Sentiment Analysis of Twitter Data for Poll Prediction

Vivsvaan Sharma
Department of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Sector 26, Chandigarh

Ashlin K Siby
Department of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Sector 26, Chandigarh

Abhinav Khetarpal
Department of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Sector 26, Chandigarh

Kamalpreet Singh
Department of Computer Science and Engineering
Chandigarh College of Engineering and Technology
Sector 26, Chandigarh

***Abstract*—Social media platforms like Twitter, Facebook, etc. are increasingly being used by people to express their opinions and interests. This has piqued the interest of businesses and other social entities, which are employing very sophisticated tools to gather and make sense of the large amounts of data available on these platforms. Sentiment analysis of the data obtained from such social media platforms has proven to be an effective way of capturing the opinion of the people and predict trends, thereby improving the decision making the process. One such application of this is the Prediction of Election Results.**

We propose a machine learning model to predict the number of votes obtained by two major parties Bharatiya Janata Party(BJP) and Indian National Congress for Lok Sabha Elections of 2019 in India. We mined data from Twitter for General ELections of 2019 as well as General Elections of 2014. Mined data was cleaned and relevant data was fed into Vader for sentiment analysis. The polarity of tweets obtained was used in a Regression model. Data pertaining to the General Elections of 2014 was used for training the model, which was then used to predict the results of the 2019 elections.

Prediction is done for each State and Union Territory of the country and accuracy of the model is obtained by comparing the results with the actual values of the votes obtained by the two parties in each state in the 2014 polls.

Keywords— sentiment analysis; prediction; elections; Twitter; Regression.

I. Introduction

1.1 Motivation

The election process is an integral part of the fabric of a country. It is a complex process and many socio-political institutions, including political parties, attempt to predict the trends and mood of the people to gain an insight into the process and make use of the gathered information. Machine Learning tools are perfect for analysis of such a process and far surpass the traditional means.

The rapid increase in the number of users in social media has provided users a powerful platform to voice their opinions. Social media Platforms like Twitter is being actively used to share ratings, reviews, and recommendations. Various psephologist has agreed to the fact that this vast array of information can be actively used for marketing and social studies. Political Campaigns have exploited this vast array of information available on the above platforms to draw insights about user opinions and thus design their marketing campaigns. Huge investments by politicians in social media campaigns right before an election along with arguments and debates between their supporters and opponents only enhance the claim that views and opinions posted by users have a bearing on the results of an election.

The ability to extract insights from social data is a practice that is widely gaining momentum throughout the world and forms a fascinating area of study with the ability to provide a wider view on how public opinion is shaped.

1.2 Contribution

This work provides a new approach to tackle the task of prediction of polls using data gathered from social media site Twitter. In contrast to the existing work where sentiment analysis is performed on gathered data and classification methods like Naive Bayes or SVM are applied to categorize the data into one or more classes, the problem has been tackled as one of Regression where a Machine Learning has been used to measure the correlation between sentiment expressed on Twitter and the number of votes a party receives. Moreover, when tweets come into the picture, hashtags themselves become important features. And no other data set can provide hashtags as features except the data that has been mined from Twitter for that specific application. Hence, it becomes necessary to devise a labeling technique for the mined Twitter data which can strike a balance between speed and accuracy.

The remainder of this paper has been organized into 4 sections.

'Related Work' discusses the existing literature and previous work done in the domain. Further, it is discussed how our work builds on top of the current work done.

'Material and Methods' section describes the approach used and provides a step by step account of the process followed as well as details about the machine learning methods used.

In the 'Results and Discussions' section analyzes the performance of the machine learning model as well as our predictions for the 2019 General Elections.

'Conclusion' section summarizes the work done.

II. Related Work

The major portion of related corpus focuses on the process of Data Gathering and subsequent Sentiment Analysis of preprocessed data.

