Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 11, November : 2023

# ELECTION RESULT PREDICTION USING SUPERVISED LEARNING ALGORITHMS

Chiravarapu Surya tejaswini[1], & Dr.T.CHARAN SINGH[2], And P. Hymavathi[3]

1Research Scholar, Department of CSE, Sri Indu College of Engineeringand Technology, Sheriguda (V), Ibrahimpatnam(M), RR District –501510, Telangana, India

2Associate professor , Department of CSE, Sri Indu College of Engineering and Technology, Sheriguda (V), Ibrahimpatnam(M), RRDistrict – 501510, Telangana, India

3Assistant professor , Department of CSE, Sri Indu College of Engineering and Technology, Sheriguda (V), Ibrahimpatnam(M), RRDistrict – 501510, Telangana, India

**ABSTRACT**

The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Platforms like Facebook, Twitter and Google+ are being actively used to share ratings, reviews and recommendations. The authors in suggest how 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 Major supervised learning text classification algorithms rely on extracting features from training data set, assigning weights to the features (depending on their frequency or some user criterion) and then using the weighted features to classify test data set.

## 1. INTRODUCTION

The proliferation of social media in the recent past has provided end users a powerful platform tovoice their opinions. Platforms like Facebook, Twitter and Google+ are being actively used to share ratings, reviews and recommendations. The authors in suggest how 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 Major supervised learning text classification algorithms rely on extracting features from training data set, assigning weights to the features (depending on their frequency or some user criterion) and then using the weighted features to classify test data set.

Due to a lack of contextually relevant training set, researchers generally use a cross domain training set for performing text classification as illustrated in. The most common example of this technique is using the popular IMDB data set which consists of 25000 manually labeled movie reviews. This technique however misses out on an important aspect of contextual relevance because the features

extracted from movie reviews need not necessarily match the features of the target data set. Moreover, when tweets come into picture, hash tags themselves become important features. And no other data set can provide hash tags 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 rest of the paper is organized as follows. In section II, we discuss papers which propose a few solutions to the lack of training data. Section III discusses our methodology to create a training data set relevant to elections. Section IV discusses the final machine learning model and the metrics to determine the winning candidate. Finally, the conclusion and future scope is discussed in section.

## 2. PROBLEM STATEMENT

For unsupervised learning, the approaches by Turney, Harb et al. and Taboada et al are discussed. Turney calculates the semantic orientation (Point wise mutual information w.r.t a positive and a negative seed word) of adjectives and verbs in a sentence and determines the overall polarity by adding up the independent values for semantic orientation. They achieved accuracy of 74% by using this technique. Harb et al. used Google search engine to define associations for positive and negative words. They then counted the total positive and negative words to determine the overall polarity of a blog. Taboada et al. used dictionaries of positive and negative words to and integrated intensifiers and negation words to determine the polarity. On an average, 68% accuracy was achieved using this technique. For cross domain sentiment analysis, the approaches by Wu and Tan and Liu and Zhao are discussed. Wu and Tan 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. 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. Liu and Zhao also propose a two-stage framework. At the first stage of their method, they used a feature translator to translate a feature in source domain to a feature in target domain. In the succotash tag clustering. Often while mining data from Twitter, users can find multiple tweets consisting of the same hash tag. For instance, consider the hash tag #MakeAmericaGreatAgain which is the official hash tag for the U.S. Presidential Candidate, Donald Trump.

Now since this is the official hash tag Donald Trump, it is obvious that any person who tweets with this hash tag is in favor of Trump. Hence, all tweets consisting of the hash tag #MakeAmericaGreatAgain must be labeled positively for Trump. So just by associating a label with a hash tag, thousands of tweets consisting of labeled via a code. However, before using this technique, it

is necessary to sort the hash tags in their decreasing order of frequency. This will make sure that higher frequency hash tags get labeled prior to lower frequency hash tags.

