IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-10, May 2019 E-ISSN: 2347-2693

A Proposed Method for Predicting Indian Loksabha Election Using Machine Learning in Social Media

Deeksha Vishnoi[1*], Brajesh Patel[2]

[1,2] CSE, SRIT, RGPV, Jabalpur, India

Corresponding Author: deekshavishnoi11@gmaill.com,

DOI: https://doi.org/10.26438/ijcse/v7si10.135143 | Available online at: www.ijcseonline.org

Abstract— Sentiment analysis is the computational study of opinions, sentiments, evaluations, attitudes, views and emotions expressed in text. It refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Sentiment analysis over Twitter offers people a fast and effective way to measure the public's feelings towards their party and politicians. The primary issues in previous sentiment analysis techniques are classification accuracy, as they incorrectly classify most of the tweets with the biasing towards the training data. Sentiment analysis is considered to be a category of machine learning and natural language processing. It is used to extricate, recognize, or portray opinions from different content structures, including news, audits and articles and categorizes them as positive, neutral and negative. It is difficult to predict election results from tweets in social media using different platforms.
We performed data (text) mining on thousands of tweets collected over a period of a month that referenced five national political parties in India, during the campaigning period for general state elections in 2018. We made use of both supervised and unsupervised approaches. We utilized Dictionary Based, Logistic Regression algorithm as the main algorithm to build our classifier and classified the test data as positive, negative and neutral. We identified the sentiment of Twitter users towards each of the considered Indian political parties. The result of the analysis was for the BJP (Bhartiya Janta Party). Proposed algorithm predicted a chance that the BJP would win more elections in the general election. Therefore, here we adopt a lexicon based sentiment analysis method, which will exploit the sense definitions, as semantic indicators of sentiment. Our method also uses a negation handling as a pre-processing step in order to achieve high accuracy.

***Keywords*:** Negation Handling; Sentiment Analysis; WordNet; SentiWordNet; Word Sense Disambiguation.

I. INTRODUCTION

In the modern society, the international relations have changed enormously, which makes the world to undergo a profound reform and enter the new stage of economic turbulence. Both of traditional security and non-traditional security threats coexist, so the major powers and emerging forces pay close attention to adjust the domestic and international policies to adapt to such changing. Under this circumstance, the competition of comprehensive national strength becomes more intense. For example, the change of each state regime, the replacement of the ruling party and the leadership of the general, they will generate great impacts on the whole world's situation. Therefore, it has great significance of grasping the changes of the international pattern in developing the diplomatic strategy and handling the international situation.

Generally speaking, a policy is proposed by the House of Representatives and voted by the Senate. Therefore, it has great value to predict the election of Senate and House of Representatives (SHR). Twitter is one of the largest social networks, providing a friendly platform for people to express opinions and share views on a variety of topics and issues. Prediction based on Twitter data analysis has drawn much attention in recent years, especially in predicting results of political events. In 2015, the mainstream media polls were wrong in the prediction of the UK general election. Traditional polling methods for election prediction analyse data from questionnaires by phone call or pedestrian survey and are usually biased in sampling and prediction process as well [1]. Nowadays, social networks provide valuable information for predicting outcomes of social or political events, which may not be obtained from the mainstream media or traditional polls. For example, many people supported Trump in the US presidential election 2016, but they might not be willing to say so in public for some reasons. Thus, prediction based on social media analysis can result in new outcomes, complementary with or even more accurate than traditional prediction poll results. In Twitter based election prediction it is critical to extract informative keywords or features reflecting true sentiment of voters. In addition, traditional prediction models may not be suitable for the data from social networks. In this thesis, a new

method for election prediction based on Twitter data analysis is proposed and applied to predict the 2019 Indian Loksabha Election.

Due to the rapid growth of the Internet and online activity, Social Media are popular as these services allow users to share information and express opinions on specific topics. Social Media are one of generating sources of Big Data and it can offer business insights by analyzing the public opinions. Twitter is one of the popular social media, which combines features of blogs and social network services. Twitter was established in 2006 and experienced rapid growth of users in the first years. Twitter has 330 million monthly active users [2] and they posts 500 million tweets every day. Twitter is a good source of information in the sense of snapshots of opinions and feelings as well as up-to-date events and current situation commenting. The best solution to identify the opinions of people is Sentiment Analysis. The sentiment analysis methods can be divided into two kinds: machine learning method and lexicon based method. The lexicon based method is performed by calculating the weight of the relevant words in the dictionary. The classification of a text depends on the total score it achieves. In the machine learning approach the task of Sentiment Analysis is regarded as a common problem of text classification [3] and it can be solved by training the classifier on a labeled text collection. The machine learning approach applicable to Sentiment Analysis mostly belongs to supervised classification. A number of machine learning techniques have been adopted to classify the reviews.

