International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Empirical Study of Sentiment Analysis of Multi-Lingual Real Time Tweets for Election Forecasting

Ayush Chhajer[1], Ajam Bhandari[2], Sudhanshu Ranjan[3], Patrick D'souza[4], Prof. Rachana Sable[5]

Student, Department of Information Technology Engineering, G.H. Raisoni Institute of Engineering and Technology, Wagholi, Pune, India[1,2,3,4]

Professor, Department of Information Technology Engineering, G.H. Raisoni Institute of Engineering and Technology, Wagholi, Pune, India[5]

ABSTRACT: Due to the properties of social media being interactive in nature and real time, it has grabbed much attention which makes it best suited for political analyzing related work. Recent research have probed that social media platforms can record the mindsets of netizens which in turn are the valuable assets to be considered for elections. A large amount of raw data is generated by these social media platforms that can be used for making decisions which can be turning point in election.

KEYWORDS: sentiment analysis; prediction; elections; Twitter; Regression.

I. INTRODUCTION

Image Social media platforms like Twitter, Facebook, etc. are increasingly being used to express their opinions and interests by people. This has piqued the interest of businesses and other social entities, which are employing very sophisticated tools to gather and make use of the large amounts of data available on these social media platforms. "Big data" is referred to as the huge amount of data generated by these platforms. After scrutinizing a lot of research related to election prediction, a survey paper is created in which every work related to psephology using social media is incorporated.

Sentiment analysis of the data obtained from social media a platform that 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.

Our prediction for the maximum livelihood of a party winning will be done by our proposed machine learning model. We mined data from Twitter. Mined data was cleaned and relevant data was fed into Text Blob for sentiment analysis. The polarity of tweets obtained will be used in many models. Real time data pertaining to the Elections was used for training the model, which was then used to predict the results of the 2019 elections.

Prediction is done and accuracy of the model is obtained by comparing the results with the values of the votes obtained by the two parties in polls.

CONTRIBUTION

The work provides a brand 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 will be tackled as one of Regression where a Machine Learning has been used to compare the correlation between sentiment expressed on Twitter and the number of votes a party receives. Moreover tweets, hashtags are the important features which are mined from Twitter for that specific application. Hence, it is mandatory to implement a labeling technique for the mined Twitter data which can make a balance between accuracy and speed along with all of this we've also added the feature of generalization and Graphical User Interface to our system.

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. 'Conclusion' section summarizes the work done.

II. RELATED WORK

The major portion of connected corpus focuses on the method information|ofknowledge|of information} Gathering and subsequent Sentiment Analysis of preprocessed data. In cross-domain dataset has been accustomed train the sentiment analysis model, to create up for the shortage of availableness of contextually relevant information. In these product reviews (15166), however web and NTUSD, Chinese Original Basic Lexicon (OBL) datasets ar used. It focuses on options classification into four categories: Sentiment words (Number of positive words, negative words in document, total of P and N, distinction of P and N, look variety of sentiment words), n-Gram(Number of 2-Grams in document), applied mathematics data (Number of question marks, exclamation marks, negations in document, Length of document,Number of disjunctive conjunctions,adjectives in document) and Results (Score, variety of sentences judged to be positive, negative & neutral) of lexicon-based methodology options. It's accuracy for Hotels: H (0.858), Books: B (0.929), Electronic devices: E (0.816). It are often created not inferior to the domain of natural philosophy. [1]

The authors discuss however gathered information are often effectively labeled and ready for classification. any comparison of various classification models has been provided. during this Twitter dataset (30,000 tweets for the coaching information set) is employed. It focuses on Hashtags options. It's accuracy for MNB (nltk)-0.54 MNB (Scikit-learn)-0.97 SVM (nltk)-0.58 SVM (Scikit-learn)-0.99. we will use such algorithmic program packages which offer a lot of accuracy for classification. [2]

In information in Bengali corpus has been classified in six feeling categories and to annotate the sentences. It uses Twitter information as dataset. It focuses on Annotators from web log as options. In future, it will adapt a corpus-driven methodology for building a lexicon of feeling words and phrases and extend the feeling analysis tasks in Bengali language. [3]

Authors haven't thought of the emojis that ar a relevant facet once explaining the polarity of a tweet. Since information was labeled manually the dimensions i.e. 36,465 wasn't large enough to produce a lot of accuracy, thus we will fetch a lot of tweets and label them. It uses Twitter information as Dataset. It focuses on words and pharses that are taken out from the tweets ar contemplate as options. It's accuracy generated were Naïve Thomas Bayes (62.11%), SVM (78.42%), lexicon based mostly (34.1%). [4]