Depending on the application, developers or analysts can even choose to not label lower frequency hash tags or hash tags which are ambiguous unlike #MakeAmericaGreatAgain since they will be handled in our next section. The emphasis on manual labeling is bolstered by the fact that a candidate maybe linked to a scam or an initiative. In the case of Benghazi controversy,which is negatively linked to Joe Biden, tweets consisting of #Benghazi cannot be labeled negatively for Biden by using a cross domain data set. Such contextual data pertaining to scams, controversies or political initiatives, which is a quite common in politics, can be handled only by human intervention. Stage 2: Using VADER to label remaining tweets Vader (Valence Aware Dictionary and sentiment Reasoner) 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 developed by Hutto and Gilbert. Vader takes a sentence as input and provides a percentvalue for three categories - positive, neutral and negative and compound (overall polarity of the sentence).

For performing sentiment analysis, a training data set should consist of sentences that are unambiguously either positive or negative. Hence, based on our observations, only those sentences having compound value >=0.8(highly positive) or compound value <=-0.8(highly negative) should be included in the training data set and the remaining sentences can be discarded. Python implementation of Vader is readily available on GitHub as an open-source project. Thus, the two-stage framework proposed above can be used to create a training data set for Twitter.

**LIMITATIONS :**

1. For prediction of sentiment of documents, supervised machine learning approaches are used.

2. The problem of unbalanced dataset in sentiment classification is solved efficiently and appropriately.

3. Naive Bayes classifier seems insensitive to the unbalanced data and gives more accurate results than the support vector machine and K-NN which are sensitive to the unbalanced data. Multilingual sentiment classification is carried out successfully.

## 3. PROPOSED SYSTEM

Now that we have a dataset, and the dataset is labeled in two stages, it can be used to train a supervised machine learning model to perform public sentiment analysis and predict election outcome. We split the dataset in 80:20 ratios to prepare the training and testing sets Based on the metric of F-1 score, we selected the SVM with linear kernel as our entity and sentiment classifier. x The entity classifier gave an accuracy of 0.98 when we used a training data of 50,433 tweets and testing data

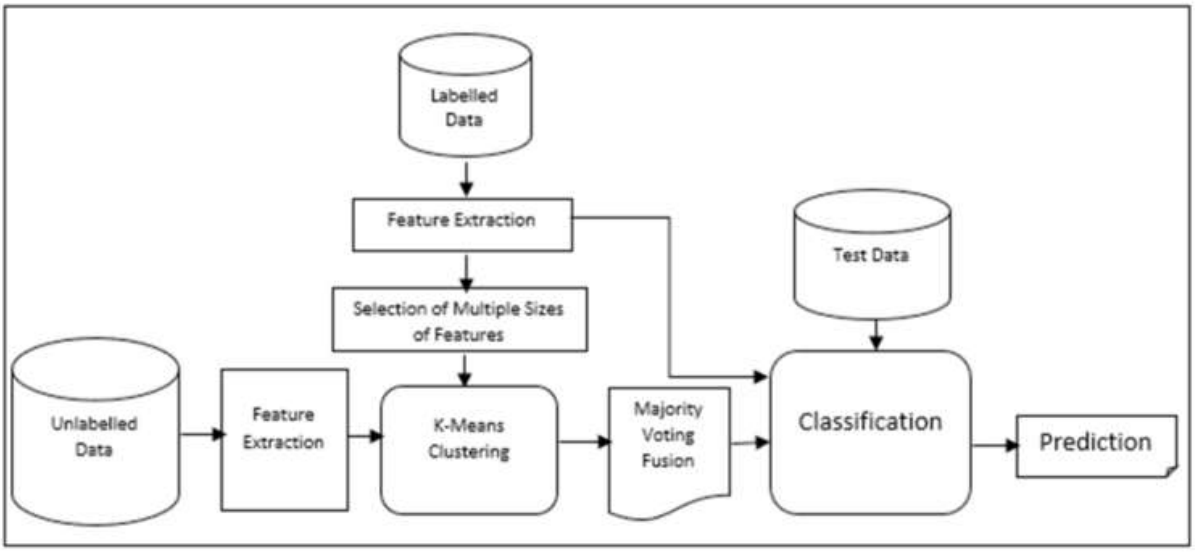
5,603 tweets for classifying 'Joe Biden' and 'Donald Trump' x The sentiment classifier gave an accuracy of 0.99 when we used a training data of 19,589 tweets and testing data of 4,898 tweets of 'Donald Trump' x The sentiment classifier gave an accuracy of 0.97 when we used a training data of 25,240 tweets and testing data of 6,310 tweets of 'Joe Biden'. x the outputs of the testing data from both the sentiment classifiers were as in Table VI C. Aggregation The winner was decided as the person having the higher Positive versus Total count ratio (PvT Ratio), calculated as Ratio = |P| / |T (1) Here, P constitutes the tweets classified to be positive for the candidate (by the candidate's sentiment analyzer), T constitutes all the tweets classified as related to the candidate (by the entity classifier).

