An Efficient Approach for the Prediction of Election Result using Machine Learning

Madhuri Nag[1], Neha Khare[2], Deepak Agrawal[3]
[1]Research Scholar, Takshshila Institute of Engineering and Technology, MadhyaPradesh India
[2]Professor, Takshshila Institute of Engineering and Technology, MadhyaPradesh India
[3]HoD, Takshshila Institute of Engineering and Technology, MadhyaPradesh India

Abstract- *In Sentiment analysis, study of opinions, sentiments, attitudes, views and emotions expressed in text. Classification is done to find out the polarity of the word and then categorize them into positive or negative sentiment. Sentiment analysis is done on Facebook and Twitter which offers an important way to calculate the public's feelings towards their party and politicians. Classification problem is the biggest challenge for the classification of different tweets.*

Keywords- Sentiment Analysis, Machine Learning, Data Mining, Random Forest, Naive Bayes, Support Vector Machine.

I. INTRODUCTION

Social media is revolutionizing the ways of social life of individuals by creating a virtual community for self-expression, connecting and collaborating with others. The sharing of information on social media results into a huge amount of user-generated content which discloses personal information and interests that can be used in making predictions inconspicuously. The data collection and analysis from social media data is time and cost-effective when compared with the traditional approaches. The researcher reported that the predictive power for measuring individual attributes from social media data does not affect by the representative nature of the sample to be a true representative of the population [1]. Twitter is one of the most frequently used social networking platforms being for data analysis and has approximately 100 million active users on daily basis. The online website's data can also be utilized for predicting and reporting job opportunities [2, 3]. The Twitter data has been analyzed by researchers in healthcare [4], classification and comparative analysis [5] and predicting elections [6]. In addition to making predictions, many researchers also investigated the uncertainty of results [7]. Presently, political parties increasingly rely on social media platforms i.e. Twitter and Facebook for political communication, interacting with voters and promotions. This increased use of social media by political candidates for attracting potential voters has been reflected by the 2011 general elections in New Zealand. In a modern democracy, the problem of predicting electoral results is very popular and has attracted the attention of researchers from computing domain to predict election results based on data collected from social media platforms [8, 9]. Prior studies i.e. Skoric et al., 2015 [10], on predicting electoral results reported accurate outcomes using data collected from social media networks and techniques such as sentiment analysis, volumetric analysis and social media analysis for countries including United States, United Kingdom , Ireland and India [11]. The electoral studies are very different from other studies being conducted in political science domain, as their goal is not to explain the election results but to forecast outcomes.

II. SOCIAL OPINION MODEL

Sentiment-related phenomena can be explained as the process of evaluation of events, objects or persons [12], [13]. The opinions are caused by the subjective evaluation of the "raw" stimuli. The "raw" stimuli may have no intrinsic emotional meaning, but will be appraised by personal relevance and implications [14]. For social opinion, the "raw" stimuli are only text and its features which are difficult to expound the corresponding sentiment related phenomena. In fact, there are less than 5% of directly emotional words of a text in daily speech, emotional writing, and affect-laden poetry [15]. In journalism domain, a lower percentage is undisputed. It is rarely influenced by the personal relevance under the social community. To simplify the problem, we focus on implications without personal relevance. According to the cognitive approaches, the result of voting is "the person's experience, goals and opportunities for action" [16]. It is process that evaluates an event by dimensions such as urgency, consistency with goals, etc. All the social opinions share the similar emotional experience, goals and opportunities for action with each other. From the NLP perspective, the models are inexplicable but feasible. From psychology and linguistics perspective, the models are explicable but lack of use in the service. Based on the general characteristic, similarity is one of six principles that guide human perception of the world in Gestalt theory, [17]. We can predict social opinions by measuring the semantic similarity between events.