Current Sentiment Analysis solutions and researches are needed to be scale up to perform well on social data. Big Data Analytics has become popular for analyzing and managing large volume of social data. Google introduced the Map Reduce paradigm [4] for parallel and distributed execution of an application over the commodity cluster. Several systems had implemented Map Reduce paradigm for parallel and distributed processing of batch data on multiple machines.

In this paper, Sentiment Analysis system is proposed to extract up to date valuable information from Social Data. To achieve the promising accuracy in Real-time multiclass Sentiment Analysis, is implemented by combining lexicon and learning based sentiment classification with Multi-tier architecture. The comparison of three machine learning based techniques (Naïve Bayes, Linear SVC, Logistic Regression) is performed to find the suitable classifier SA.

1.2 BACKGROUND OF SENTIMENT ANALYSIS

Sentiment analysis, also called opinion mining, is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term sentiment analysis is more commonly used, but in academia both sentiment analysis and opinion mining are frequently employed. They basically represent the same field of study. The meaning of opinion itself is still very broad. Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments. To do an analysis, classification plays a key role in opinion mining. A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations.

Opinions are central to almost all human activities because they are key influencers of our behaviours. Whenever there is a need to make a decision, others' opinions are required. In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies.

Opinion summarization summarizes opinions of articles by telling sentiment polarities, degree and the correlated events. With opinion summarization, a customer can easily see how the existing customers feel about a product, and the product manufacturer can get the reason why different stands people like it or what they complain about. A seller's job can be quite complicated or it can be quite easy. The two contradictory terms define the selling experience, based on the fact as how seller interprets the consumer interests. Unless one is a psychic or knows how to get into others mind the actual demand of the customer's and the product can't be collaborated. Having a right product is important and equally important is to present it before the right customer (one who actually needs it or is interested in it). The product should put on positive feeling of ownership among the individuals. And such feelings are clearly expressed in opinion mining polls. Sentiment analysis is a process of finding user's review towards a website or a product. Sentiment analysis is classified into positive comment, negative comment or neutral comment. Figure 1.1 shows the complete process of

sentiment analysis that refers how the input is being classified on the various steps.

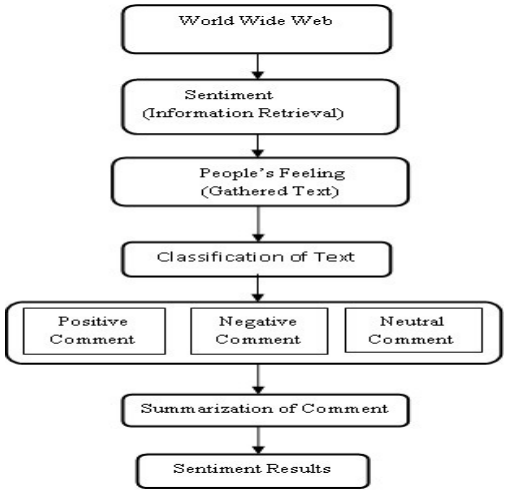

Figure 1.1: Steps of Sentiment analysis.

The sentiment analysis summarization process contains three main steps, first is Sentiment information retrieval, second is Sentiment classification and third is Sentiment summarization. Review text is retrieved from review websites such as Twitter, Facebook, Amazon and News Sites etc. Sentiment text in the blog, reviews, comments, microblogs etc. contains subjective information about the topic or issue. Sentiment results are generated based on features (sentiment sentences) selection about a matter.

In general, feature based opinion mining involves three subtasks viz.

(i) To correctly identify the opinionated and product specific features,

(ii) To identify the review sentences attributing positive/ negative opinions to the extracted features and

(iii) To generate a feature based summary from the information extracted.

The aim is to improve the accuracy and simplify the task of mining the opinions of customer reviews with respect to the features extracted.