Consider a scenario where 50,000 tweets for Donald Trump are mined out of which only 10,000 are positive and 30,000 tweets are mined for Joe Biden out of which 9,000 are positive then direct comparison of positive tweets would yield incorrect results since the percentage of positive tweets for Joe Biden is much higher.

**ADVANTAGES:**

The under-sampling method is complex to classify the sentiments and it is a time-consumingprocess. Supervised methods require excessive quantity of labelled training dataset which are very expensive. It may fail when training data are insufficient.

## 4. SYSTEM DESIGN



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 11, November : 2023

## 5. IMPLEMENATION

### 5.1 Data Collection

Twitter data for two candidates – namely Donald Trump and Hillary Clinton were collected for the dates March 16th, 2016 and March 17th, 2016. We used the Twitter Streaming API to fetch data relevant to the presidential candidates. The Streaming APIs give developers low latency access to Twitters global stream of Tweet data. The input parameters to the streaming functions were the names of Presidential candidates and other keywords like “Democrats”, “Republicans”. Tweets corresponding to the given parameters were returned in JSON format. The JSON result basically comprised of key-value pairs. Some keys were created at, id, re-tweeted, screen name, location etc. The JSON responses were culled to extract only the body of the tweet and stored ina CSV file.

### 5.2 Data Preprocessing

In this stage, the tweets were stripped off special characters like '@' and URLs to overcome noise. Additionally, in the Machine Learning modules, to improve the classifier accuracy, we employ the TF-IDF (term frequency - inverse document frequency) technique, to identify terms which are more relevant to sentiments.

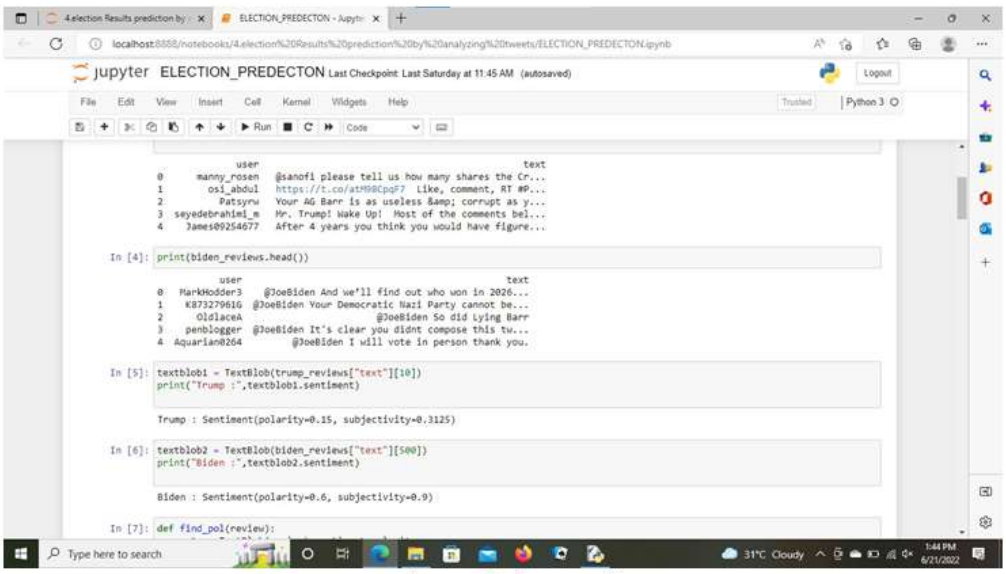
### 5.3 Manual Labeling using hash tag clustering

