Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree

Ferdin Joe John Joseph
Faculty of Information Technology
Thai-Nichi Institute of Technology, Bangkok
ferdin@tni.ac.th

Abstract – Social Media is a huge corpus of raw data available to process and analyze the mood of the people. Various institutions and corporations use social media data to understand market response of their products. Twitter data is widely available one for this purpose. Similar to market analysis, this data is used to map political mood of the general public. Most of the existing methodologies use tweets downloaded on a particular criterion. There is very limited study done on daily mood mapping of political views among people. This paper addresses a methodology to predict the outcome of the 2019 Indian general elections using the sentiment analysis of twitter data. Decision tree classifier is used to train and test data and the predicted outcome is found to be close to that of the actual outcome and most of the pre poll analysis done so far. The experiments reported in this paper are only on tweets in English language and having the most number of retweets by the users. This methodology is efficient enough to map the mood of people over a timely basis across various phases of polls.

Keywords – Sentiment Analysis, Decision Tree, Twitter Data, Election Prediction

I. INTRODUCTION

The Indian general elections are held once in 5 years to elect the Member of Parliament (MP) from 541 constituencies all over the country. These elected MPs will elect the Prime Minister who will rule the country for the next 5 years. For confirming the term, the Prime Minister who took oath should prove his majority in Lok Sabha, the lower house of the Indian Parliament. In this motion, the Prime Minister must have support of at least 272 MPs including the vote of the elected speaker. This process is happening since 1952 when the country was declared a republic. The country had elections in 1952, 57, 62, 67, 71, 77, 80, 84, 89, 91, 96, 98, 99, 2004, 09, 14 and 19 respectively. Until 2009 elections, the election campaigning happened on party and ideology centric. Since 2014 the trend shifted towards Prime Minister candidate. There are many processes carried out by various independent agencies to predict the outcome of polls in the past. Pre poll analysis was conducted weeks or months before the election and Exit poll was conducted on the day of election with the voters returning from the voting booth as sample space. This is done on all the 7 phases of election conducted in various parts of the country with different timetable. These kind of poll predictions were banned by the Election Commission of India under the model code of conduct for the elections to happen without bias in the largest democracy of the world. The internet users in India has increased exponentially over the past decade and it is expected to continue the trend in the future. Though the process of poll predictions is banned due to the model code of conduct, it applies to twitter based poll predictions as well. So this process is carried out during the polling season and the results are declared after the model code of conduct is lifted by the Election Commission of India, which is now legal. This paper discusses about the prediction of election outcomes from twitter data in the past and proposes a new methodology good enough for a long electoral process done in the Republic of India during the year 2019. The results obtained are for the incumbent ruling party against all the opposition parties combined. For a country with nearly 1.3 billion populations, it is not practical enough to conduct pre poll and exit poll surveys in all 543 constituencies with a neutral and transparent manner. So the agencies managing the political campaign and the media house need to gain knowledge through freely available raw data from which the exact outcome is understood. The section II will discuss on the various methodologies used so far to predict elections all over the world and the section III discusses the proposed methodology in this paper. Then the results are visualized and the method is justified over in section IV.

II. RELATED WORK

There are many methodologies proposed for election outcome prediction from tweets. A handful of literature is available for Indian elections. Election outcome prediction using twitter has a history dating back to 2012. It was for Queensland state election, the methodology focused on the issues raised in twitter and the popular mention [1]. A substantial study on the sentiment on twitter was done to forecast 2013 Pakistani and 2014 Indian elections [2]. A Diffusion centric model was created to map the sentiments of people in election centric issues. This has no support to the popularity or any forecast in the respective elections. An unsupervised learning based methodology was used for 2016 US presidential poll campaign based tweets [3]. This was done for two days of tweets with around 60000 tweets each for Donald Trump and Hillary Clinton.

Swedish elections were predicted in [4] using the frequent mapping of political behavior of users such as retweet count, likes and other aspects. A support vector machine and Convolutional Neural Network based classification was done to forecast elections in UK [5]. This methodology obtained around 80% accuracy. Normally these twitter based election

978-1-7281-1019-6/19/$31.00 ©2019 IEEE

classification methods are good for countries with two party system. For countries like India having multi-party democracy, forecasting of election outcome is still a challenge.