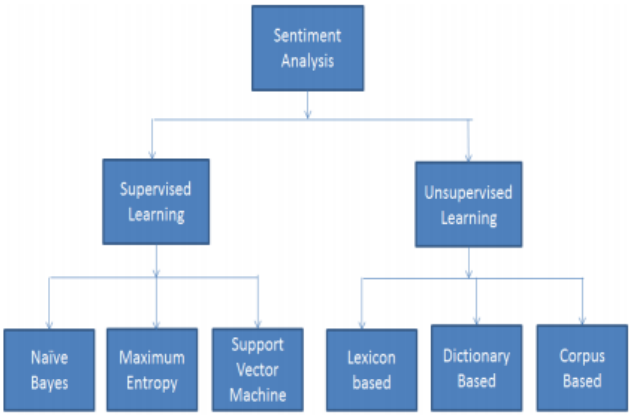
The social cognitive process can be modeled based on a stereotypical knowledge set consisting of social opinion: $P=\{\langle event, f, s, t\rangle_1, \langle event, f, s, t\rangle_2, \cdots, \langle event, f, s, t\rangle_n\}$

Instead of establishing appraisal criteria, the cognitive process can be regarded as the neighbor analysis in set P. The set P can be interpreted as social experience. The cognitive process can be simplified as matching the "raw" stimuli between the priori social opinions in set P. The social community has stabilized emotion towards specific events. It can be explained by a social psychology behavior named "stereotype" which is a fixed view of people, groups, events, institutions, or problems [18], [19]. Stereotype widely exists in media [18], [19]. To be more accurate, the task is modeling the relationship between current event and priori social opinions based on semantic similarity.
As we defined, the social opinion is a quadruple {event, f, s, t}. In social opinion model, the first step is called words-extract in which the raw feature (f) of social event is processed as Bag of Words (BOW) which is widely used for document representation. Considering the online performance of the algorithm, we utilize the term frequency (TF) instead of term frequency–inverse document frequency (TF-IDF) for measuring the importance of a word in corpus. TF-IDF is not available in real-time updated data. Because we need to know the global distribution of words to calculate TF-IDF. By contrast, TF can be explained as social community ratio of attention to the words. Based on the explanation of attention, it is persuasive that the readers concern each news equally, in other words, pay equal attention to each news. Based on it, the assumption is justifiable that the BOW f should be normalized. So far, the preprocessed i f in social opinion quadruple, {event f s t} is a normalized histogram which is finite-dimensional vectors with nonnegative coordinates whose sum is equal to 1.

III. SENTIMENT ANALYSIS

The area of study that interprets people's opinions, against any particular topic, about any event etc. in text mining it is known as opinion mining or sentiment analysis. It produces a vast problem zone. There are also various names and having different tasks, e.g., sentiment analysis, opinion extraction, opinion mining, sentiment mining, affect analysis, subjectivity analysis, review mining, etc. [26] Levels of Analysis: In general, sentiment analysis is categorized into mainly three different levels: A. Document Level Analysis: This level classifies that whether the complete document gives a positive sentiment or negative sentiment. The document is on single topic is considered. Thus texts which comprise comparative learning cannot be considered under document level. B. Sentence Level Analysis: The task of this level is sentence by sentence and decides if each sentence represents opinion into negative, positive, or neutral. Neutral, if sentence does not give any opinion means it is neutral. Sentence level analysis is related to subjectivity classification. That expresses factual information from sentences that gives subjective aspect and opinions. i.e. good-bad terms. C. Entity/Aspect Level Analysis: Both the document and the sentence level analysis don't find peoples like and dislikes. Entity/Aspect level gives throughout analysis. Entity/Aspect level was earlier called feature level. The core task of entity level is to identification constructs, aspect level straightforwardly gives attention at the opinion or sentiment. It is based on the concept that an opinion resides of an attitude and a destination of opinion.

The fundamental task in Sentiment Analysis is classifying the polarity of a given tweets feature. The polarity is in three classes i.e. Positive, Negative and Neutral. Polarity identification is done by using different lexicons e.g. Bing Lui sentiment lexicon, SentiWordNet etc. which help to calculate sentiment strength, sentiment score, etc.



Two fundamental approaches are there in sentiment analysis i.e. Supervised learning Approach and unsupervised learning Approach. Sentiment classification of twitter data is done using supervised machine learning approaches like Naïve-Bayes, SVM, and Maximum-Entropy etc. Efficiency of classifier is built upon which dataset is used for which classification methods. In the case of Supervised machine learning approaches to train the classification model Training dataset is used which then help for classification of test data. [27].