The first stage of this framework comprises of manually labeling the Twitter data. However, the entire Twitter data set need not be labeled manually. We introduce a technique called hash tag clustering. Often while mining data from Twitter, users can find multiple tweets consisting of the same hash tag. For instance, consider the hash tag #MakeAmericaGreatAgain which is the official hash tag for the U.S. Presidential Candidate, Donald Trump. Now since this is the official hash tag for Donald Trump, it is obvious that any person who tweets with this hash tag is in favor of Trump. Hence, all tweets consisting of the hash tag #MakeAmericaGreatAgain must be labeled positively for Trump. So just by associating a label with a hash tag, thousands of tweets consisting of the same hash tag can be automatically labeled via a code. However before using this technique, it is necessary to sort the hash tags in their decreasing order of frequency. This will make sure that higher frequency hash tags get labeled prior to lower frequency hash tags. Depending on the application, developers or analysts can even choose to not label lower frequency hash tags or hash tags which are ambiguous unlike #MakeAmericaGreatAgain since they will be handled in our next section. The emphasis on manual labeling is bolstered by the fact that a candidate maybe linked to a scam or an initiative. In the case of Benghazi controversy, which is negatively linked to Hillary Clinton, tweets consisting of #Benghazi cannot be labeled negatively for Hillary by using a cross domain data set. Such contextual data pertaining to scams, controversies or political initiatives, which is a quite common in politics, can be

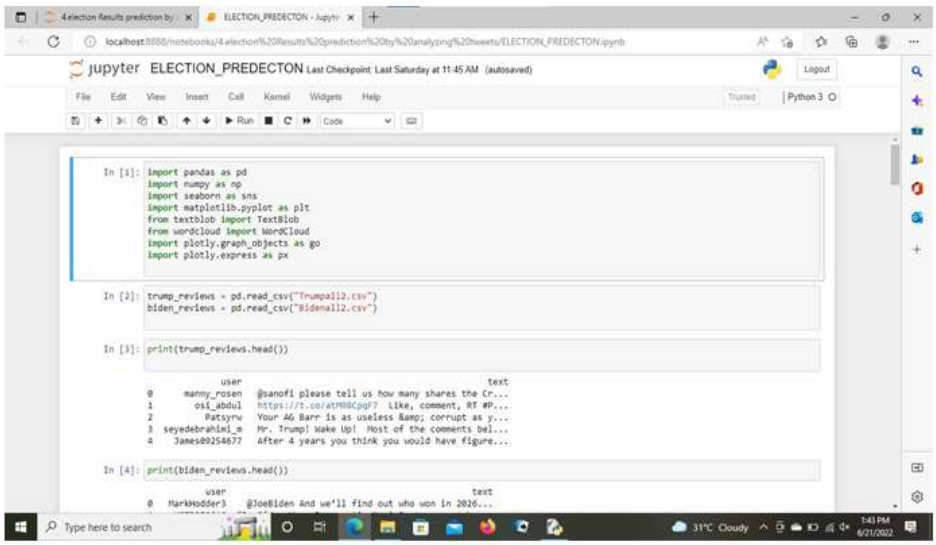
Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 11, November : 2023

handled only by human intervention.