German federal elections were predicted using twitter text. This attempt was done on the twitter mentions and they got an error rate of 1% [6]. However, the advent of twitter bots and IT wing of political parties make this method not convincing enough to predict.

From all the literature studied above, it is clear that a specific methodology Is needed to predict in a multi-party democracy using twitter feed. This process of forecasting elections has an impact in stock exchanges as the indices of market get affected to a considerable extent when an unexpected outcome is going to come in the actual result of the election.

III. PROPOSED METHODOLOGY

The proposed methodology consists of the 3 phases. The first two phases are performed every day during a fixed time.

A. Data Collection

Tweets are downloaded from the twitter database using twitter API connectivity. Jupyter Notebook with tweepy [7] library is used. The collected tweets are stored in Mongo DB. This is done using the pymongo [8] library. Every day during a particular time, 5000 tweets each for the ruling and opposition parties most famous twitter handles were extracted. Tweets with twitter handles of the then Prime Minister, ruling party leaders and the party itself are taken as ruling party tweets and the tweets with twitter handles of the leader of opposition, opposition party major stakeholders and the opposition and regional parties are taken as opposition party tweets. These tweets are extracted with the conditions of most popular tweets with the most retweets and English as the language of tweets. In the experimentations on the proposed methodology, tweets in English language alone are taken. The sentiment classifier and lexical analysis is available for English language but not available especially for Indian languages.

B. Preprocessing

Preprocessing is done to prune regular expressions and emoji's not available in ASCII or Unicode. These symbols are not interpretable to any kind of sentiment and were responsible for exceptions during the preliminary experimentation. After pruning of unusable regular expressions, the tweets are sorted based on the total number of retweets. Then the attributes ID, text and lang are taken for easier processing of data. This is made to include a sparse data processing and management [9]. Inorder to obtain an efficient decision, stopwords are removed using the nltk corpus [10] of stopwords. After removing stopwords, regular expressions, emojis, Unicode and punctuation marks are pruned. Then the text is checked whether it has any non-English words and then pruned. After this procedure the resultant text is then subjected to normalization by tokenizing the words and making it parse-able by the classifier.

C. Sentiment Analysis

Long before the data collection, preliminary experimentations were done using classifiers like Artificial Neural Network, Naïve Bayes Classifier and SVM. These methods were not good enough to experiment by creating a tree of grammar from the Parts of Speech (POS tag). Decision tree was good in proving the necessary scores in polarity and subjectivity to map the mood of people. This was evident with a rough estimation during the State Assembly elections in end of 2018. However, this process is not a refined one to report as a methodology. With this result of primary investigation, decision tree classifier is used to predict the polarity of tweets in the proposed methodology. Decision Tree classifier in Textblob [11] library is used to classify the text extracted. Using this classifier, the polarity and subjectivity of sentiments from each tweet is calculated. From the scores of polarity and subjectivity, the tweet is classified as positive, negative or neutral. The popularity score for each day is calculated on the tweets downloaded on that particular day. This process was carried out for 50 days during the election season.

Popularity = ((0 x Negative tweets) + (Neutral tweets / 2) + Positive tweets)/Total tweets

Negative tweets are not given any score. The proposed methodology is designed to find only the popularity. Whatever be the lower bound or higher bound weights to negative tweets gave similar trend. This popularity score is recorded for tweets in favor of ruling party keywords and other parties' keywords separately. For each phase of the poll, the average of popularity scores was recoded and they were used to calculate the number of seats possible to be won on those getting polled on that phase of polls. The number of seats predicted was taken to project how many seats possible to be won by the respective party. The results obtained was compared to the pre poll survey and the actual result obtained on the counting day.

IV. RESULTS

The popularity score obtained every day for ruling party and other parties are listed in Fig 1. There might be some ups and downs in the trend but the popularity scores' difference was always positive for ruling part and it is evident from the trend obtained from Fig 2. The highest popularity obtained by any party in a particular day was 72% by the ruling party and the lowest was recorded by the opposition as 49%. This is the daily trend of popularity. The popularity difference is obtained over various phases and it is the average of popularities from day 1 to the respective phase of election.

