TWITTER BASED PREDICTION AND ANALYSIS OF ELECTION

Shwetha BM
Dept. Of CSE, MTECH
SSIT, Tumkur

Dr. H Venugopal
Professor, Dept. Of CSE, MTECH
SSIT, Tumkur

***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 language. We performed data (text) mining on tweets collected over a period of a time that referenced three national political parties in India, during the campaigning period for general state elections in 2019. We made use of both supervised and unsupervised approaches. We utilized Dictionary Based, Naive Bayes 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.**

Keywords—*Sentiment Analysis; Twitter; Indian Elections; Naive Bayes;*

1. 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 congress and labelled them as negative, neutral and positive. The aim of the paper was to analyze trends in the Indian General Election 2019 using location as a filter. They employed a supervised approach by applying a Naïve Bayes classifier. 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. 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.

2. LITERATURE REVIEW

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

Data mining is a wide area, but there have not been many experiments done in the Hindi language or any other Indian languages. 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 EnglishHindi 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% simultaniety 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 treebanks (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 Hindilanguage 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.

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.

3. OUR EXPERIMENT

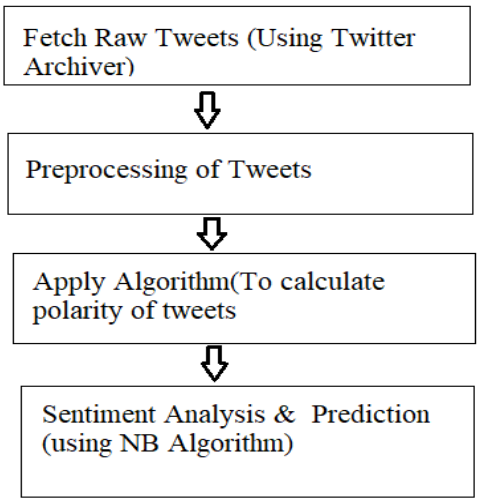

Fig1: Steps and techniques used in experiment

A. Data Collection

We collected tweets corpus utilizing Twitter Archiver [17]. It was collected using Google Spreadsheet which established the connection to Twitter using a Google script by finding key details from a Twitter account and importing all the search results. It connected to Twitter every few minutes and fetched the recent tweets. The query was placed in the Twitter archiver using tweets as a filter. The tweets all discussed different Indian political parties. Using this information, we were able to identify the number of tweets for and against the different political parties.

B. Prepocessing

Text processing is a major phase of text mining for data analysis. Preprocessing includes the following steps such as remove website urls, remove hashtags, twitter mentions(), stopwords, emoticons and special character and punctuations[17].

C. Negation Handling

In every language there are certain words like "No"," Not" in which can revert meaning of the sentence. So, these words also help in finding the polarity of tweets.

D. Algorithm Used

We used a supervised approach such as classification algorithms Naïve Bayes, unsupervised approach as dictionary based. We took tweets with the names of Indian political parties such as BJP, congress etc.

a) Dictionary Approach

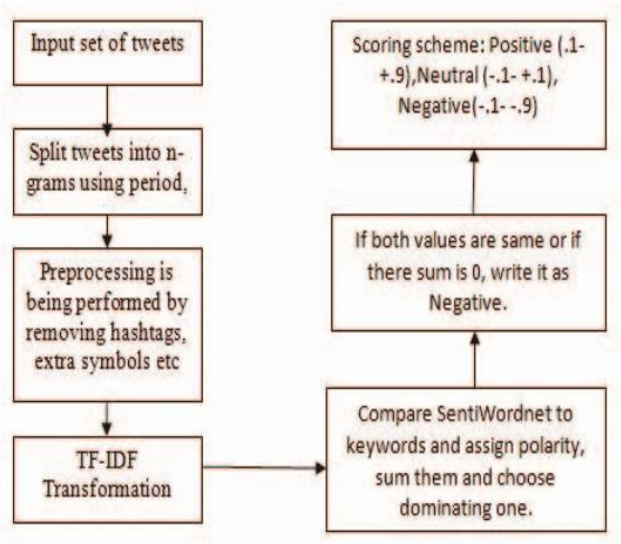

Fig2: Flow diagram of using Dictionary Based Approach.