Authors uses dictionaries of excellent and dangerous words to see the polarity. It uses Twitter dataset (9021 tweets from twenty one totally different domains) and Twitter dataset (16,454 tweets). Generic lexicon within the language unit house are often use in future for classification, wherever it cannot distinguish between the assorted senses of a word. an idea enlargement approach, to expand the feature vectors, might persuade be helpful. this is often thanks to the intensive world data embedded within the tweets. [5]

Authors Taboada, Brooke, and Stede justify the model that integrate pack of-words in speak markers with the slant demand by five-hitter exactitude. It focuses on weight, multiple cut-offs as options. It uses Polarity Dataset. In future it's not increased that may are done. With multiple sources of information (Taboada, Brooke, and Stede 2009). [6]

Authors suggested depicting Hindi reviews as positive, neutral, negative. They crack another score smashing purpose and used it for 2 totally different ways. Moreover, fusion takes place between the POS labelled Ngram and central N-gram. Twitter information is employed as dataset. [7]

Authors seek 2 non-identical methods, Multinomial Naïve Thomas Bayes (MNB) and SVM. They found that MNB methodology surpasses SVM on scaled scale areas with short meterial. Twitter information is employed as dataset. Its accuracy is seventy four.85 %. Disadvantage is we will take solely one thousand documents per category in line with the flick review corpus. This allow us to to get rid of any primary sentiment angle which can be gift. [8]

To recognize the sentiment from Hindi content, Authors offerrise to an well organised approach. By adding more opinion words, they developed Hindi language entity and improve the present Hindi Senti Word Net (HSWN). 80% precision has been measured by them. It uses annotated dataset. It focuses on words in the document as features. Its accuracy is 80.21%. Disadvantage is that the dataset is not extended for the better and generalized results. [9]

Authors equate quantities of public opinion estimated from survey with sentiment computed from text. They scrutinize some studies on public faith and ministerial ideas from year 2008 to 2009, and discover that they collectively agreed to sentiment idiom frequencies in coetaneous Twitter messages. While our outcome differs over datasets, in individual instances the mutuality are high about 80%, and seize noteworthy extensive drift. The outcome focuses the prospects of text streams as a replacement and extension for traditional ballot. [10]

Authors come up with an interpreted record on the subject of electoral prediction using Twitter data. The order is sequential, In the middle of papers there are cross-references, and few of them are just assigned for the advantage of comprehensiveness. In addition to that, seminal papers not entirely related to the topic are included plus papers on associated sectors such as integrity, gossips, and anthropology of twitter users. [11]

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

Authors inspected over 100,000 tweets, bring up political organizations in advance to the German federal election 2009. All-inclusive, we observe that Twitter is as a matter of fact used as a podium for state meditation. The bare number of tweets comes adjacent to traditional opinion polls and revolves voter fondness, while the emotion of tweets intimately equivalent to political agendas, applicants portrait, and affirmation from the media covering scope of the offensive series. [12]

Authors coated some a hundred and fifteen studies on the employment of Twitter in politics. For this discussion, studies area unit sorted in 3 topical categories: studies addressing the employment of Twitter by politicians and campaigns; studies addressing the employment of Twitter by varied publics throughout election and issue campaigns; and comments on Twitter throughout campaign events–such as televised debates, party conventions, and polling day coverage. [13]

Authors, during this work, we've studied the employment of twitter by house senate and politician candidates throughout the midterm (2010) elections within the U.S. our information is comprehensive of virtually 700 candidates and over 690k documents that they made and cited within the three.5 years resulting in the elections. we've used graph and text mining techniques to research variations between Democrats, Republicans and party candidates, and recommend a completely unique use of language modelling for estimating content cohesiveness. Our findings have shown vital variations within the usage patterns of social media, and recommend conservative candidates used this medium additional effectively, conveyancing a coherent message and maintaining a dense graph of connections.

Despite the shortage of party leadership, we discover party members show each structural and language-based cohesiveness. Finally, we tend to investigate the relation between network structure, content Associate in Nursingd election results by making a proof-of-concept model that predicts candidate conclusion with an accuracy of eighty eight.0%.[14]

Authors operate the current Irish General Election as a chronology for exploring the capability to represent ministerial emotions by way of excavation of social media. Our perspectives unite emotion studies using supervised learning and volume-based initiatives. We appraise in opposition to the stereotypical public opinion polls and the final election outcome. [15]

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 according to specific lingual.