In K. Mao, J. Niu, X. Wang, L. Wang, and M. Qiu[1] cross-domain dataset has been used to train the sentiment analysis model, to make up for the lack of availability of contextually relevant data.

In Jyoti Ramteke, Darshan Godhia, Samarth Shah and Aadil Shaikh[2] the authors discuss how gathered data can be effectively labeled and prepared for classification. Further comparison of different classification models has been provided.

In D. Das and S. Bandyopadhyay[3] data in Bengali corpus has been classified in six feeling classes and to annotate the sentences.

A. Bakliwal, P. Arora, and V. Varma[4] generated a word reference by using fundamental graph traversal of antonym words and they are further used to generate a subjectivity vocabulary.

S. Mukherjee and P. Bhattacharyya[5] used dictionaries of positive and negative words to determine the polarity.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede[6] explains about the model updates that combine pack of-words in talk markers with the slant demand by 4% exactness.

A. Bakliwal, P. Arora, and V. Varma[4] suggested depicting Hindi reviews as positive, neutral and negative. They figured out another score breaking point and used it for two different techniques. Moreover, they used a fusion of the POS Tagged Ngram and central N-gram approaches.

Birmingham and Smeaton [7] tested two distinct strategies, Multinomial Naïve Bayes (MNB) and SVM for web pages and scale blog. They found that MNB methodology outperforms SVM on scaled scale areas with short substance.

Mittal et al. [8] generated an efficient approach to identify the sentiment from Hindi content. They built up Hindi language corpus by adding more opinion words and improve the present Hindi SentiWordNet (HSWN). Their algorithm showed 80% precision on the course of action of studies

Pak and Paroubek in [13] utilized tweets which end with emoticons like ":)" ":-)" as positive, and ":(" ":-(" as negative. They accumulated models including Max Entropy, Support Vector Machines (SVM) and Naive Bayes and concluded that SVM performed the best amongst various others, attaining more precision which leads SVM to be the best performer of all the classifiers. They recorded that all distinctive models were beaten by the unigram model.

For cross-domain sentiment analysis, the approaches by Wu and Tan [9] and Liu and Zhao [10] are discussed. Wu and Tan [9] use a two-stage framework as follows: At the first stage, an association is created between the source and the target domain by applying a graph ranking algorithm [11]. Then some of the best seeds from the target domain were selected. At the second stage, they used the essential structure to calculate the sentiment score of each document and then the target-domain documents were labeled based on these scores. In both the techniques used above, the overall accuracy has been roughly 70% which is less in comparison to supervised learning methods. Thus, it reinforces the claim that an accurate and contextually relevant training data set is vital to achieving highly accurate results for text classification as illustrated in Neethu, M. S., and R. Rajasree[13]. However, the authors achieve only a sparse data set of 1000 tweets which fail to satisfy the quantitative aspect required by a supervised learning algorithm.

Existing work is limited due to the very fact that it focuses solely on the sentiment analysis phase and simply classifies the gathered data into classes.

In this text - sentiment analysis, although a very important phase, has been looked at like the first step in preparing a machine learning model. Gathered tweets were first organized according to the states from which they originated. By performing sentiment analysis on the tweets, the overall positive and negative sentiment towards the two major parties, BJP and Congress, is determined in each state by assigning a score to the two polarities. Then a Regression model has been establish how the positive and negative sentiment as expressed by users on Twitter relates to the number of votes that the two parties obtain. This is in contrast to existing classification centric models used as it goes beyond simply analyzing the data and predicts how the opinion of public affects the number of votes obtained.

III. Materials and Methods

A. Data Collection

This task included a series of decisions that needed to be made. The Twitter Developer platform is a powerful tool that offers different approaches according to the goals and demands of each project. Calls to the Twitter API are free, upon request and can be addressed using two different approaches The Streaming API and The Search API

The Search API was the optimum for this research project's needs. But there were some other challenges that had to be addressed. The tweets had to be originated from Indian citizens located within India. Twitter data for two parties - namely BJP (Bharatiya Janata Party) and Congress (Indian National Congress) were collected.