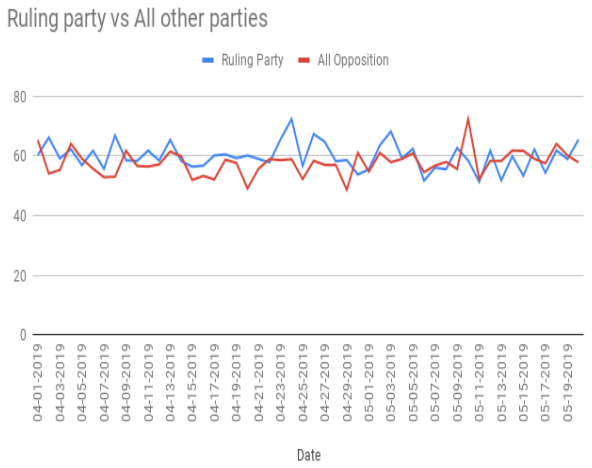

Fig 1: Popularity scores obtained by ruling and other parties every day.

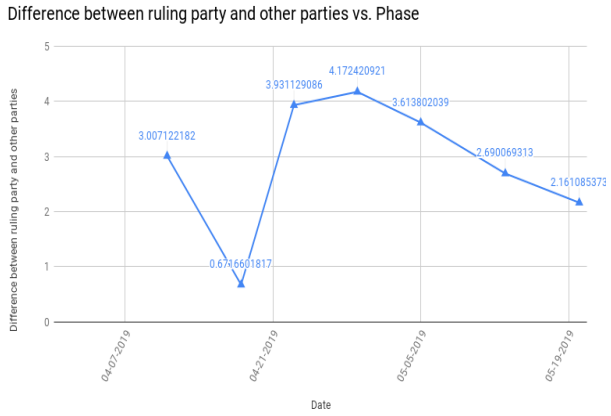

Fig 2: Difference between ruling party and other parties' popularity during various phases.

The lowest performance was recorded during the second phase of polls. It was the phase when the seats taking poll gave results in favor of the opposition parties. The seats won by the ruling party on the phase 2 also had a little difference when compared to those won in other phases.

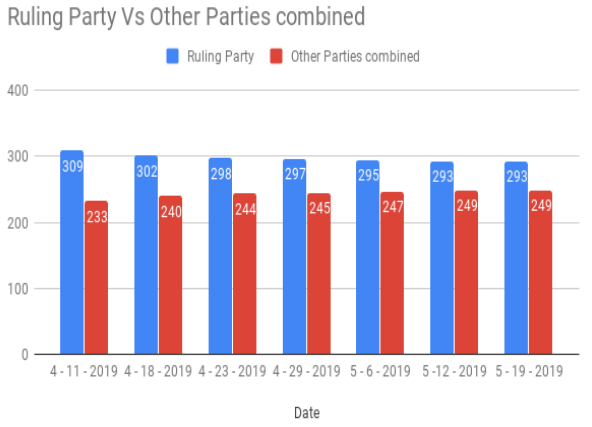

Fig. 3: Projected seats to be won by ruling party and other parties during various phases.

The projected seats over the phases shows the seats possible for ruling or other parties to win during that particular phase of poll and it is evident from Fig. 3.

TABLE I: PREDICTION OF PROPOSED METHODOLOGY COMPARED WITH VARIOUS SURVEYS AND ACTUAL RESULT OBTAINED.