IV. LITERATURE REVIEW

Parama Fadli Kurnia, Suharjito[2018] In this paper , a business intelligence dashboard is created to observe the performance of each topic or channel of news posted to social media account such as Facebook and Twitter.

Topical performance in social media is the number of Topics in articles posted to social media getting like, share, comment etc To be able to know the Topic of a news post in social media, used some text classification techniques such as Naive Bayes, SVM and Decision Tree. The comparative results

of the algorithms are taken which has the best accuracy of SVM for subsequent implementation in the data warehouse.

Sourav Das, Anup Kumar Kolya[2017], In this paper, we present a simple and robust work to gather, analyze and graphically represent people's opinion about India's new taxation system using Naive Bayes algorithm.

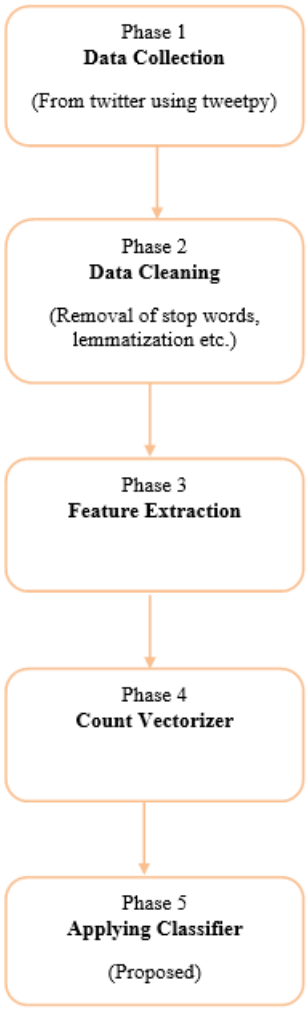
Limitations- Naive Bayes Classifier primarily works on the conditional probability theory. It offers the assumption of a particular feature from a class of features. But not necessarily, it will come out as the accurate one. It only work if the probability that is already recorded for a particular class

Huma Parveen, Prof. Shikha Pandey(2016), The Naive Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it permits us to capture uncertainty about the model in a principled way by determining probabilities. It helps to solve diagnostic and predictive problems.

Limitations- Firstly the data is downloaded from the twitter. They are stored into the HDFS for analysis. Before evaluation the tweets, we need to be pre-processed in order to remove the noise from the data. Pre-processing is complex.

Rohit Joshi, Rajkumar Tekchandani, (2017). In this paper we applied supervised machine-learning algorithms like support vector machines (SVM), maximum entropy and Naïve Bayes to classify data using unigram, bigram and hybrid i.e. unigram + bigram features. Result shows that SVM surpassed other classifiers with remarkable accuracy of 84% for movie reviews. In this paper, we have done comparative analysis on supervised classifiers like Naïve Bayes, support vector machines and maximum entropy using unigram, bigram and hybrid (unigram + bigram) feature. Naive Bayes is used for classification

V. PROPOSED MODEL



VI. PROPOSED WORK

Step I. Data Collection:

The data/tweets are collected from twitter platform by using tweepy library. Tweepy is a library used to import tweets from tweeter. The data will be imported from tweeter only after authentication using consumer key and consumer secret key. The imported data is stored in the form of comma separated file. The tweets are collected similar to Indian parties for election prediction such as #BJP, #Congress, #aap and #others.

Step II. Pre-Processing:

The Data Processing segment is required for transformation, standardization, and cleaning the information.

In this phase the unwanted reviews are removed from the collected data that are not used for analysis parameters. This step involves various steps like deleting URLs, deleting the stop words, porter stemming, and so on. Also 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.

Step III: Feature Extraction

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 . 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.