A recent study focused on cost-effective values of online reviews and provides deep understanding between product reviews and their sales performance [5]. People tend to read online reviews understanding the opinions and sentiments and trust them as much as they are recommended by their friends or families. Twitter, a social networking service plays significant role in social networking research. Tweets give rich information about movie, product, or service [6].

Sentiments perform very important role for predicting future sales performance, a mix of good and even bad reviews will create a positive effect on the sales performance and sales prediction. In this thesis, the different types of issues are encountered like modelling the online reviews and tweets, sales prediction, and deriving the actionable knowledge. The sentiments, movie past sales performance these factors are important for predicting sales performance of movies [5]. A most current sentimental approach is that classifies the reviews into positive and negative which not gives complete understanding of sentiments. So that we done mining of sentiments based on Probabilistic Latent Semantic Analysis (PLSA) is S-PLSA model, which is totally different from traditional PLSA model [7].This model consists of the sentiments from reviews and tweets as the joint result of hidden factors and handle multifaceted nature of sentiments. S-PLSA model considered sentiments as a substitute of topics. So instead of using bag of words only sentimental words are considered. In this model only appraisal words from reviews and tweets are exploited for composing the feature and used for gathering the hidden sentiment factors.

1.3 PREDICTION BASED ON TWITTER DATA ANALYSIS

Prediction based on Twitter data has been studied in recent years, especially for predicting election results [8]-[9]. Burnap et al. [8] proposed a prediction model for using Twitter as an election forecasting tool and applied it to analyse the UK general election 2015. They assigned a string of text a positive or negative score ranging from 5 to +5 according to its election related sentiment. Each score is based on words in the string that are known to carry emotive meanings. For example, the score for love is 5 and for hate is 4. Firstly, they calculated a score for each tweet and produced a list of all tweets with positive and negative scores. After that, they consolidated the scores for each party and its leader, based on which they predicted the change of seats in the parliament. However, it is not very clear how the scores were consolidated.

Another election prediction model based on Twitter data was proposed by Cameron et al. [9], in which the authors tried to answer the following questions: What are the links between political information made available through social networks and the voting choices of citizens? Does an online presence and a social media strategy matter? Is online activity an indicator of support and does it influence election results? They counted the number of friends and followers of each electorate candidate in Facebook and Twitter. However, majority of candidate profiles in social networks were not complete. To test the relationship between the number of supporters for each candidate in social networks for a specified date and the election result, two regression models were proposed: a linear OLS model of vote share and a logistic regression model with election outcome as the dependent variable. However, their result showed that their models were not effective in predicting election results based on social media.

Gayo-Avello et al. [10] also showed that Twitter data seemed to be a poor electoral predictor because of demographic bias. However, the Twitter data used in their studies were collected before 2010 and there have been dramatic changes in social media since then, and it is worth revaluating the predictive power of social media in political events such as elections.

Le et al. [11] leveraged The American Voter as an analysis framework in the realm of computational political science, employing the factors of party, personality and policy to structure the analysis of public discourse on online social media during the 2016 US presidential primaries. They illustrated the importance of multifaceted political discourse analysis by applying regression to quantify the impact of party, personality, and policy on national polls. Their analysis found that tweets revealed continuing importance of party, personality, and policy [12] and their finding was also enriched by the application of sentiment analysis techniques [13].

1.4 Machine Learning Approach

Firstly, we will be classifying the tweets on the basis of Support Vector Machine (SVM) and adaboosted Decision Tree individually. The tweets i.e. string input will be taken which will be converted to numeric type by TF-IDF. Then a hybrid technique will be applied by feeding the outputs obtained earlier as the input to the Decision Tree.

A. Tf–idf

In context of information retrieval, tf–idf or TFIDF (term frequency–inverse document frequency) is used which provides a statistic of how a particular word is crucial for the given document dataset. Tf-idf uses weights for text mining and information retrieval and its value is directly dependent on the number of a particular word within the given dataset and we count the frequency of each word in the given document for a better information retrieval that regulates more frequently appearing words generally. Tf-idf weight is used in the search query when searching for a given document relevant to the given user query and is used by the search engines in the ranking of the document. Term Frequency (tf) describes how frequently a word appears in a document. Inverse document frequency (idf) describes how important a particular word is with respect to the given document. One of the best classification methods is worked out by adding the tf–idf or every search query word. Besides this, other subtle classification methods are variations of this easy model. [14]

B. Support Vector Machine (SVM)

Support Vector Machines employ the technique from computational learning theory which basically aims at reducing the structural risk. Finding the decision boundary that maximizes the distance between two classes is the basic principle of SVM. The vectors that define this decision boundary are termed as support vectors. SVM algorithm builds the classification model that assign test examples to one of the predefined class categories making it a non-probabilistic linear classifier. The basic principle of SVM involves three steps:

(1) Finding out the optimal decision boundary which maximizes the distance between two classes.