Table I

Approximate data for candidates

Party Name	Total Tweets
BJP	3471
Congress	3125

B. Data Preprocessing

Real-world data is often incomplete, inconsistent and/or lacking in certain behavior/trends and is likely to contain many errors. The search API returns a massive volume of metadata for every tweet. The fields that we were interested in were Tweets Location, Date posted, Tweet's Text and User's followers count

The collected raw data were transformed into an understandable format, and stored in CSV files. It included the following steps:

Data Cleaning - This process included filling in missing values, smoothing the noisy data (all the tweets were stripped off special characters like '@' and URLs to overcome noise), resolving inconsistencies in data.

Data Integration - Data with different representation are put together and conflicts were resolved.

Data Transformation and reduction - Data is normalized, aggregated and generalized. Then it is represented in a reduced form and stored in DataSet.

C. Data Labeling

Vader (Valence Aware Dictionary and sEntiment Reasoner)[12] is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is basically a sentiment intensity polarizer. Vader takes a sentence as input and provides a percent value for three categories - positive, neutral, negative and compound (overall polarity of the sentence).

Polarity values range from -1 to 1, where a positive compound value indicates that the overall sentiment expressed in the sentence is positive and vice versa.

Apart from the polarity value, each tweet has associated with it the Indian State where the tweet was made from and the number of followers of the account from which the tweet was made.

Table II

Vader Sentiment Analysis Examples

Sentence	comp	pos	neg	neu
He is smart and funny	0.83	0.75	0.0	0.254
A horrible Book	-0.822	0.0	0.79	0.20
It sucks, but I'll be fine	0.22	0.274	0.195	0.53

The above table provides three examples of sentences analyzed using Vader. The first sentence is highly positive, second highly negative and the third is neutral. For performing sentiment analysis, a training data set should consist of sentences that are either positive or negative.

Thus the method proposed above was used to calculate the polarity of each tweet.

D. Creating a Training Set

The Dataset for 2014 Elections was used to create our training dataset. From the State Wise Result of 2014 Elections, date related to Total valid votes polled in states and Total valid votes polled by parties was collected and was integrated with our twitter dataset.

Table III

Example of Training Dataset

location	**party name**	**sentiment**	**follower count**	**Total votes polled in the state**	**Total votes polled by parties**
Bihar	BJP	0.75	5541	42126	124721
Delhi	Cong	0.41	2145	74125	512346
Delhi	BJP	-0.21	8520	321546	221346

E. Design

The proposed model can be divided into 3 main stages on the basis of the nature of the task to be performed.

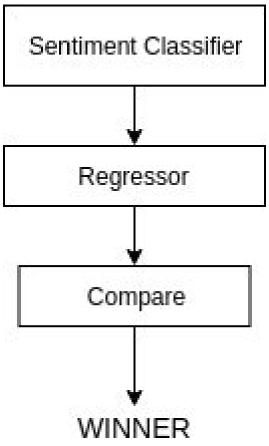


Fig 1 Model for Election Prediction

Sentiment Analysis phase involves attributing the data with its associated polarity value.

Final data set created after sentiment analysis and data labeling is fed into a regressor. The regressor is trained on the data collected for General Elections of 2014 and then is used to predict the results of 2019 elections.

One of the regressors used is Multivariate Linear Regression. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Formally, the model for multiple linear regression, given *n* observations, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \textbf{ for } i = 1,2, \ldots n.$$

In the least-squares model, the best-fitting line for the observed data is calculated by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. The least-squares estimates b_0, b_1, ... b_p is usually computed by statistical software.

The values fit by the equation $b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$ are denoted $\hat{y}_i$, and the residuals e_i are equal to y_i - $\hat{y}_i$, the difference between the observed and fitted values. The sum of the residuals is equal to zero.

