraction technique shows the combination of Decision tree with tf-idf to be better than other combinations to predict electoral results of Lok-Sabha elections held in 2019.

II. Related Work

There is an extensive research on prediction of election results using machine learning models for sentimental analysis. Researchers have used from simple machine learning models to advance methods of deep learning for predicting election results of various elections held in different areas around the globe [1][6]. The studies on predicting elections have been

carried out from early 2000's and various approaches and researches have been proposed for the same.

The French elections of 2017 were predicted using popularity of tweets[7]. However, their model was only applicable on tweets of French language. Salunkhe et.al [4] further explained how the multiclass classification on two political entities or parties be performed on tweets related to them and can be compared to find winner. Patrick Lai [2] worked on trends of twitter that are related to presidential performance. He has extracted strong sentiments of twitter data related to elections by using lexicon- based approach and have used date-time series method for predicting final outcome. Ramzan et.al [8] stored tweets in MongoDB and have categorized tweets using a sentiment analyzer. The authors used tweets of all the parties involved in election rather working on particular parties.

The authors in [5] used concept of polarity and subjectivity for predicting president of United States of America. They analyzed elections between Donald Trump and Hilary Clinton of 2016 and predicted Hilary as a winner. They also used lexicon analysis for sentimental analysis of tweets. MengHsiu Tsai [9] digged a little deeper in Deep learning and used trained Recursive Neural Tensor Network (RNTN) model to classify tweets on elections held in America. A different approach for the same election was conveyed by Boyi Xie [10] who used Tree Kernel and Polarity Scoring. He formed two classifiers, one which can classify text into positive and negative class and second, which can classify text into positive, negative and neutral class. He named the former as binary classifier and latter as a 3-way classifier.

This paper works on Indian Lok-Sabha election of 2019 and converts problem of unsupervised learning to supervised learning by using python [17] tool such as VADER and using different machine learning models with two feature extraction techniques for achieving final results.

III. Proposed Methodology

The main approach of this paper is to convert an unsupervised problem to a supervised learning problem and then perform sentimental analysis on the data followed by visualization and prediction of results.

A. Dataset Collection and Cleaning

The data of Election of Lok-Sabha 2019 has been collected from different sources and final dataset is formed using various approaches such as twitter library "tweepy" [15] and Github [16]. The main parties involved in central election were BJP and Congress. The tweets related to these two political parties are collected using "tweepy" library. However, this feature of twitter only allows access to limited tweets. To overcome this, open source platform such as Github [16] is used for collection of data and to form a complete dataset.

The main objective after collection of data is to preprocess it. As tweets are largely unstructured data, therefore, the raw data is cleaned by removal of hashtags, stopwords, punctuations, and conversion of entire tweet to lowercase.

B. Splitting of dataset and Applying VADER

The approach after preprocessing of dataset is to split whole dataset into two parts containing BJP and Congress tweets in separate sets. The set containing BJP tweets is taken for labelling sentiment for each tweet by applying VADER which uses lexicon and rule-based analysis to label a sentence. It gives score to each tweet based on their polarity and sentiment [12]. It is to be noted that VADER is applied to only BJP set to use this set as labeled dataset. The labeled BJP set is used to train five different machine learning model, each model combined with two feature extraction techniques.

C. Model Training

For the proposed work, five machine learning models have been analysed.

- **Logistic Regression:** The approach of Logistic Regression is to classify an event into different classes, as in this work, the tweets (event) are classified into positive or negative (different classes) based on the probability.
- **Decision Tree:** Decision Tree machine learning models uses logic and math to generate result, rather than selecting them on the basis of intuition and subjectivity. It uses a tree-like structure for making decisions of an event.
- **Naive-Bayes:** The basic idea of Naive Bayes model is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes. This model classifies using probabilities of events.
- **LinearSVC:** This model uses vector method to classify a tweet into various classes. A decision boundary is constructed between two classes and a data-point is classified between these two classes.
- **XGBoost:** In this model, the gradient boosted decision trees are arranged and designed for better execution and production.

The two feature extraction techniques combined with each model are BOW and tf-idf.

The BOW approach uses a bag like structure to store words of a text (such as a sentence or a document) and uses term frequency, namely, the number of times a term appears in the text to give polarity to a sentence [13].

The tf-idf technique is a numerical statistic procedure that reflects the importance of a word in a document [12] and uses weightage and frequencies of words in a sentence for giving polarity.

The model and feature extraction technique combination with the best accuracy is used as the final model for labelling data of Congress set. These two sets of BJP and Congress are then combined for further visualization and prediction.

D. Model Training

The proposed approach is shown in Figure 1.