Initially we had to construct SentiWordNet which contained synonyms and antonym of the respective words along with the score of that particular word. The polarity of tweets was calculated by above method.

b) Naïve Bayes classifier

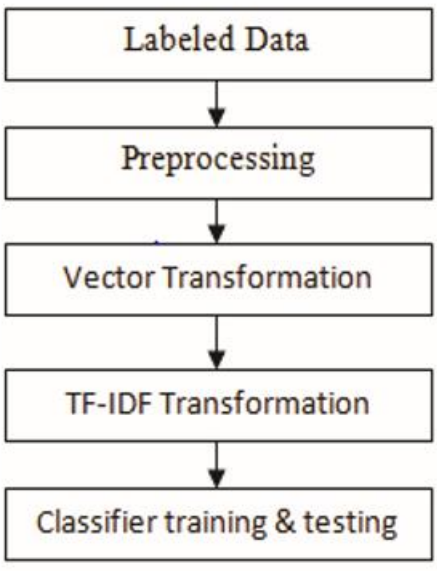

Fig3: Steps and techniques used in sentiment classification

It is a simple probabilistic classifier based on the Baye's theorem. It assumes every feature is independent of each other. To assign labels for every input vector features is utilized using the formula below.

P (label | feature) =P(label)*P (feature | label)/P(features)

4. RESULT

We collect data through twitter API, after that we performed Preprocessing on the data. For collected data for the Indian election we classified polarity. Classifying tweets in three categories positive, negative and neutral.

In analysis of sentiment analysis in Indian election, we considered two parties BJP (Bhartiya Janta Party) and Indian National Congress party. After fetching the tweets from tweeter API, we apply the sentiment analysis for the captured tweets.

The below table present the analysis of tweets of BJP.

Total Count of Tweets	124
Positive Count of Tweets	100
Negative Count of Tweets	2
Neutral Count of Tweets	25
Undecidable Count of Tweets	4

The below table present the analysis of tweets of congress.

Total Count of Tweets	124
Positive Count of Tweets	90
Negative Count of Tweets	10
Neutral Count of Tweets	35

Undecidable Count of Tweets	8

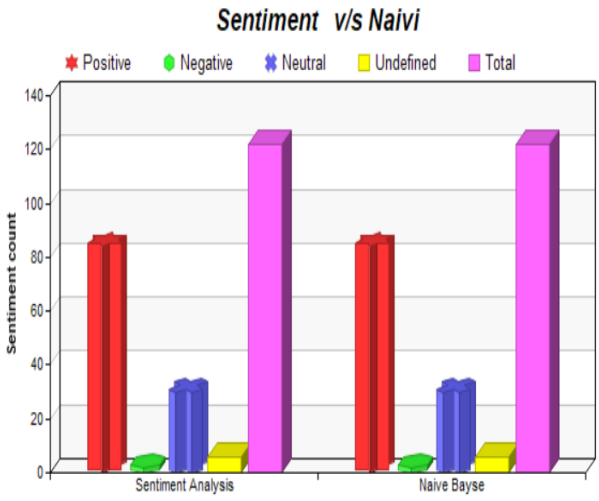

Chart: Indian election sentiment analysis

Finally, we show analysis of both the parties and high chances of which party may win.

5. CONCLUSION

As It is very difficult to predict the result of elections using other methods, including prevalence of social media, such as Facebook and Twitter, utilize sentiment analysis of Twitter tweets to predict the results of the Indian General election.

6. FUTURE WORK

There could be many other prospective areas to conduct this research in, including the data from other big social media sites like facebook to increase the size of the data set. We have more space to work with the training dataset such as considering the sample dataset in which the certain number of features of an algorithm is already defined. More machine learning algorithms such as Regression, Random forest can also be considered for classification and further prediction.

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. Varma, “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. Varma, "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. Varma, “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. Bermingham, 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] Wong, F. M. (2013). Quantifying Political Leaning from Tweets and Retweets. ICWSM.

[17] Boutet, A. K. (2012). What's in your Tweets? I know who you supported in the UK 2010 general election. Proceedings of the International AAAI Conference on Weblogs and Social Media.

[18] Golbeck, J. &. (2011). Computing political preference among twitter followers. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

[19] Pennacchiotti, M. &. (2011). Democrats, republicans and starbucks afficionados: user classification in twitter. Proceedings of the 17th ACM SIGKDD international conference on

[20] Tumasjan, A. S. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM, 10, 178-185.