Survey Of Election Prediction Using Machine Learning

Madhuri Nag[1], Neha Khare[2]
[1, 2] Dept of CSE
[1, 2] Takshshila Institute of Engineering and Technology, Jabalpur

Abstract- *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 different Indian languages. We used Twitter Archiver tool to get tweets in Hindi language. We performed data (text) mining on election related tweets collected over a period that referenced five national political parties in India, during the campaigning period for general state elections in 2020. We made use of both supervised and unsupervised approaches. We utilized Dictionary Based, Naive Bayes, SVM algorithm and Random Forest Classifier 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 results of the analysis were for the BJP (Bhartiya Janta Party), As it turned out, BJP will win in the 2020 general election, far more than any other political parties.*

Keywords- Sentiment Analysis; Twitter; Indian Elections; Naive Bayes; Support Vector Machine.

I. INTRODUCTION

Natural Language Processing (NLP) can be classified into opinion mining and text mining. It is used in segregating the views of people's postings with respect to different social media applications like Facebook, Twitter, etc. Text or Sentiment mining is also helpful in different situations such as analyzing people's feelings about a movie, product, song, etc. and to differentiate between positive, neutral and negative reviews. It can be used in places like the stock market, ecommerce websites, song recommendations, etc. for better predictions and recommendations."

There has been much research already conducted on Sentiment Analysis in the English language. Almatrafi et al [1] collected tweets using the Twitter API that considered only two major parties BJP (Bhartiya Janta Party) and AAP (Aam Aadmi Party) and labelled them as negative, neutral and positive. The aim of the paper was to analyze trends in the Indian General Election 2014 using location as a filter. They employed a supervised approach by applying a Naïve Bayes classifier. "The problem statement: Is it probable to predict the popularity of any political party and therefore extrapolate their chances of winning the election by utilizing sentiment analysis of Twitter data? To answer this question, it is imperative to analyze Twitter tweets to learn and study the sentiments of people in terms of positive polarity, neutral polarity and negative polarity. To analyze the problem statement, the authors obtained tweets, filtering for Hindi language and then applied sentiment mining and prediction operations. This takes us to certain research queries, for example, how to anticipate and break down what strategy is being accomplished? What steps are suitable for the task of election prediction? Using tweets, we can analyze the positive or negative feelings or opinions posted by people on social media. Furthermore, the preprocessing techniques, such as removal of emoticons, repeated words, Twitter mentions, Hindi stop words etc. are applied to dataset (tweets) and then classification models are applied for predicting the results.

II. LITERATURE REVIEW

This part of the paper is used to explain the related study of opinion mining in different Indian languages, related techniques, micro-blogging system tasks and algorithms to fulfill those tasks. Furthermore, it talks about certain significant categories that emerged from this study. It involves the analysis of Indian languages such as Hindi, Marathi, etc. to predict the results of the upcoming general elections.

A. Sentiment Analysis in Local Language

Data mining is a wide area, but there have not been many experiments done in the Hindi language or any other Indian languages. Using some early study for languages such as Bengali and Hindi. Das and Bandopadhya [2] prepared a Bengali SentiWordnet (a dictionary that includes the sentiment scores of word). A word level lexical-exchange system has been connected to every passage in the English SentiWordNet utilizing an English-Bengali Word reference to acquire a Bengali SentiWordNet.

To understand the sentiment of a word four procedures were discussed by Das and Bandopadhya [3]. The