In this text - sentiment analysis, although a very important phase, has been looked at like the first step in preparing a machine learning model. Gathering of tweets happened.

Tweets were used so that sentiment analysis could be performed, the overall positive, negative, neutral sentiment towards the two major parties, BJP and Congress, is determined by assigning a score to the three polarities. Then a Regression model will be established how the positive, neutral and negative sentiment as expressed by netizens on Twitter relates to 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 selections that needed to be made. The Twitter Developer platform may be a powerful tool that gives different approaches consistent with the goals and demands of every 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 (BharatiyaJanata Party) and Congress (Indian National Congress) were collected.

We've generalized a system to predict either products or brands sentiment among people. With the help of which users can predict for the review of movie, for the review of electronic brands, for the review of an automobiles, etc.

Table I: Tweets, Likes &Retweets for BJP and Congress

Party	Tweets	Likes	Retweets
BJP	200	8452	16480
Congress	200	10858	14645

B. *DATA PRE-PROCESSING*

Real-world data is usually incomplete, inconsistent and/or lacking in certain behavior/trends and is probably going to contain many errors. The search API returns a huge volume of metadata for each tweet. The fields that we were

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

interested in were Tweet's ID, Date posted, Tweet's Text, Tweet's length, Tweet's Source, Tweet's retweets and Tweet's likes.

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

Data Cleaning - This process included 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.

C. DATA LABELING

TextBlob may be a Python (2 and 3) library for processing textual data. It provides an easy API for diving into common tongue processing (NLP) tasks like part-of-speech tagging, phrase extraction, sentiment analysis, classification, translation, and more. It is specifically attuned to sentiments expressed in social media. It is basically a sentiment intensity polarizer. TextBlob 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 tweet features like Tweet's ID, Date posted, Tweet's Text, Tweet's length, Tweet's Source, Tweet's retweets and Tweet's likes.

Following figures shows an analysis of tweets which gives us the information of likes, retweets, tweet id, tweet length, date of tweet posted, source of tweets, and sentiment of tweets for BJP and Congress.

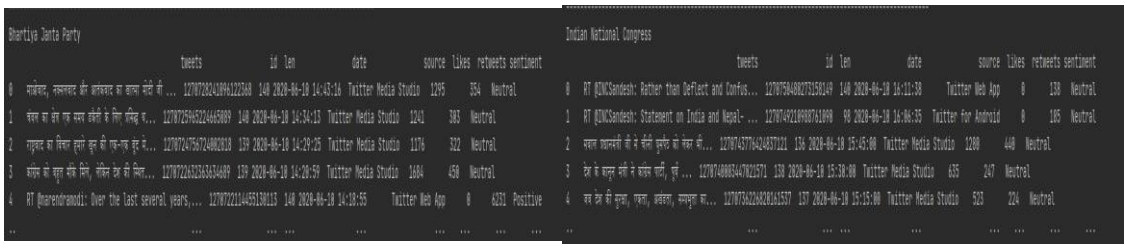


Fig 1.BJP AnalysisFig 2. Congress Analysis

As our system is generalized we can use it to predict the trend of a particular product or brand. So here is an example of our generalized system, we've analyzed the tweets related to 3 idiots movie.

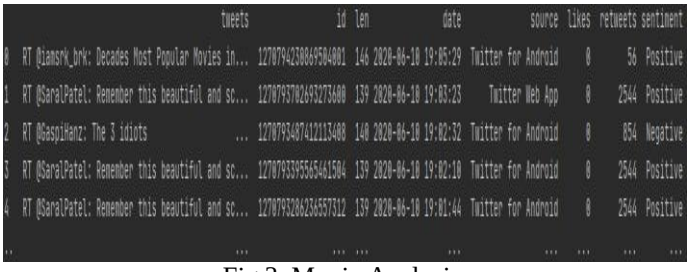


Fig 3. Movie Analysis

Table II: Text Blob Sentiment Analysis Examples

Sentence	Comp	Pos	neg	neu
He is smart and funny	0.89	0.85	0.0	0.273
A horrible Book	-0.84	0.0	0.78	0.19
It sucks, but I'll be fine	0.23	0.265	0.189	0.59

The above table provides three examples of sentences analyzed using TextBlob. 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

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

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

We are using TextBlob which is an already trained model for performing sentiment analysis on the dataset. Our dataset is nothing but an text file named as "Tweets.txt" where we store real-time tweets related to two major parties BJP and Congress. The real-time dataset was used for our system to work on.