The variance σ^2 may be estimated by

s^2 $\frac{\sum e_i^2}{n - p - 1}$ =,, also known as the mean-squared error (or MSE).

The estimate of the standard error *s* is the square root of the MSE

Finally, the accuracy of the regression model is measured and results are compared.

F. Implementation

To implement the supervised Regression model design, the performance of the Multivariate Linear and other various regression models was compared.

TABLE IV

SUPERVISED REGRESSION TECHNIQUES COMPARISON

Regressor	Score
Multivariate Linear Regression	87.2%
Decision Tree Regression	99.4%

Based on the metric of the score, we selected Linear regression.

IV. RESULTS AND DISCUSSIONS

After comparing the accuracy it is clear that Decision Tree model is way better and optimum than Linear Regression but here is the twist, using Decision Tree model on the 2014 dataset (by splitting it into train and test set) will give in a logically incorrect result (i.e. 100% accuracy always). It should be trained using 2014 dataset and should be tested using 2019 dataset. So for now, only the Multivariate Linear Regression Model is considered.

The Multivariate Linear Regression Model is trained on the 2014 election dataset the features are location, party, sentiment, and followers, and values to be predicted are total votes polled in states. Now for testing this trained model we a test set is needed. Sentiments Dataset for 2019 is used here as a test set, in which the total votes polled by parties in states is predicted.

Since the state wise result of 2019 elections is not yet declared, the comparison of the predicted and actual values (here votes) cannot be made, thus accuracy also cannot be calculated.

To resolve this issue, the dataset for 2014 elections is split into training and test set and the linear regression model is trained on that training set and then it predicted values (total votes polled by parties in states) for the states in the test set. To avoid repetitions of states in training and test set, data of 3 states (i.e. Andaman & Nicobar Islands, Andhra Pradesh, and Arunachal Pradesh) was removed from the training set and data of all other states, except for these, was removed from the test set. So basically, here the model is being trained on all the states except for those 3, and then it is being tested on those 3 states.

After the prediction of total votes polled by parties in states is made by our model, the predicted and actual values compared and Mean Squared Error and Accuracy Score is calculated.

Accuracy Score - 87.2%

Mean Squared Error - 0.1279 (after feature scaling)

A graph is plotted showing the comparison between total predicted votes polled by BJP and Congress in those 3 states.

TABLE V

TOTAL VOTES POLLED IN 3 STATES

State	Cong	BJP
Andaman and Nicobar	30987811	83448690
Andhra Pradesh	82413193	129013303
Arunachal Pradesh	19452099	18350990

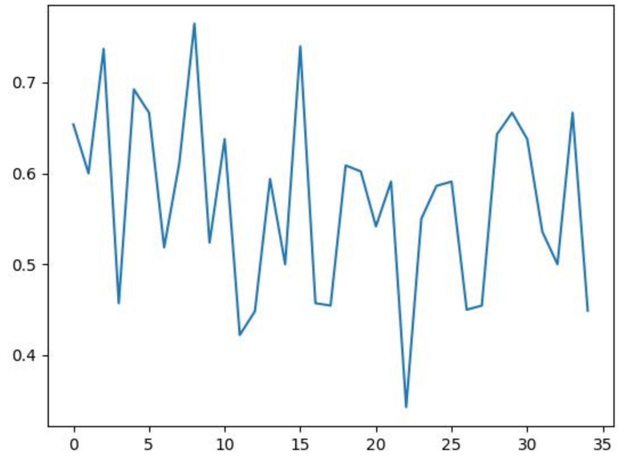


Fig 2 Fraction of positive sentiment for BJP(Refer Table 6)