(2) Enhancing the same for non-linear separable problems.

(3) map data to high dimensional space so that data is classified easily for linear separable and non-linear separable cases.

C. Decision Tree

A decision tree is basically used to represent choices and their subsequent results in the form of graphs. Nodes of the graph represent an event and edges represent decision condition. The complex problems of branches are solved by segmenting it. A decision tree model is generated with training data and some sets of validations are used to check and improve the performance of decision tree model. It is the work of the decision tree to clarify all the branches which give us the answer to a complex solution. The most popular classification technique is the DECISION TREE in data mining. In a tree form they form some of the classification rules, and have several other advantages when compared to other techniques.

- Easily understood because of it's the simplicity of its presentation
- Decision tree can be applied to any data types like nominal, ordinal, and numerical, etc is punctuated within the parentheses.)
- Test data classify very fast with the help of decision tree algorithm.

D. ADABOOSTED DECISION TREE

Adaboost is a machine learning algorithm; it is a solution to the problem created by "boosting" weak classifiers into strong classifiers, which can be done by weighting their respective outputs. The input to the boosting algorithm is the training data in the form of t (x1, y1), (xm, ym) where xi represents instant space X and yi represents class label set Y. Let say Y ={0, 1}. AdaBoost algorithm recursively iterates the basic learning technique with t iterations denoted as 1 to t. Distributing weight set over the entire training dataset is the main objective of boosting algorithm. In the initial phase, if Dt (i) is the distributed weight of training dataset I on t iterations, all weights are equally distributed and, in every iteration, it gradually increases the weight of incorrect classified tuples and the weak learners are focused more in the training examples.

1.5 REVIEW OF LITERATURE

In the recent time, research on Sentiment Analysis in information retrieval sphere concentrate on classifying

sentiment according to their polarity either as positive or negative or neutral. Earlier researches on Sentiment Analysis are based on subjectivity or sentiment level. One standard technique was developed by [15] which outperformed human produced baselines as it analyzed the movie reviews on the basis of different Machine Learning classifiers. For many other researchers, this technique is functioned as a baseline like [16] as he intended to develop a "Distant Supervision Learning" method to handle noisy labeled tweet data. As for sentiment level detection, [17] described a system that detects the sentiment of message-level task and the sentiment of term level task by creating two different SVM classifiers. But [18] used aspect based sentiment scrutiny technique which was applied on datasets for different languages and domains. On the other hand [19] evolved a structure to handle the target tweets. Few researchers reconnoitre features trade along with combination [16] where they developed a feature combination scheme which utilizes the sentiment lexicons and the extracted feature combinations [20] where they developed a feature combination scheme which utilizes the sentiment lexicons and the extracted tweet unigrams of high information gain by evaluating the performance. Current trend of Sentiment Analysis shows that consequences of micro-blogging sentiment classification are broadly used in different social media applications, including tracking sentiment towards products, movies etc.

Many different approaches are used for the processing of text in sentiment analysis. Constructing lexical chains, machine learning and many more are very useful approaches for the purpose. Others could be statistical approaches, domain knowledge driven analysis. Such approaches proved to be very advantageous in the task of sentiment analysis. Work has been accomplished by researchers in many different languages such as Thai, Nepali, Bengali, Malayalam etc. But very small amount of work has been done in Hindi Language. Results provided by processing of sentiment analysis are much time saving and accurate too. The very first work was done in Hindi, Marathi and Bengali. But the level of work in Hindi language at this time is not much appreciable. So the need of the same has been realized as the result of the various surveys.