- **Dataset Collection:** Data is collected using "tweepy" library provided by twitter. Some data is collected through Github too.
- **Data Cleaning:**All tweets contained in dataset have been preprocessed and cleaned by removing stopwords,

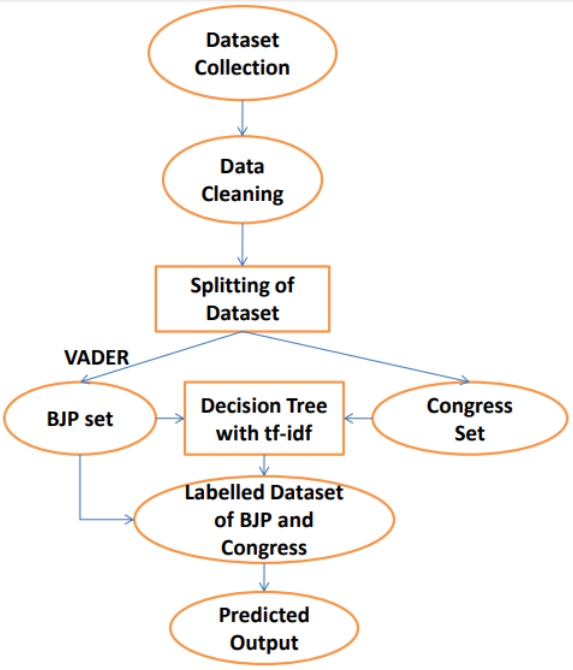

Fig. 1. Proposed Approach

hastags, punctuations, and conversion of entire tweet to lowercase.

- **Splitting of Dataset:** Data has been split into BJP and Congress set. The BJP set have been labeled by VADER and this set is used to train five models. The model with better accuracy is used to label Congress set.
- **Labelled Dataset:** In this module, both the sets of BJP and Congress have been combined with labelling of '1' to positive tweet and '-1' to negative tweet
- **Predicted Output:** The number of positive tweets of both parties have been calculated and the party having higher number of positive tweets is predicted as final output of this work.

IV. Results and Discussions

The election prediction of Lok-Sabha elections of 2019 has been done on open source web application "Jupyter Notebook" on a 64-bit processor [19].

Dataset has been formed containing 41,265 total tweets of BJP and Congress using "tweepy" and Github. The dataset collected has columns of location, time, party which that tweet belongs to, and a whole tweet. The columns of whole tweet and their party are taken into account for this analysis. The neutral tweets of both parties have been ignored as they hardly contribute in final results.

The accuracies of five machine learning combined with BOW feature exraction technique is shown in Figure 2 and Table 1. BOW model has accuracy of 81.7% with Logistic Regression, 84.5% with XGBoost, 86.0% with Decision Tree, 79.4% with Naive-Bayes and 82.2% with LinearSVC. The Decision Tree with BOW has highest accuracy among all combinations of models with BOW.

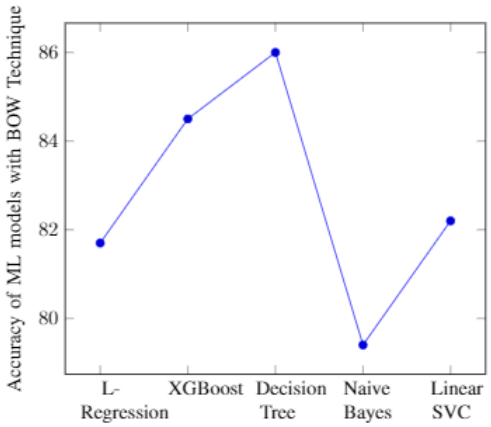

Fig. 2. Accuracy of ML Models combined with BOW Technique

TABLE I
Accuracy of Machine Learning Models with BOW Technique

Machine Learning Models (Combined with BOW)	%ageAccuracy
Linear Regression	81.7
XGBoost	84.5
Decision Tree	86.0
Naive-Bayes	79.4
LinearSVC	82.2

The accuracies of tf-idf with five models has been depicted in Figure 3 and Table 2. tf-idf model has accuracy of 80.9% with Logistic Regression, 84.9% with XGBoost, 86.3% with Decision Tree, 79.1% with Naive-Bayes and 80.2% with LinearSVC. The Decision Tree with tf-idf has highest accuracy among all combinations of models with tf-idf technique.

The accuracies of Decision tree with both Bag-of-Word and tf-idf is greatest in their respective sets and compared in Figure 4. The model of Decision Tree with tf-idf is better with accuracy of 86.3%. The model of Decision Tree with tf-idf is used to label Congress set. The most occuring words of Congress set has been shown in Figure 4.