## 6. OUTPUT RESULTS

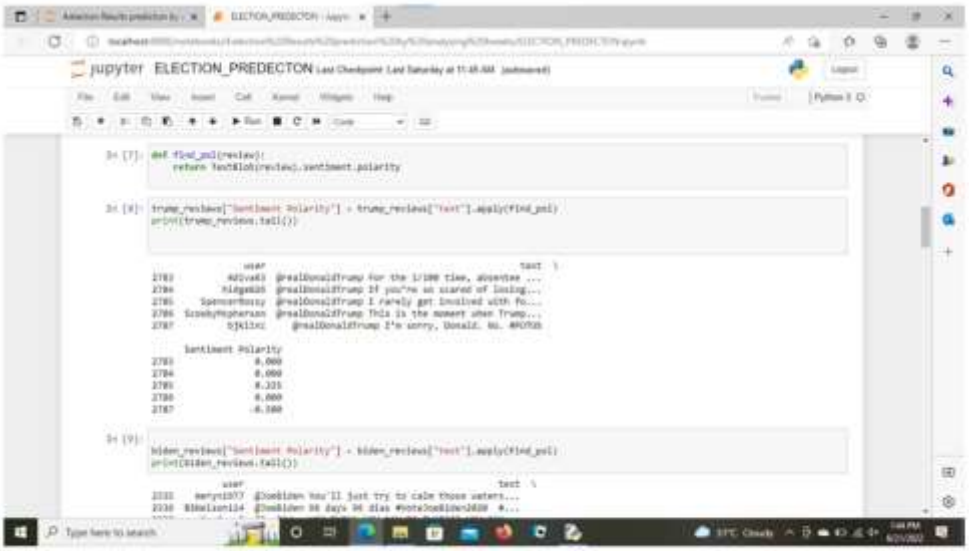



Review Code for JeoBiden

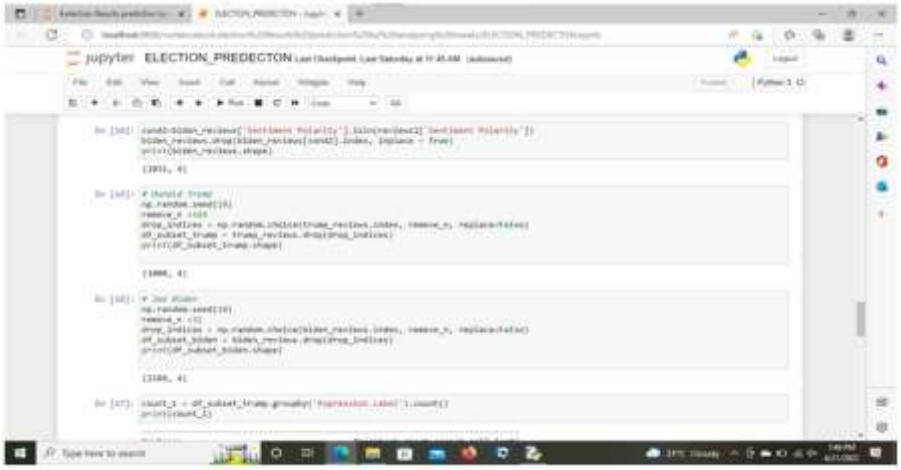



CSV File

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 11, November : 2023

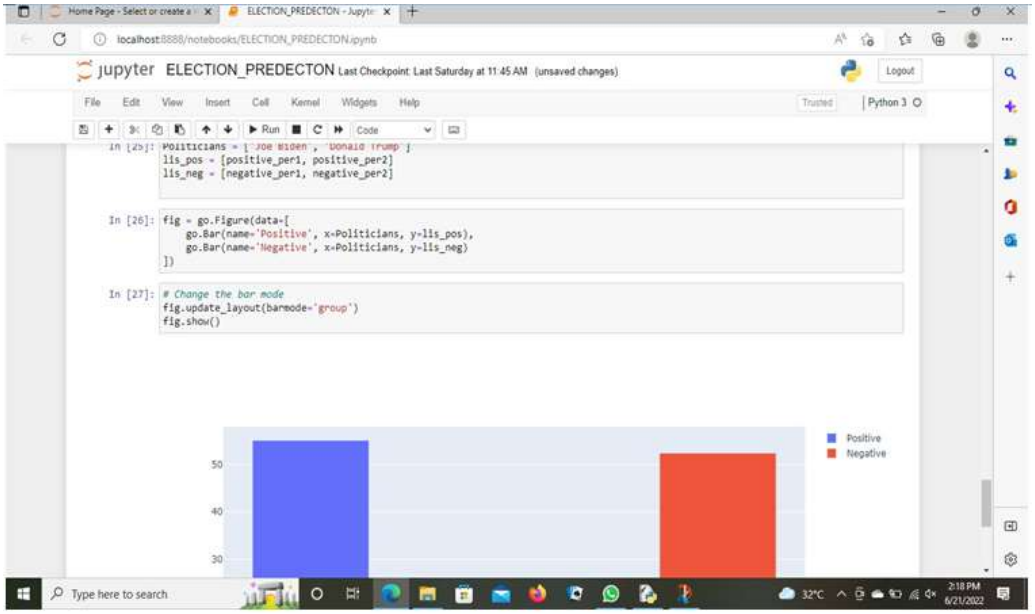



Code for trumph

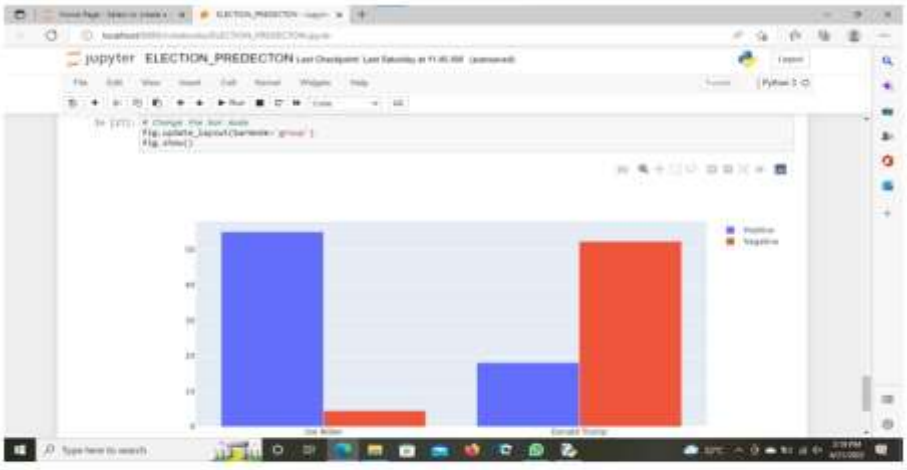



Election prediction

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 11, November : 2023




Sentiment Analysis



Bar Graph

## 7. CONCLUSION

The use of social media for prediction of election results poses challenges at different stages. In this paper, we first tackle the scarcity of training data for text classification by providing a two-stage framework. Finally, we propose our model for election result prediction which uses the labeled data created using our framework. While our 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 3 days. However, this model can be extended in the future to create an automated framework which mines 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 existing set of features. Some similarity metric can be used to compare the new and old features. Only in cases where themetric value crosses a threshold, the newly mined data should be labeled using the two-stage framework. 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.

## 8. FUTURE ENHANCEMENT

In the future, we would like to identify the age of the user, so that during data filtering process we can eliminate the tweets that come from twitter handles that have an age of less than 18. Dividing our corpus-based on gender, caste and community based on user identification algorithms can help us to analyse voting patterns in India better. In this study we demonstrate a text mining and sentiment analysis framework for finding most popular topics about contesting parties discussed on Twitter and classifying the tweets under the assumption that each tweet is mixture of weighted and labelled topics. We believe that this study contributes towards new research possibilities in the field of election prediction using social media data. Nevertheless, there are few limitations to our study. First of all, the data size isreduced because of geo tagging 20. Collected Twitter data may not completely represent voting population. Specifically, rural regions are not represented by Twitter and other social media websites. Secondly, small