Das and team used English–Bengali dictionary for creating a sentiwordnet for Bengali which consisted of 35,805 words created by them [21]. Work was further preceded by giving four strategies which predicted the sentiment of word [22]. Yakshi Sharma, Veenu Mangat, Mandeep Kaur, used unsupervised lexicon method for classification so as to compare the proposed algorithm with unigram presence method. The positive and negative words are then counted to choose the dominating one [23]. Huangfu et.al furnished an improved news sentiment analysis method. In this, a news sentiment analysis is done by making an in-depth study to Chinese news by chopping the title and text distinctly and in both the areas two different algorithms are applied. Neutral news determination method has been proposed for title part and a subjective sentence recognition algorithm is used for the news text. Finally, a different weight for title and text sentiment is used to compute final sentiment value for news [24]. Bhoir and Kolte proposed a system in which the movie reviews are summarized at a level which helps a user to easily find out which aspects of movie are liked and disliked by the user. Subjectivity Analysis is performed for the proposed system which is one of the vital and useful tasks in sentiment analysis. They implemented two different methods for finding subjectivity of sentences after that to find feature –opinion pair, rule based system is used and finally orientation of extracted opinion is revealed. Phenomenon like Naïve Bayes and Sentinel are used here along with precision, recall, Fmeasure. System performance is increased in terms of efficiency and accuracy and here they showed that Naïve Bayes classifier is giving more accurate results than SentiWordnet [25].

Joshi et.al proposed a strategy which claimed to be a fall back strategy for Hindi language. Three approaches are followed by their strategies which were: In-language Sentiment Analysis, Machine Translation, Resource based Sentiment analysis. They developed a Hindi SentiWordnet (HSWN) by replacing words of English WordNet by their Hindi equivalents. Finally an accuracy of 78.14% has been received by them Xiadong Yan, Tao Huang proposed and implemented a system which is a Tibetan sentence sentiment tendency judgment system that is based on maximum entropy and at the same time, parallel a corpus of 10000 Tibetan sentiment sentences is also tested resulting efficiently. Chandan Prasad Gupta, Bal Krishan proposed a system in which it is assured that this is the first work carried out for detecting sentiment in Nepali texts and the results reflect that the Machine learning approach performs better than the rule based approach and have great impacts on the accuracy and efficiency of the system. They had developed Nepali Sentiment corpus and Nepali Sentiword net. Firstly an approach is seen where a lexical resource called Bhavanakos is developed. Secondly, an approach to train a machine learning based text classifier is concerned by them.

Zhongkaihu, JianqingHu, WeifengDing, XiaolinZheng imported a deep neutral network which is effectively appropriate for high dimension data analysis, and a framework of sentiment analysis is developed based on deep learning. Different methods of constructing feature vector for document level sentiment analysis are analyzed ensuring the performance by remarkably reducing the training time; sub-network for each feature vector is separately trained [38]. Yan Wan, Hongzhuruinie TianguangLan, Zhahui Wang developed an algorithm in order to highlight the implicit features that are emotional words here. An example is taken so as to illustrate significant value in fine grained sentiment

analysis of online reviews, where they help the producers to make improvements clear and on the second hand helping the consumers to understand advantages and disadvantages of the target product, thus making a wise decision. Andréa ensures that the carried work explores the usage of several classifiers and feature extraction methods for classifying the large set of business text reviews.

II. PROPOSED METHODOLOGY

This stage will give explanation of the process. Before that try to look at the general process in this system in Figure 3.1.