The two sets of BJP and Congress are combined for prediction of final results. The tweets having label as '1' is considered as positive tweet. The number of tweets having label as '1' are calculated for both BJP and Congress. The number of positive tweets of both parties have been shown in Figure 6. The Congress party has 10111 positive tweets and BJP has 20130 positive tweets. The party having higher number of positive tweets is predicted the winner of the

TABLE II
Accuracy of Machine Learning Models with tf-idfTechnique

Machine Learning Models (Combined with BOW)	%ageAccuracy
Linear Regression	80.9
XGBoost	84.9
Decision Tree	86.3
Naive-Bayes	79.1
LinearSVC	80.2

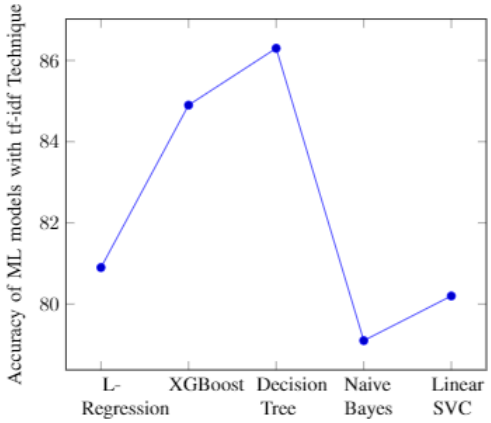

Fig. 3. Accuracy of ML Models combined with tf-idf Technique

Fig. 4. WordCloud of weighted words of Congress set

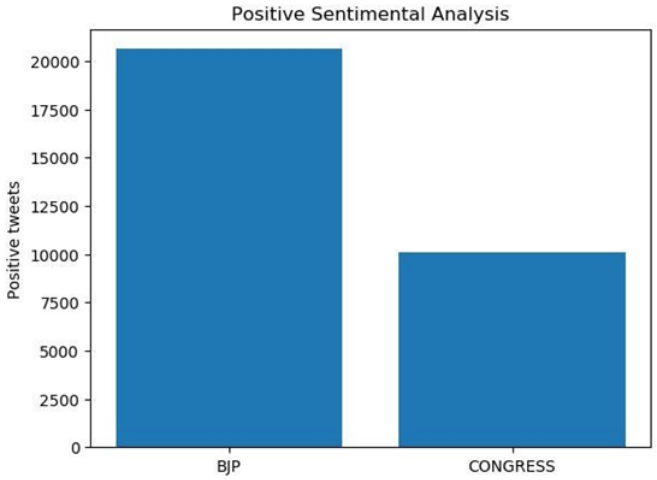

Fig. 5. Number of positive tweets of BJP and Congress

elections as the higher positive tweets clearly shows the popularity and positivity of that party among twitter users. As it comes out the number of positive tweets of BJP are very much greater than number of positive tweets of Congress, BJP party is predicted as final winner.

V. Conclusions

This work analyses two feature extraction techniques namely BOW and tf-idf combined with five machine learning. models on a set which has been labeled by VADER. Among five combination, Decision Tree with tf-idf is better and is used as final model which is applied on tweets of Indian Loksabha elections 2019 which predicts BJP win over Congress.

This work can further be extended on tweets of different regional languages of Indian states other than English to improve accuracy

References

[1] A. Sarlan, C. Nadam, and S. Basri, "Twitter sentiment analysis," in *Proceedings of the 6th International conference on Information Technology and Multimedia*, pp. 212–216, IEEE, 2014.
[2] P. Lai, "Extracting strong sentiment trends from twitter," 2010.
[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *Fourth international AAAI conference on weblogs and social media*, Citeseer, 2010.
[4] P. Salunkhe and S. Deshmukh, "Twitter based election prediction and analysis," *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, p. 10, 2017.
[5] F. Nausheen and S. H. Begum, "Sentiment analysis to predict election results using python," in *2018 2nd international conference on inventive systems and control (ICISC)*, pp. 1259–1262, IEEE, 2018.
[6] F. J. J. Joseph, "Twitter based outcome predictions of 2019 indian general elections using decision tree," in *2019 4th International Conference on Information Technology (InCIT)*, pp. 50–53, IEEE, 2019.
[7] L. Wang and J. Q. Gan, "Prediction of the 2017 french election based on twitter data analysis," in *2017 9th Computer Science and Electronic Engineering (CEEC)*, pp. 89–93, IEEE, 2017.
[8] M. Ramzan, S. Mehta, and E. Annapoorna, "Are tweets the real estimators of election results?," in *2017 Tenth International Conference on Contemporary Computing (IC3)*, pp. 1–4, IEEE, 2017.
[9] M.-H. Tsai, Y. Wang, M. Kwak, and N. Rigole, "A machine learning based strategy for election result prediction," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1408–1410, IEEE, 2019.
[10] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, pp. 30–38, 2011.
[11] V. Kharde, P. Sonawane, *et al.*, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.
[12] https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/.
[13] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.
[14] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting tf-idf term weights as making relevance decisions," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, pp. 1–37, 2008.
[15] https://www.tweepy.org/.
[16] https://github.com/.
[17] https://www.python.org/.
[18] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," *Journal of Big Data*, vol. 5, no. 1, p. 12, 2018.
[19] https://jupyter.org/.
[20] K. Das and R. N. Behera, "A survey on machine learning: concept, algorithms and applications," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 2, pp. 1301–1309, 2017.