first approach for determining the sentiment was an interactive game was proposed that annotates the words with their respective polarity. In the second approach, bilingual dictionary of English and Indian languages was used to assign the polarity. In the third approach, WordNet was used to assign the polarities. In the fourth approach they decided the polarity, using pre-annotated corpora. Das and Bandopadhya [4] recognized enthusiastic expressions in the Bengali corpus. They arranged the words in six feeling classes with three sorts of intensities to perform sentence level annotation. A fallback procedure was proposed by Joshi et al. [5] for the Hindi language. Using three methodologies: Machine Translation, Resource Based Assumption Analysis and Language Sentiment Analysis. In this system, a lexical resource of Hindi SentiWordNet (HSWN) was created, utilizing its English format. H-SWN (Hindi-SentiWordNet) was created by lexical resources such as English and English-Hindi WordNet. English SentiWordNet words were supplanted by their equivalent words in Hindi to get H-SWN by utilizing Wordnet. The precision of their test was 78.14%. By considering a framework, Bakliwal et al. [6] generated a word reference. They used fundamental graph traversal of the antonym words and Proportionate word further will be used to generate the subjectivity vocabulary. 79% precision is achieved by the proposed algorithm in order of surveys and gives 70.4% simultaneity with public reviews. Mukherjee et al. [7] explains about the model updates that combine pack of-words in talk markers with the slant demand by 4% exactness. Bakliwal et al. [8] suggested depicting Hindi reviews as positive, neutral and negative. They figured out another score breaking point and used it for two different techniques. Moreover, they used a fusion of the POS Tagged Ngram and central N-gram approaches.

In a different study Ambati et al. [9] proposed a way to deal with known errors in tree banks (text corpus that explains syntactic or semantic sentence structure). The suggested technique can decrease the validation time. They experimented with Hindi data and could see a 76.63% rate of errors at the dependency level. Arora et al. [10] described a diagram based system which is used to collect a subjective dictionary for Hindi, using WordNet. The subjective vocabulary of the Hindi language is made with dependence on WordNet. Initially they considered a small wordlist containing some opinion words using WordNet and then added antonyms and synonym of those words and updated the wordlist. Wordnet like a diagram which is being crossed by words where each word in a Wordnet was seen as a center point, which is then combined with antonyms and similar words. They achieved 74% exactness and 69% precision when synchronization with public comments in Hindi. Gune et al. [11] implemented the parsing of the Marathi language and then built a parser which has a Chunker and Marathi POS tagger. In their described framework, morphological analyzers provide the ambiguity and suffixes for extracting feature sets. Mittal et al. [12] generated an efficient approach to identify the sentiment from Hindi content. They built up Hindi language corpus by adding more opinion words and improve the present Hindi SentiWordNet (HSWN). Their algorithm showed 80% precision on the course of action of studies.

B. Sentiment Analysis Using Twitter

Recent research based on sentiment analysis says that the analysis of opinion utilizes simultaneous learning. Pak and Paroubek in [13] utilized tweets which end with emoticons like ":)" ":-)" as positive and ":(" ":-(" as negative. They accumulated models including Max Entropy, Support Vector Machines (SVM) and Naive Bayes and concluded that SVM performed the best amongst various others, attaining more precision which lead SVM to be the best performer of all the classifiers. They recorded that all distinctive models were beaten by the unigram model. To gather subjective information, they compile the tweets ending with emoticons comparatively as Go et al. In [14]. To attain the target result, they moved Twitter records of well understood papers like The New York Times etc to a database. They concluded that both bigrams and POS help (regardless of results displayed in [2]). Both bigram and POS methods are categorized by n-gram models.

Birmingham and Smeaton [15] tested two distinct strategies, Multinomial Naïve Baye's (MNB) and SVM for web pages and scale blog. They found that MNB methodology outperforms SVM on scaled scale areas with short substance. Wang and Can et al. [16] build a reliable structure for the 2012 US races to recuperate political suppositions at work using Twitter. In the present systems, they are considering real time tweets, keeping location as filter and then analyzing people's sentiments.