and regional parties are not popular on Twitter, therefore making the prediction results slightly less accurate. Finally, the sentiment towards a political party fluctuates heavily, hence for best results; we opt to collect data immediately before elections and also during month long elections. To improve twitter-based sentiment analysis in future, researchers can use domain specific lexicons. This study can be extended to monitor elections, for instance Obtm can be used to find positive and negative topics using a twitter stream. Hence, regularly recording sentiment scores, volume, and other variables to find strengths and weaknesses of a contesting party. We can easily extend vote share prediction to seat share prediction by estimating electoral swing 7. Although, it is challenging to apply geo-tagging for each constituency to predict elections 20, results obtained from this methodology can be more reliable and can match up to traditional polls

## 9. REFERENCES

[1] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), may 2010.

[2] Y. Yang and F. Zhou, "Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification", 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES) ,2015.

[3] M. Haj Mohammadi, "LACK OF TRAINING DATA IN SENTIMENTCLASSIFICATION: CURRENT SOLUTIONS", IJRCCT, vol. 1, no. 4, pp. 133-138, 2012.

[4] 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.

[5] P. D. Turney, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania,2002.

[6] A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset, and P.Poncelet, "Web opinion mining: how to extract opinions from blogs?," presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary scienceandtechnology, Cergy- Pontoise, France, 2008.

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

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

[9] K. Liu and J. Zhao, "Cross-domain sentiment classification using atwostage method," presented at the Proceedings of the 18th ACMconference on Information and knowledge management, HongKong, China, 2009.

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

[11] "The Streaming APIs | Twitter Developers", dev.twitter.com, 2016. [Online]. Available: https://dev.twitter.com/streaming/overview. [Accessed: 25- Apr- 2016].

[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 socialmedia (ICWSM-14). Ann Arbor, MI, June 2014.