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]).
```

Step IV: Count Vectorizer

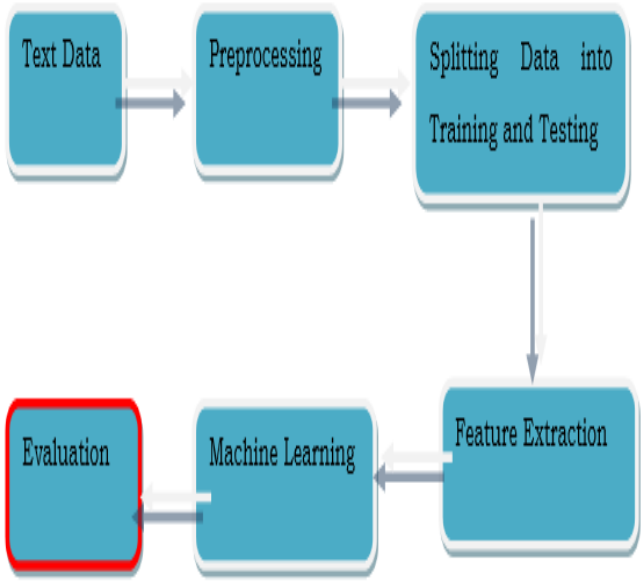
Text files are actually series of words. In order to run machine learning algorithms the text files are converted into numerical feature vectors using BoW/CV/TF-IDF representation models. Count Vectorizer creates Document Term Matrix with token counts to represent the feature collection. TF-IDF Vectorizer is an alternative for Count Vectorizer that constructs as normalized TF-IDF score for each word instead of frequency count. Briefly, each text file is segmented into of TF-IDF features matrix. Each unique word in the dictionary will correspond to a feature (descriptive feature). Scikit-learn is a Python open source library is used to realize the text classification algorithm. It has a high level component named as TF-IDF VECTORIZER which will create feature vectors.

Step V: CLASSIFYING TWEETS

Tweets Classification (TC) is a significant machine learning problem. If there is a trained set of documents in predefined classes, the task is to automatically classify a new document into one of these known classes.

Random Forest: As the name suggests, Random Forest algorithm generates the forest with a number of decision trees. So it is the collection of decision trees. Decision trees are attractive classifiers among others because of their high execution speed. Based on random samples from the database a random forest classifier averages multiple decision trees. Generally, the more trees in the forest is the sign that forest is robust. Similarly in the random forest classifier, high accuracy is obtained by higher the number of trees in the forest. While concurrently creating a tree with decision nodes, a decision tree breaks the dataset down into smaller subsets. The decision root node is selected through highest information gain and leaf nodes based on a pure subset for each iteration simultaneously. Calculation of Information Gain (IG) requires impurity measure (Entropy) of that node. There are various indices to measure the degree of impurity. A leaf node represents a category or pure subset. The trees in a random forest are created under random data so there might be chances to be lack meaning and noisy. In

order to make a model with low variance random forest averages these trees. The irrelevant trees drop each other out and the staying meaningful trees yield the final result.
The stages of Tweets Classification process are shown in Figure 3.2 below



- **Proposed Algorithm**

Input = test data.
Procedure: C.V = CountVectorizer (Input) a vectorized matrix created from the input test data.
S.M = (np.array(C.V)) a sparse matrix is generated from the vectorized data.

This sparse matrix will be passed to different trained machine learning algorithms to get the result.
NaiveBayes () = trained (multinomial naive bayes algorithm).

NaivePred = Naive (S.M) result from the trained naive bayes algorithm.
SVM () = trained (Support Vector Machine algorithm)

SM= result from trained svm algorithm.
Decision = trained (Decision Tree algorithm)

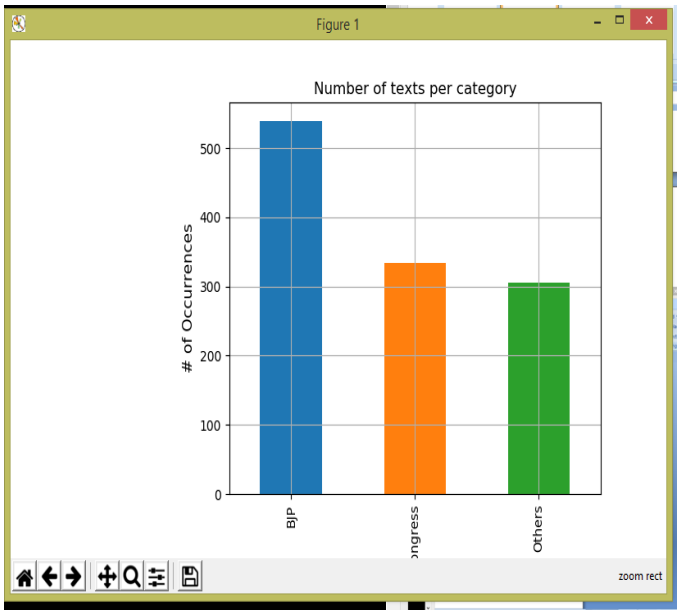
DecisionPred = result from the trained Decision Tree algorithm.
RandomForest() = trained(Random Forest algorithm)

RandomForestPred = result from the trained Random Forest algorithm.

VII. RESULTS AND EVALUATION

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.