Previously, the polling data has been considered as the best estimator for forecasting electoral results. But, recently the polling data has been considered partial and erroneous. Therefore, a research, investigated the accuracy of polls in comparison with sentiment analysis results performed on Twitter tweets. Surprisingly, the results reported that Twitter was 3.5% more biased in popular votes and 2.5% more biased in state results when compared with the polls. The study concluded that the predictions based on Twitter data are worse when compared with polling data [47]. To examine the effectiveness of the previously proposed methods for predicting electoral results i.e. sentiment analysis, by applying on a new dataset of Twitter tweets related to politics. The

dataset of 234,697 tweets was collected using the Twitter streaming Twitter API. The data preprocessing was performed on the collected tweets by removing the hashtags, links and names of accounts. Emotions and similes were replaced by full form i.e. "-" to. The study reported that there have been many limitations attached to the previous methods due to which they were inadequate in predicting electoral results using social media data. It has been suggested that to improve the accuracy of predictions the researchers should not only rely on the polarity of words alone. The approaches of POS Tagging, Sense disambiguation should be adopted in preprocessing along with considering the contextual and lexical features of words [48]. The traditional techniques of predicting election results i.e. polling have become unreliable with the frequently changing technology. Due to the increased use of smartphone and easily available internet facility, the social and digital media has become the platform for presenting political views. Numerical comparisons have been done for the slogans used in Twitter tweets for US elections 2016 and visualized using WordCloud. The results reported by the study were inconsistent then the actual outcomes of the elections and it were suggested by the researchers to further consider and evaluate qualitative aspects for making electoral predictions. The Trump win was not predicted by the Twitter for the states of Michigan and Wisconsin by any of the approaches being utilized by the study [49]. A study explored the relationship between the size of the social network of candidates and the chances of winning the elections using regression analysis and data collected from Facebook and Twitter. Three models were proposed in which the number of votes was taken as a dependent variable and number of Facebook connections, along with other factors were taken as independent variables. The results reported the size of the network and the chances of a win are significantly correlated to each other. Hence, the election results were reported to be inferred by looking at the size of social network. However, the magnitude of the effect is very small and the social media data was reported to be predictive for elections with close competition. The proposed model is elaborated in Figure 2.1 [50].

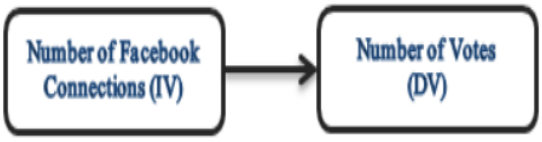

Figure 2.1: Regression model representation.

Figure 2.1: Regression model representation IV & DV Social network techniques i.e. volumetric analysis, sentiment analysis, has been utilized by authors to evaluate the predictive power of Twitter data for inferring electoral results for three countries, Pakistan, India, and Malaysia. The data preprocessing was performed on approximately 3.4 million Tweets collected using Twitter streaming API. To separate the tweets in English language a natural langue toolkit of python was used. 90% of the Tweets from Pakistan and India are English, but on the other hand only 23% of the tweets were in English. The equation (1) shows volumetric analysis approach, where Volx for any party x represents the volume of tweets and Cx represents tweets count.

Equation (2) shows sentiment analysis whereas equation (3) represents social network analysis approach. In equation (2) $Sent_t$ is for sentiment of tweets pos_t represents positive tweets where as neg_t represents negative tweets. In equation (3) the centrality score of a party x is represented by Net_x out of n number of parties and the raw score is represented by s ९.The performance was measured using Mean Absolute Error (MAE) and is represented by equation (4). The results reported that the Twitter data was not effective for making election predictions for Malaysia, but in the case of Pakistan and India, it appeared as an effective and efficient for electoral predictions. By combining multiple techniques the proposed model for predicting electoral outcomes was also effective for candidates and parties having small vote count [51].

$$Vol_x = \frac{c_x}{\sum_{j=1}^{n} c_j}\% \qquad (1)$$

$$Sent_t = \{\, 1, pos_t > |neg_t| - 1, pos_t < |neg_t|, pos_t = |neg_t| \,\} \qquad (2)$$

$$Net_x = \frac{s_x}{\sum_{j=1}^{n} s_j}\% \qquad (3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i| \qquad (4)$$