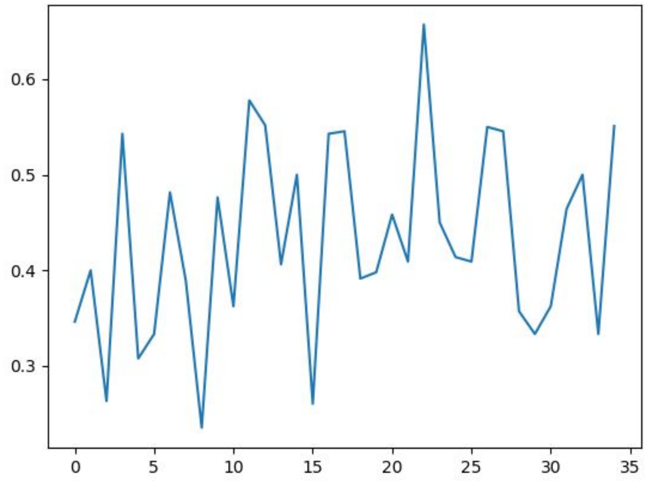


Fig 3 Fraction of positive sentiment for Congress (Refer Table 6)

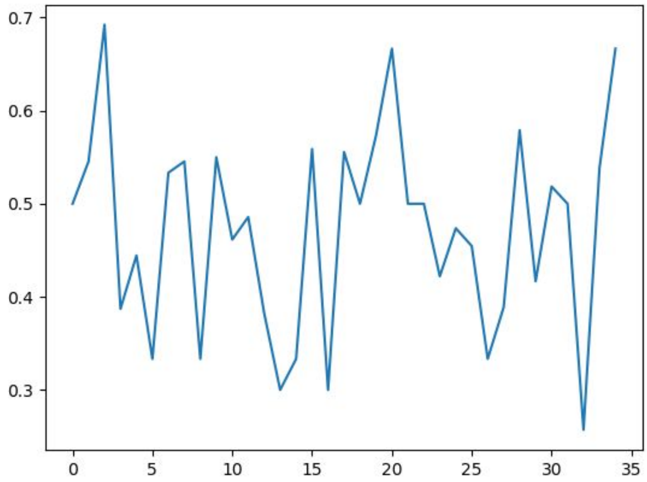


Fig 4 Fraction of negative sentiment for BJP (Refer to Table 6)

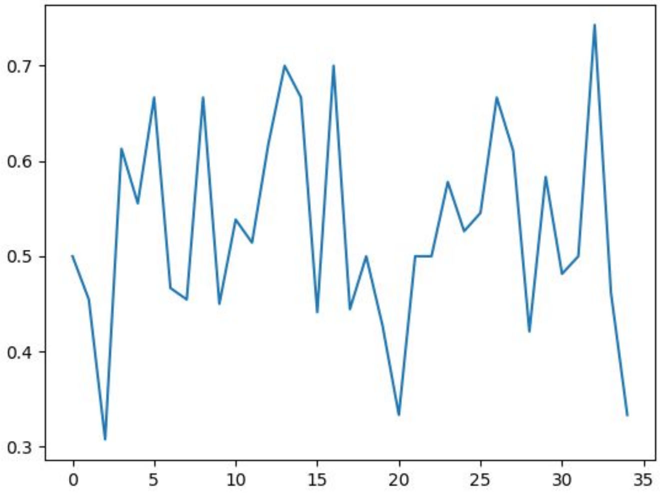


Fig 5 Fraction of negative sentiment for Congress

(Refer to Table 6)

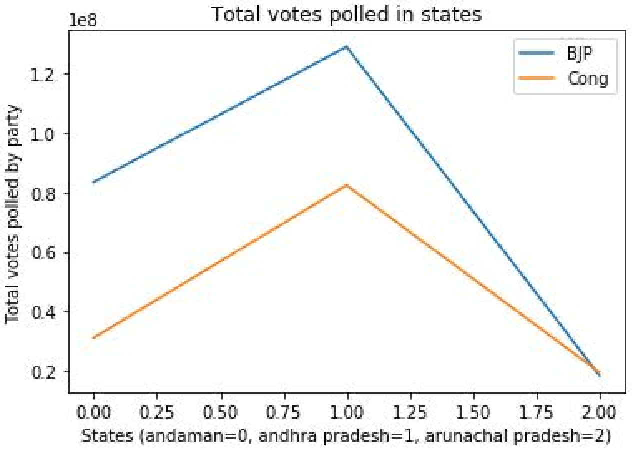


Fig 6 Total Votes polled in 3 states (Refer Table 6)

On seeing the graph, it is clear that the BJP polled more votes than Congress.