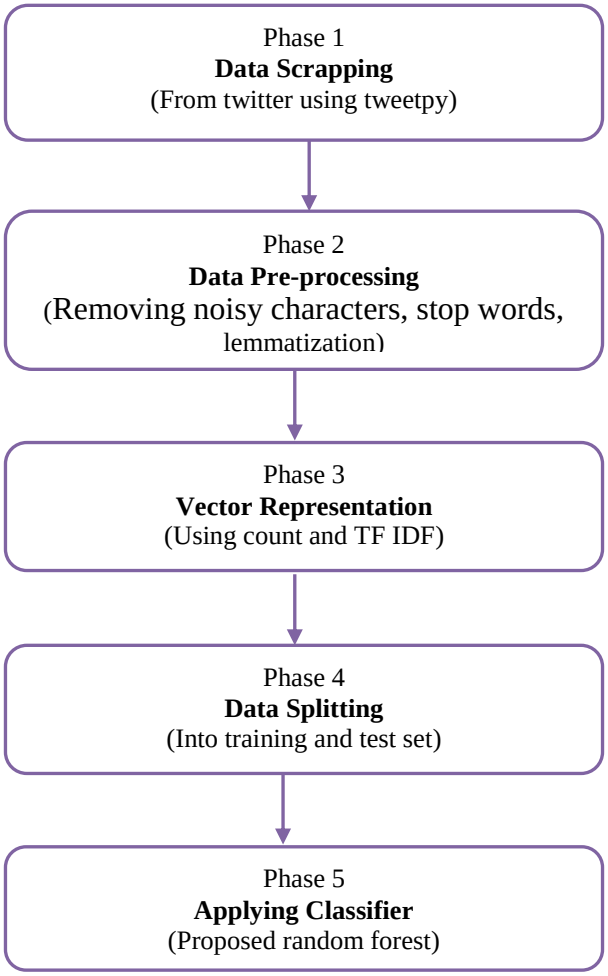

Figure 2.1: Proposed Model.

Details the process description:

1. Retrieved of raw data in the form of tweet using API from twitter then data stored in .csv form. For labeled data is training data and unlabeled data is data testing. We took tweets with the names of Indian political parties such as #BJP, #Congress, #others.

2. In the data testing or training data is done preprocessing. The process of preprocessing involves deleting URLs, punctuation, deleting stop words, changing the word slang to raw and stemming.

Twitter provides an API to search its data based on keywords or tags so as to extract tweets for specific purposes. Using time tags and Indian e tweets related to the 2019 Loksabha election and posted for the politicians were collected and sorted in terms of dates and for our experiments. The collected tweets were then classified as positive, negative, or neutral to specific candidates using keywords reflecting positive and negative sentiments as features, which were extracted by data preprocessing or domain knowledge. For example, based on domain knowledge, vote, win or wins, and lead can be positive words, whilst not, bad, attack, betray can be negative words.

Term frequency was adopted to get more keywords in our study. According to the frequency of words in the collected tweets, most frequently occurred words with sentimental meanings were considered as features for classifying tweets. However, word frequency usually changes with time. Some words may have high frequency on one day, but may not appear in any tweet on another day, such as "Gathbandhan" leads in the Loksabha election 2019."

It first identifies the presence of URL using a regular expression and removes all the URLs from the extracted tweet. Then it removes all the private usernames by identifying "@username". Then it removes all the hashtags identified by the symbol "#" and all the special characters. Refined tweets are then classified using classification scheme. Negation handling is one of the factors that contributed significantly to the accuracy of our classifier. The major problem occurs during the sentiment classification is in the negation handling. When we use each word as a feature, the word "win" in the phrase "not win" will be contributing to positive sentiment rather than negative sentiment .This will lead to the errors in classification. This type of error is due to the presence of "not" and this is not taken into account. To solve this problem we applied a simple algorithm for handling negations using state variables and bootstrapping. We built on the idea of using an alternate representation of negated forms [7]. This algorithm stores the negation state using a state variable. It transforms a word followed by an't or not into "not"+ word form. Whenever the negation state variable is set, the words read are treated as "not"+ word. When a punctuation mark is encountered or when there is double negation, the state variable will reset. Many words with strong sentiment occur only in their normal forms in their training set. But their negated forms would be of strong polarity. We solved this problem by adding negated forms to the opposite class along with normal forms during the training phase. It means that if we encounter the word

"fail" in a negative document during the training phase, we increment the count of "fail" in the negative class and also increment the count of "not fail" for the positive class. This is to ensure that the number of "not" forms is sufficient for classification. This modification resulted in a significant improvement in classification accuracy due to bootstrapping of negated forms during training.

3. Then performed feature extraction process on tweet that has been clean result of preprocessing. The feature extraction process includes word grouping with the Bag Of Words method and feature weighting with Tf-idf.

Following features are extracted for BJP, Congress and Others political parties.