The Facebook data has not been much utilized for predicting electoral results. However, a research predicted the congressional polls by using a proposed model, Senate Vote = *f* (partisan voting index + incumbency + participation advantage) and data collected from Facebook. Where, senate vote is the percentage of forecasted votes won by either two of the major parties, Partisan Vote Index (PVI) is the past election results. The metric from Facebook was used to calculate the incumbency and participation advantage. The incumbency was added in the proposed Facebook model to overcome the limitation of PVI. The participations factor was used to enable the model for forecasting trends. It represents the real-time success rate of each political campaign. The components attached with the participation variable include likes, active post engagements, and time slices. The track of

fans and the post engagements has been continuously kept by the Facebook pages. The metrics used by the research are expected to be influential for conducting political campaigns to outreach potential supporters. The results of the study reported that Facebook data accurately predicted the outcomes for US Senate elections 2012. It is recommended to further investigate the usefulness of Facebook data for predicting electoral outcomes to verify the effectiveness and accuracy of the proposed model [52].

REFERENCES

[1] O. Almatrafi, S. Parack, and B. Chavan, "Application of Location-Based Sentiment Analysis Using Twitter for Identifying Trends Towards Indian General Elections 2014," Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication, Article No. 41, Jan. 2015.

[2] A. Das and S. Bandyopadhyay, "SentiWordNet for Indian languages," Proceedings of the 8th Workshop on Asian Language Resources, pp. 56–63, Aug. 2010.

[3] A. Das and S. Bandyopadhyay, "SentiWordNet for Bangla," Knowledge Sharing Event-4: Task, Volume 2, 2010.

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

[5] A. Joshi, B. A. R, and P. Bhattacharyya, "A fall-back strategy for sentiment analysis in Hindi: a case study," Proceedings of ICON 2010: 8th International Conference on Natural Language Processing, Dec. 2010.

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

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

[8] A. Bakliwal, P. Arora, A. Patil, and V. Verma, "Towards enhanced opinion classification using NLP techniques," Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP, pp, 101–107, Nov. 2011.

[9] B. R. Ambati, S. Husain, S. Jain, D. M. Sharma, and R. Sangal, "Two methods to incorporate local morphosyntactic features in Hindi dependency parsing," Proceedings of the NAACL HLT 1st Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL), pp. 22–30, June 2010.

[10] P. Arora, A. Bakliwal and V. Verma, "Hindi Subjective Lexicon Generation using WordNet Graph Traversal," International Journal of Computational Linguistics and Applications, Vol. 3, No. 1, pp. 25–39, Jan-Jun 2012.

[11] H. Gune, M. Bapat, M. M. Khapra and P. Bhattacharyya, "Verbs are where all the action lies: Experiences of shallow parsing of a morphologically rich language", Proceedings of the 23rd International Conference on Computational Linguistics, pp. 347–355, Aug. 2010.

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

[13] A. Pak, and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), pp. 1320–1326, May 2010.

[14] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, Stanford University, pp. 1–12, 2009.

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

[16] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp 115–120, July 2012.

[17] Anuta, D., Churchin, J., & Luo, J. (2017). Election bias: Comparing polls and twitter in the 2016 us election. ArXiv preprint arXiv: 1701.06232.

[18] Chung, J. E., & Mustafaraj, E. (2011, April). Can collective sentiment expressed on twitter predict political elections? In AAAI (Vol. 11,pp. 1770-1771).
Hinch, J. (2018). # Make Americas Polls Great Again: Evaluating Twitter as a Tool to Predict Election Outcomes. The Geographical Bulletin, 59(1), 45-54.

[19] Cameron, M. P., Barrett, P., & Stewardson, B. (2016). Can social media predict election results? Evidence from New Zealand. Journal of Political Marketing, 15(4), 416-432.

[20] Jaidka, K., Ahmed, S., Skoric, M., & Hilbert, M. (2018). Predicting elections from social media: a three-country, three-method comparative study. Asian Journal of Communication, 1-21.