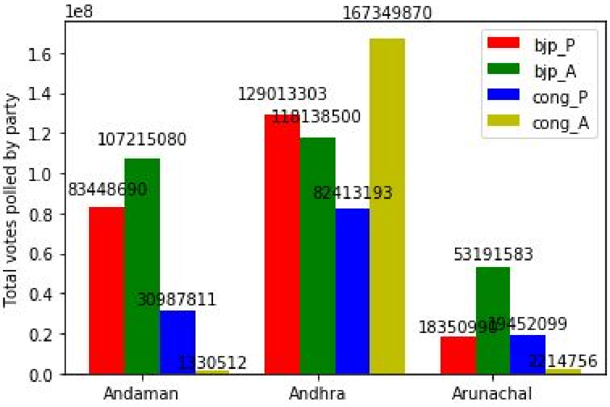


Fig 7 Total Votes polled in Andaman & Nicobar Islands, Andhra Pradesh, Arunachal Pradesh

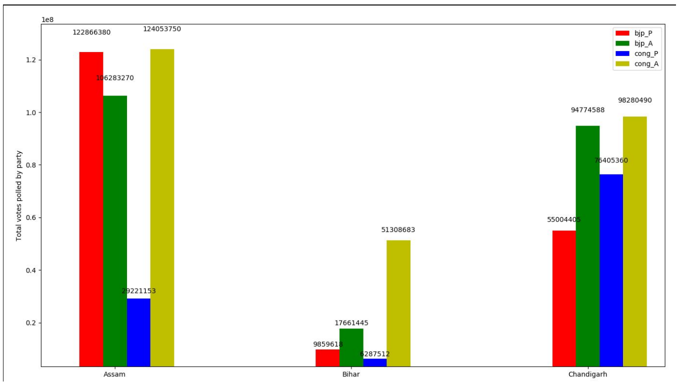


Fig 8 Total Votes polled in Assam, Bihar, Chandigarh

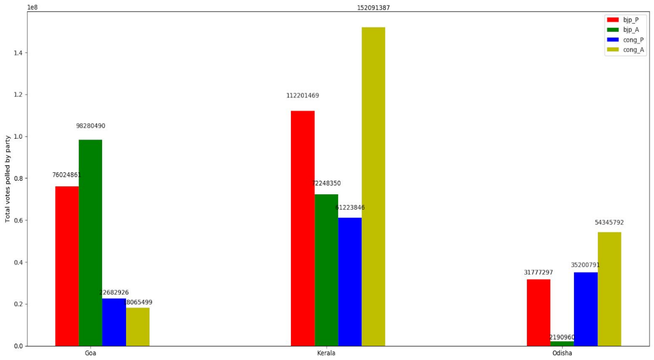


Fig 9. Total Votes polled in Goa, Kerala, Odisha

After training the model on the 2014 data set we tested it for the 2019 dataset and our prediction is depicted below.