This data contains the tweets related to what is being predicted and we are usingsTextBlob which analyses the sentiments of each tweet and also gives us the information of like, count of retweet, id, date of publication, source, and length of each tweet. TextBlob gives us original data removing all the duplicate data or duplicated tweets or retweets.

We also have a dataset in which there is data of two political party polarities which is describing the sentiments of people related to past events done by our indian political parties for a particular event that affected our nation.

E. *Design*

The proposed model can be divided into 4 main stages on the basis of the nature of the task to be performed.

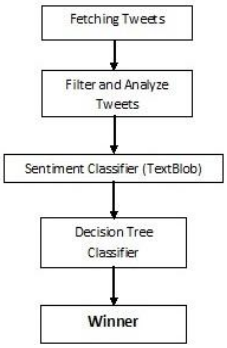
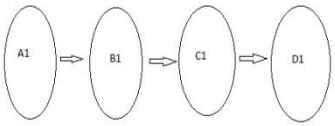


Fig 4. Model for Election Prediction

Sentiment Analysis phase involves attributing the data with its associated polarity value.Final data set created after sentiment analysis and data labeling will be fed into a Decision Tree Classifier. The Decision Tree Classifier is trained on the data collected for the election prediction.

Once the sentiment analysis is done for two particular parties like BJP and Congress in our case then these result of polarity is fed into the Decision Tree Classifier whose accuracy is far much better than linear regression.

Mathematical Model



Where,
A1= Establishing connection with twitter database
B1 = Analysis of tweets
C1 = Assigning Polarity
D1 = Comparing Result

Set Theory :
S={s,e,X,Y,F}
S=Set of system
s=start of the program
X=Input of the program
X={A1}
Where A1={Access Token,AccessSecret,Consumer key ,Consumer Secret}

F={B1,C1}
Where F= Set of functions that are implemented to get output
Y = Output of program
Y= {D1}
Where D1={Comparing results among parties and declaring winner}
e=End of program
So, in this case the problem is p class problem

F. IMPLEMENTATION

To implement the Supervised Decision Tree Classifier Model which will give us better accuracy, the performance of the Decision Tree Classifier and Simple Linear regression were analyzed. After analyzing various machine learning algorithms we discovered that Decision Tree Classifier model gives better accuracy than other algorithms. So we decided to use Decision Tree Classifier regression model for comparing and testing the result.

We've generalized our system for the things which can be predicted such as you can predict for a movie to check the review and decide whether to watch it or not, you can also predict the products popularity among people by understanding the sentiments.

G. GUI REPRESENTATION

For the ease of making our system user friendly we have built GUI. GUI is nothing but a graphical user interface through which user can be more comfortable with the machine and it is a best form of representation. Due to GUI a user is expected to use system more.
Following figures shows the representation of UI.

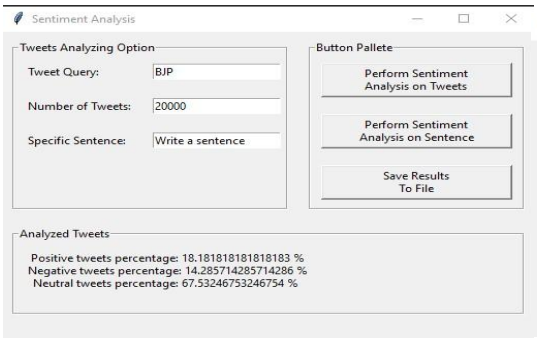
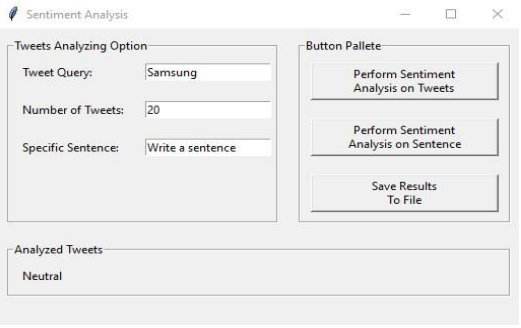


Fig 5.UI – IFig 6. UI – II



IV. RESULTS AND DISCUSSIONS

The TextBlob Model is trained on a dataset which gives us features to do task related to NLP i.e. natural language processing.After comparing the accuracy it is clear that Decision Tree Classifier model which is a binary classifier have is way better and optimum than other machine learning algorithms. The polarity dataset of previous elections is used for training and testing of our model which is being used on our real-time polarity. The Decision Tree Classifier model is used to compare and give us the result which is giving the most probable chances for a political party to win the upcoming election.