BJP=['modi', 'modiji', 'win', 'vote', 'namo', 'support', 'good', 'best', 'bjp', 'modi ji' 'bhakt', n 'hindu', 'love', 'like', 'amit', 'pappu']

CONGRESS=['congress', 'rahul', 'win', 'feku', 'good', 'best', 'love', 'like', 'bjp lose' 'support', 'fekugiri']

OTHERS=['arvind', 'aap', 'feku', 'bjp lose', 'pappu', 'congress lose', 'fekugiri']

Then the negative words are removed from tweets. The algorithm is mentioned below:

```
Negative_words=['not']

With open ('Feautures.csv', 'w', newline='') as file:
   writer = csv.writer(file, delimiter = ',')
   x = ['Bjp','Congress','Others']
   writer.writerow(x)
   for i in range(len(corpus)):
      b=0
      c=0
      o=0
      temp = corpus[i]
      for word in temp.split ():
         If word in BJP:
            b+=1
         if word in CONGRESS:
            C+=1
         if word in OTHERS:
            o+=1
         if word in Negative_words:
            O-=1
            B-=1
            C-=1
      writer.writerow ([b,c,o])
```

4. Tweet that is already a collection of features in the Bag of Words and has been given a weighting using Tf-idf classified using the Multinomial Logistic Regression method.

5. The classification process produces accuracy and sentiments of tweet.

3.2 Proposed Algorithm for Sentiment Calculation:

Algorithm: Calculating sentiment scores of a candidate.
Input: Election candidate Ec, Set of candidate-related tweets, Positive Lpos and negative lexicon Lneg.
Output: Positive sentiment score of candidate c Sent.posEc, Negative sentiment score of candidate Sent.negEc.

1: initial Sent.posEc =0; Sent.negEc =0;
2: initial posEc=0; negEc=0;
3: for every tweets do
4: words ← Tweets (.CSV);
5: for every word w that belongs to words (tweets) do
6: if w € Lpos then
7: pos ← pos + 1;
8: if w € Lneg then
9: neg ← neg + 1;
10: Sent.posc ← Sent.posEc + pos
11: Sent.neg ← Sent.negEc+ neg;
12: return Sent.posEc;
13: return Sent.negEc;
14: End Algorithm.

3.3. Methods for Popularity Prediction

To calculate the popularity of a candidate in the election, this thesis proposes the following formula:

$$\text{Popularity (A)} = \left\{ \frac{\text{Pos (A)}}{\text{Pos (A) + neg(A)}} \right\} \left\{ \frac{\text{N (A)}}{\text{N(A) + N (B)}} \right\}$$

Where N(A) and N(B) are the number of neutral tweets are related candidate A and B respectively. Pos(A) and neg(A) are the number of positive and negative tweets that are related to candidate A.

III. RESULTS AND DISCUSSION

The classification performance can be evaluated in terms called: accuracy, which is shown below in table. Accuracy explains correctly classified instances. Table below shows the accuracy of all algorithms by applying count vectorization method. Table shows that proposed algorithm has highest accuracy of all.

Table 3.1: Showing accuracy.

Classifier	Accuracy (in %)
Naive Bayes	40.675
Decision Tree	54.018
KNN	56.425
LR	63.497
SVM	51.612

ADC	51.539
Proposed	**64.88**

Table below shows the accuracy of all algorithms by applying tfidf vectorization method. Table shows that proposed algorithm has highest accuracy of all.

Table 3.2: Showing accuracy.

Classifier	Accuracy (in %)
Naive Bayes	42.062
Decision Tree	58.718
KNN	58.093
LR	69.031
SVM	46.375
ADC	61.843
Proposed	**73.562**

IV. CONCLUSION AND FUTURE SCOPE

Typically, opinion mining looks at social media content to analyse people's explicit opinions about an organization, product or service. However, this backwards looking approach often aims primarily at dealing with problems, e.g., unflattering comments, while a forwards-looking approach aims at looking ahead to understanding potential new needs from consumers. This is achieved by trying to understand people's needs and interests in a more general way, e.g. drawing conclusions from their opinions about other products, services and interests. It is not sufficient, therefore, to look at specific comments in isolation: non-specific sentiment is also an important part of the overall picture.

In this thesis, Naive Bayes Algorithm, DECISION TREE algorithm, Logistic Regression algorithm, Adaboosted Algorithm, SVM algorithm and proposed algorithm for sentiment classification model are presented for improving the overall accuracy of the classifier in the classification of tweets. For the same we apply preprocessing techniques so that accurate data is fed as an input to the training process, our proposed approach classify the tweets as Positive and Negative tweets which further helps in sentiment analysis and uses that sentiment analysis for further decision making. The work of proposed model has gone through preprocessing stage and classifiers learning stage. For analytical evaluation of the proposed classifier accuracy, precision and recalls are used. The comparative results prove that proposed model improved the overall classification accuracy and precision measure of sentiment prediction as compared to traditional existing techniques for classification.