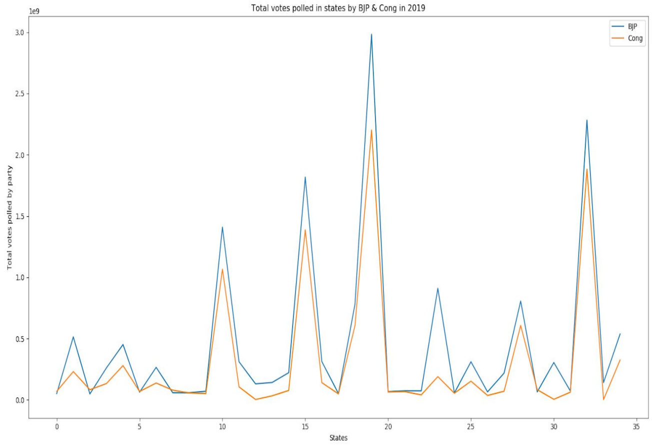
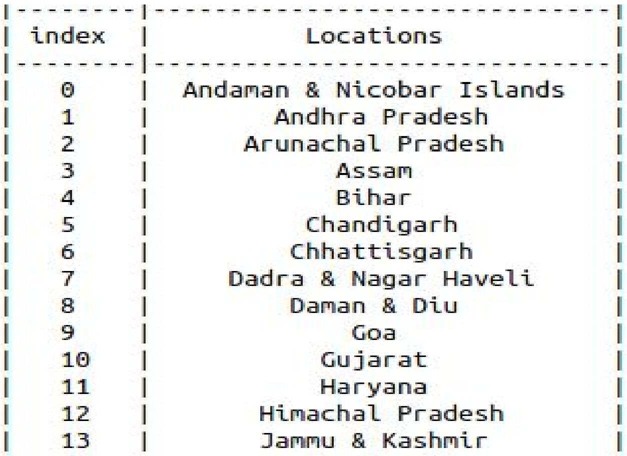


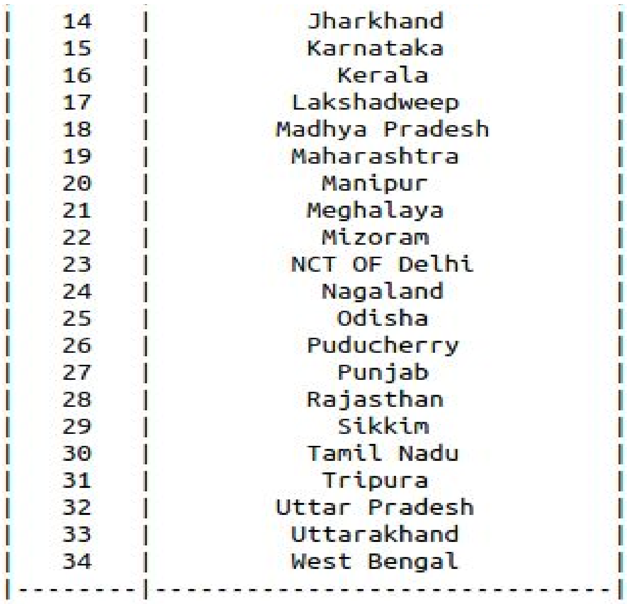
Fig 10. Total Votes polled in Goa, Kerala, Odisha (Refer Table 6)

TABLE VI

LIST OF STATES AND UNION TERRITORIES



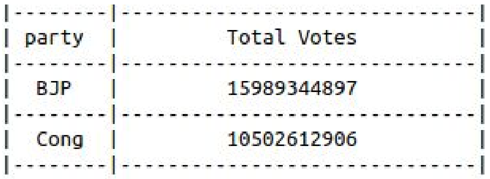
index	Locations
0	Andaman & Nicobar Islands
1	Andhra Pradesh
2	Arunachal Pradesh
3	Assam
4	Bihar
5	Chandigarh
6	Chhattisgarh
7	Dadra & Nagar Haveli
8	Daman & Diu
9	Goa
10	Gujarat
11	Haryana
12	Himachal Pradesh
13	Jammu & Kashmir



```
|   14   |          Jharkhand            |
|   15   |          Karnataka            |
|   16   |            Kerala             |
|   17   |         Lakshadweep           |
|   18   |       Madhya Pradesh          |
|   19   |         Maharashtra           |
|   20   |           Manipur             |
|   21   |          Meghalaya            |
|   22   |           Mizoram             |
|   23   |        NCT OF Delhi           |
|   24   |           Nagaland            |
|   25   |            Odisha             |
|   26   |          Puducherry           |
|   27   |            Punjab             |
|   28   |          Rajasthan            |
|   29   |            Sikkim             |
|   30   |          Tamil Nadu           |
|   31   |           Tripura             |
|   32   |        Uttar Pradesh          |
|   33   |         Uttarakhand           |
|   34   |         West Bengal           |
|--------|-------------------------------|
```