Classifier	Accuracy (in %)
Naive Bayes	40.675
Decision Tree	54.018
KNN	56.425
LR	63.497
SVM	51.612
ADC	51.539
Proposed	**65.28**



VIII. CONCLUSION

The topic of predicting elections is gaining the attention of researchers as the utilization of social media data is making an important place due to its real-time nature and easy availability. Many attempts have been made by researchers to explore and validate the reliability of social media data in predicting election results. Mostly, twitter data was used by researchers in making predictions of electoral outcomes by checking the polarity of words using sentiment analysis. It is observed that very few studies utilized data collected from Facebook in making electoral predictions. Majority of studies found the social media data effective in making electoral predictions. Whereas few of the studies also discussed and reported that social media data cannot be relied upon. Similarly, sentiment analysis is the most common approach being adopted by researchers. But most of the studies also argued about the limitations attached to the approach of sentiment analysis and

suggested to consider the linguistic and contextual feature as well. In the future work, the data sources other the social media will be discussed along with the comparison of data sources and approaches used.

REFERENCES

[1] Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, 16(2), 340-358.

[2] Lovaglio, P. G., Cesarini, M., Mercorio, F., & Mezzanzanica, M. (2018). Skills in demand for ICT and statistical occupations: Evidence from web- based job vacancies. Statistical Analysis and Data Mining: The ASA Data Science Journal, 11(2), 78-91.

[3] Bilal, M., Malik, N., Khalid, M., & Lali, M. I. U. (2017). Exploring Industrial Demand Trend's in Pakistan Software Industry Using Online Job Portal Data. University of Sindh Journal of Information and Communication Technology, 1(1), 17-24.

[4] Nawaz, M. S., Bilal, M., Lali, M. I., Ul Mustafa, R., Aslam, W., & Jajja, S. (2017). Effectiveness of social media data in healthcare communication. Journal of Medical Imaging and Health Informatics, 7(6), 1365-1371.

[5] Tiwana, M. S., Javeed, F., Lali, I. U., Dar, H., & Bilal, M. (2018). Comparative Analysis Of Context Based Classification Of Twitter. Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA 2018), IEEE.

[6] Boutet, A., Kim, H., & Yoneki, E. (2012). What's in Your Tweets? I Know Who You Supported in the UK 2010 General Election. ICWSM, 12, 411-414.

[7] Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. Social Science Computer Review, 31(6), 649-679.

[8] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm, 10(1), 178-185.

[9] Fisher, S. D., Ford, R., Jennings, W., Pickup, M., & Wlezien, C. (2011). From polls to votes to seats: Forecasting the 2010 British general election. Electoral Studies, 30(2), 250-257.

[10] SKORIC, M., Liu, J., & Lampe, C. (2015, May). Gauging public opinion in the age of social media. In International Communication Association Annual Conference.

[11] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm, 10(1), 178-185.

[12] R. E. D. L. Lopes and O. Vian, "The Language of Evaluation: appraisal in English," in symposium/workshop on electronic design, test and applications, vol. 23, no. 2, pp. 371–381, 2007.

[13] Parama Fadli Kurnia, Suharjito(2018) Business Intelligence Model to Analyze Social Media Information" Elsevier.