REFERENCES

[1] Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp.1320-1326.

[2] "Number of monthly active Twitter users worldwide from 1st quarter 2010 to 3rd quarter 2017 ", Availabe at "https://www.statista.com/statistics/282087/number-ofmonthly-active-twitter-users/. " [Online; accessed 4-December-2017].

[3] K. Bannister, "Sentiment Analysis". Available: " https://www.brandwatch.com/blog/understanding-sentimentanalysis/ ".

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in Proceedings of the 6th Conference on Symposium on Operating Systems Design & Implementation, Volume 6, pp. 10–10, 2004.

[5] Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, Aijun An, "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA, vol. 24, No.4, APRIL 2012.

[6] Andrei Oghina, Mathias Breuss, Manos Tsagkias&Maarten de Rijke. (2012) Predicting IMDB movie ratings using social media. Proceedings of the 34th European conference on Advances in Information Retrieval, pp. 503-507.

[7] Liu, Bing, and Lei Zhang. "A survey of opinion mining and sentiment analysis." In Mining text data, pp. 415-463. Springer US, 2012.

[8] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, 140 characters to victory?: Using Twitter to predict the UK 2015 General Election Journal of Electoral Studies, vol. 41, pp. 230-233, 2016.

[9] M.P. Cameron., P. Barrett, and B. Stewardson, Can social media predict election results? Evidence from New Zealand Journal of Political Marketing, vol. 15, pp. 416-432, 2016.

[10] D. Gayo- Limits of electoral predictions using Twitter Proceedings of the 5th ICWSM, 2011, pp. 178 18.

[11] H.T. Le, G.R. Boynton, Y. Mejova, Z. Shafiq, and P. Srinivasan, Revisiting The American Voter on Twitter Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 4507-4519.

[12] A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp.1320-1326.

[13] B.O. Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith," From tweets to Polls: Linking Text Sentiment to public opinion time series" Proceedings of the 4th ICWSM, 2010, pp 122 129.

[14] Rui Xia, FengXu, ChengqingZong, QianmuLi, Yong Qi, and Tao Li, August 2015," Dual Sentiment Analysis: Considering Two Sides of One Review", IEEE transactions on knowledge and data engineering, Vol.27, AnNo.8.

[15] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[16] Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1.2009 (2009).

[17] Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. "NRC Canada: Building the state-of-the-art in sentiment analysis of tweets."arXiv preprint arXiv:1308.6242 (2013).

[18]Pontiki, Maria, et al. "SemEval-2016 task 5: Aspect based sentiment analysis." ProWorkshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, 2016.
[19]Rosenthal, Sara, Noura Farra, and Preslav Nakov. "SemEval-2017 task: Sentiment analysis in Twitter." Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017.
[20]Yang, Ang, et al. "Enhanced Twitter Sentiment Analysis by Using Feature Selection and Combination." Security and Privacy in Social Networks and Big Data (SocialSec), 2015 International Symposium on. IEEE, 2015.
[21] Amitava Das, SivajiBandopadaya, SentiWordnet for Bangla, Knowledge Sharing Event -4: Task, Volume 2, 2010.
[22] Amitava Das, SivajiBandopadaya, "SentiWordnet for Indian Languages", Proceedings of the 8th Workshop on Asian Language Resources, Pages 5663, Beijing, China, August 2010.
[23]Yakshi Sharma, VeenuMangat, MandeepKaur, A practical Approach to Semantic Analysis of Hindi tweets", 1st International Conference on Next Generation Computing Technologies(NGCT-2015), Dehradun, India,Page No(677-680), September 4-5, 2015.
[24]Yu Huangfu, Guoshiwu, Yu Su Jing Li, Pengfei Sun Jie Hu, "Än Improved Sentiment Analysis Algorithm for Chinese news", 12th International Conference on Fuzzy Systems and Knowedge Discovery(FSKD), Page No(1366-1371), 2015.
[25]Purtata Bhoir, Shilpa Kolte, "Sentiment Analysis of Movie Reviews using Lexicon approach", IEEE International Conference on Computational Intelligence and Computing Research, 2015.