TABLE VII

TOTAL VOTES POLLED BY PARTIES



```
|--------|-----------------------------|
| party  |         Total Votes         |
|--------|-----------------------------|
|  BJP   |        15989344897          |
|--------|-----------------------------|
|  Cong  |        10502612906          |
|--------|-----------------------------|
```

After comparing total votes polled by both the parties (BJP and Congress), BJP will win.

V. CONCLUSION

The use of social media for the prediction of election results poses challenges at different stages. In this paper, the scarcity of training data is tackled for text classification. Finally, a model is proposed for election result prediction which uses the labeled data created using Vader framework. While this model alone may not be sufficient to predict the results, however, it becomes a crucial component when combined with other statistical models and offline techniques (like exit polls).

We implemented the proposed model on a dataset which was created by mining Twitter for 2 weeks. However, this model can be extended in the future to create an automated framework which mines the data for months since election result prediction is a continuous process and requires analysis over long periods of time. Features should be extracted from newly mined data and compared with an existing set of features. Some similarity metric can be used to compare the new and old features. Only in cases where the metric value crosses a threshold, the newly mined data should be labeled using Vader.

Thus we recommend creating an Active learning model wherein the model itself recommends what data should be labeled. This would minimize the efforts for labeling while making sure that there is no compromise on contextual relevance.

References

[1].K. Mao, J. Niu, X. Wang, L. Wang and M. Qiu, "Cross-Domain Sentiment Analysis of Product Reviews by Combining Lexicon-Based and Learn-Based Techniques", 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, 2015.

[2]. Election Result Prediction Using Twitter sentiment Analysis by Jyoti Ramteke, Darshan Godhia, Samarth Shah and Aadil Shaikh

[3]. D. Das and S. Bandyopadhyay, “Labeling emotion in Bengali blog corpus - a fine-grained tagging at the sentence level,” Proceedings of the 8th Workshop on Asian Language Resources, pp. 47–55, Aug. 2010.

[4]. A. Bakliwal, P. Arora, and V. Varma, “Hindi subjective lexicon: A lexical resource for Hindi polarity classification,” Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 1189–1196, May 2012.

[5]. S. Mukherjee and P. Bhattacharyya, “Sentiment analysis in twitter with lightweight discourse analysis,” Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp. 1847–1864, Dec. 2012.

[6]. M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon based methods for sentiment analysis," Comput. Linguist., vol. 37, pp. 267-307, 2011.

[7]. A. Bermingham, and A. F. Smeaton, “Classifying sentiment in microblogs: Is brevity an advantage?,” Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1833–1836, Oct. 2010.

[8]. N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek, “Sentiment Analysis of Hindi Review based on Negation and Discourse Relation,” Proceedings of International Joint Conference on Natural Language Processing, pp. 45–50, Oct. 2013.

[9]. Q. Wu and S. B. Tan, "A two-stage framework for cross-domain sentiment classification, "Expert Systems with Applications, vol.38, pp. 14269-14275, Oct 2011.

[10]. K. Liu and J. Zhao, "Cross-domain sentiment classification using a two-stage method," presented at the Proceedings of the 18th ACM conference on Information and knowledge management, HongKong, China, 2009.

[11]. Q. Wu, S. Tan, H. Zhai, G. Zhang, M. Duan, and X. Cheng,"SentiRank: Cross-Domain Graph Ranking for Sentiment Classification," presented at the Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology Volume 01, 2009.

[12]. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[13]. Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications, and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013