After the prediction of chances of winning the election by parties is made by our model, the predicted and actual polarity is compared and Accuracy Score is calculated.Linear Support Vector Machine (Linear SVM) is widely regarded as one of the best text classification algorithm. We achieve a higher accuracy score of 79% which is 5% improvement over naïve bayes algorithm according to our research.

Features are extracted from newly mined data and compared with an existing set of features.

We also measure the accuracy for Decision Tree Classifier model and linear regression model by implementing then we found out that the Decision Tree Classifier algorithm gives us much better accuracy over linear regression algorithm by enhancing the accuracy score by 5-15% that is 78-88%.The model visualization of our system's Decision Tree Classifier model, which helps us to predict the winning of a particular political party.

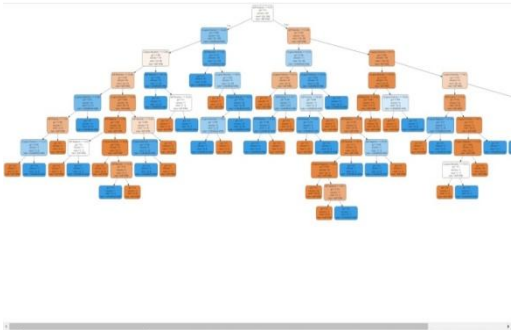


Fig 7. Model Visualization of Decision Tree Classifier Model

Following graph(7) shows number of Tweets likes and retweets of BJP and Congress. Following graph(8) shows the comparison between polarities for BJP and Congress.

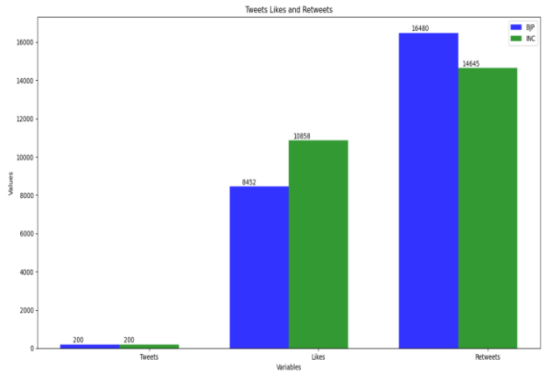


Fig 7.Number of Tweets, Likes & Retweets of Parties

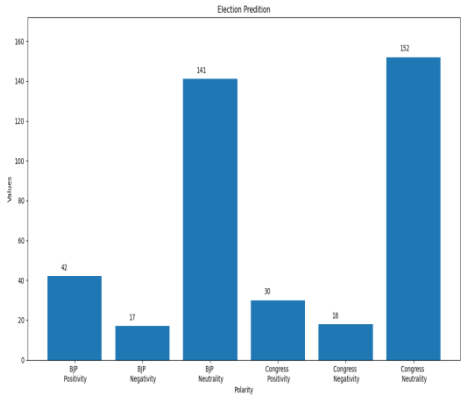


Fig 8. Comparison between Polarities for BJP and Congress

The graph below is a pie chart which shows the percentage of positivity, negativity and neutrality of tweets for two major parties BJP and Congress.

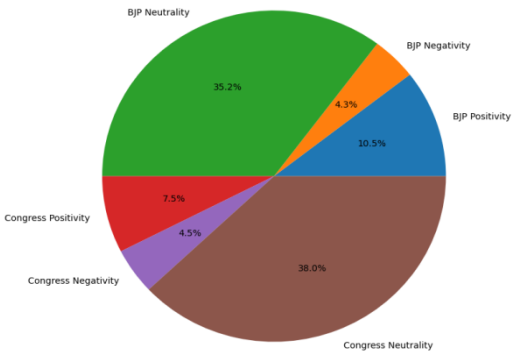


Fig 9.Pie chart showing polarities for BJP and Congress

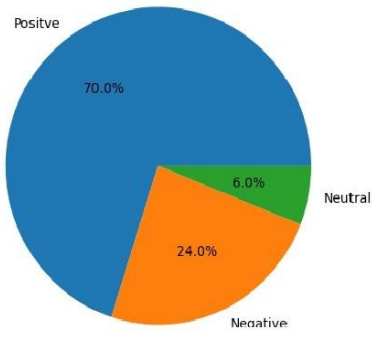


Fig 10. Pie chart showing polarities for movie 3 Idiots

Since our system is generalized, here's the proof supporting our statement. For example you can check for the reviews of 3 idiots movie.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