Date Obtained	**Survey agency**	**Ruling Party**	**All other Parties**	**Effect to Ruling Party**
23 May 2019	2019 General Election Results (Actual)	**303**	239	Win
20 May 2019	**Proposed Methodology (Prediction on Ruling Party)**	**293**	249	Win
April 2019	Times Now-VMR [12] (Total for Alliance)	**279**	264	Win
April 2019	IndiaTV-CNX [13] (Total for Alliance)	**275**	268	Win
April 2019	Jan Ki Baat [14] (Total for Alliance)	**310**	233	Win

From the Fig. 3, the latest result predicted was 293 and it is compared against various pre poll surveys and actual result. While most of the pre poll surveys gave a near majority number, the proposed methodology produced a prediction result with 97% accuracy. Fig 3 shows that during the first phase of election, the ruling party was capable of winning over 309 seats while during other phases it reduced to 303 due to some controversial statements issued by the prime campaigners. All the experiments were carried out in Jupyter Notebook using Python 3.7 on a windows environment using 4GB RAM and intel core i7 processor. Internet connection stability works fine over a fibre optic connection with atleast 30 mbps download bandwidth.

V. CONCLUSION

It is evident from the results that the proposed methodology gave a near prediction to the actual result from analyzing tweets in English language. There is a need for this methodology to analyze tweets in other languages. This will help predict elections where non English language is spoken in majority or to map the trends of people's mood in country like India over various states speaking and tweeting multiple languages. The results obtained in the proposed methodology shows that this methodology has a promising future in predicting Indian General elections. However, this trend has to be validated with the results obtained by State Assembly elections. There are many other deep learning classifiers like Convolutional Neural Networks (CNN), Recurrent CNN etc

but the proposed methodology is not dealt with any deep learning methodology.

ACKNOWLEDGEMENT

The author thanks the reviewers for their time and valuable comments while reviewing this paper. Gratitude is bestowed to the support provided by Thai-Nichi Institute of Technology, Bangkok during the entire process of research presented in this paper.

REFERENCES

[1] J. Burgess and A. Bruns, "(Not) the Twitter election: the dynamics of the# ausvotes conversation in relation to the Australian media ecology," *Journal. Pract.*, vol. 6, no. 3, pp. 384–402, 2012.

[2] V. Kagan, A. Stevens, and V. S. Subrahmanian, "Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election," *IEEE Intell. Syst.*, vol. 30, no. 1, pp. 2–5, 2015.

[3] J. Ramteke, S. Shah, D. Godhia, and A. Shaikh, "Election result prediction using Twitter sentiment analysis," in *2016 international conference on inventive computation technologies (ICICT)*, 2016, vol. 1, pp. 1–5.

[4] A. O. Larsson and H. Moe, "Studying political microblogging: Twitter users in the 2010 Swedish election campaign," *New Media Soc.*, vol. 14, no. 5, pp. 729–747, 2012.

[5] X. Yang, C. Macdonald, and I. Ounis, "Using word embeddings in twitter election classification," *Inf. Retr. J.*, vol. 21, no. 2–3, pp. 183–207, 2018.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Election forecasts with Twitter: How 140 characters reflect the political landscape," *Soc. Sci. Comput. Rev.*, vol. 29, no. 4, pp. 402–418, 2011.

[7] J. Roesslein, "tweepy Documentation," *Online] http://tweepy. readthedocs. io/en/v3*, vol. 5, 2009.

[8] A. Nayak, *MongoDB Cookbook*. Packt Publishing Ltd, 2014.

[9] F. J. John Joseph, R. T, and J. J. C, "Classification of correlated subspaces using HoVer representation of Census Data," in *2011 International Conference on Emerging Trends in Electrical and Computer Technology*, 2011, pp. 906–911.

[10] E. Loper and S. Bird, "NLTK: the natural language toolkit," *arXiv Prepr. cs/0205028*, 2002.

[11] S. Loria, "textblob Documentation," 2018.

[12] T. N. Bureau, "Times Now-VMR Opinion Poll For Election 2019: PM Narendra Modi-led NDA likely to get 279 seats, UPA 149," *Times Now News*, New Delhi, 2019.

[13] I. T. News Desk, "Lok Sabha Election 2019: NDA may get thin majority with 275 seats, BJD may retain Odisha, YSR Congress may win Andhra, says India TV-CNX pre-poll survey," *India TV*, 2019.

[14] "2019 Indian General Election," *Wikipedia*, 2019. [Online]. Available: https://en.wikipedia.org/wiki/2019_Indian_general_election.