V. ACKNOWLEGEMENT

The research and implementation of this project was guided by Prof. Rachana Sable who provided their experience and their knowledge that was a great resource of knowledge. We are also thankful to Prof. JyotiChauhan and Prof. Ganesh Kadam for their insights that guided us to improve the quality of our work.

VI. FUTURE SCOPE

The use of social media for the prediction of election results poses challenges at different stages. Finally, a model is proposed for election result prediction. 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).

Our system can be further enhanced by stating how much a particular party is expected to get seats in the parliament.

Our system can also be enhanced to deal with sarcasm and figures of speech which are present in the text.

Our system can be enhanced by adding the concept of active learning so as to not miss out on any sentiments of the people which are the critical stakeholder for determination of a political party winning the elections.

To identify features and provide analysis for each aspect based on it.

To handle positive sentences which contain a negative word.

VII. CONCLUSION

The use of social media for the prediction of election results poses challenges at different stages. Finally, a model is proposed for election result prediction. 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 our model on a dataset which was created by mining Tweets polarity. 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.

Netizens use social media platform to express their emotions for a particular event. These emotions are used for psephology. This helps us to predict the maximum likelihood for a party to win election. It also helps us to analyze how events happening in a particular geographic location ends up affecting a party standing for election.

Thus we recommend that we should be using recurrent neural network along with reinforcement learning so that our system will be capable enough to handle sarcasm, aspect based analysis, negation handling.

REFERENCES

[1].Cross-Domain Sentiment Analysis of Product Reviews by Combining Lexicon-Based and Learn-Based Techniques, by K. Mao, J. Niu, X. Wang, L. Wang and M. Qiu (2015)

[2]. Election Result Prediction victimization Twitter sentiment Analysis by JyotiRamteke, DarshanGodhia, Samarth sovereign and AadilShaikh.(2016)

[3].Labeling feeling in Bengali journal corpus - a fine-grained tagging at the sentence Level by D. Das and S. Bandyopadhyay(2011)

[4] Prediction of Indian Election victimization Sentiment Analysis on Hindi Twitter by Parul Sharma, Teng-Sheng Moh(2010)

[5] Sentiment Analysis in Twitter with light-weight Discourse Analysis by Subhabrata Mukherjee, PushpakBhattacharyya(2009)

[6] Lexicon-Based ways for Sentiment Analysis by MaiteTaboada, solon Rupert Brooke, Kimberly Voll, Manfred Stede(2012)

[7] Classifying Sentiment in Microblogs: Is Brevity a plus By Adam Bermingham& Alan Smeaton(2010)

[8] Sentiment Analysis of Hindi Review supported Negation and Discourse Relation by N. Mittal, B. Agarwal, G. Chouhan, N. Bania, and P. Pareek(2005)

[9] From Tweets to Polls: Linking Text Sentiment to opinion statistic by Brendan Flannery O'Connor, RamnathBalasubramanyan, Bryan R. Routledge and Noah A. Smith(2010)

[10] “I needed to Predict Elections with Twitter and every one I got was this Lousy Paper”. A Balanced Survey on Election Prediction victimization Twitter information by Daniel Gayo-Avello.(2011)

[11] Predicting Elections with Twitter: What a hundred and forty Characters Reveal regarding Political Sentiment by AndranikTumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe(2011)

[12] Twitter in Politics: A Comprehensive Literature Review by Andreas Jungherr (2014)

[13] The Party is Over Here: Structure and Content within the 2010 Election by AvishayLivne , Matthew P. Simmons , Eytan Adar and Lada A. Adamic (2011).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 6, June 2020||

[14] On victimization Twitter to watch Political Sentiment and Predict Election Results by Adam Bermingham and Alan F. Smeaton (2012).
[15] Predicting elections from social media: a threecountry, three-method comparative study by KokilJaidka, Saifuddin Ahmed, Marko Skoric& Martin David Hilbert (2018).