Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter

Parul Sharma
Department of Computer Science
San Jose State University
San Jose, CA, USA
parul.sharma@sjsu.edu

Teng-Sheng Moh
Department of Computer Science
San Jose State University
San Jose, CA, USA
teng.moh@sjsu.edu

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 42,235 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 2016. We made use of both supervised and unsupervised approaches. We utilized Dictionary Based, Naive Bayes and SVM 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 results of the analysis for Naive Bayes was the BJP (Bhartiya Janta Party), for SVM it was the BJP (Bhartiya Janta Party) and for the Dictionary Approach it was the Indian Nathional Congress. SVM predicted a 78.4% chance that the BJP would win more elections in the general election due to the positive sentiment they received in tweets. As it turned out, BJP won 60 out of 126 constituencies in the 2016 general election, far more than any other political party as the next party (the Indian National Congress) only won 26 out of 126 constituencies.**

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, e-commerce 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.

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

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, Marathi 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% 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 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 helps (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.

III. OUR EXPERIMENT

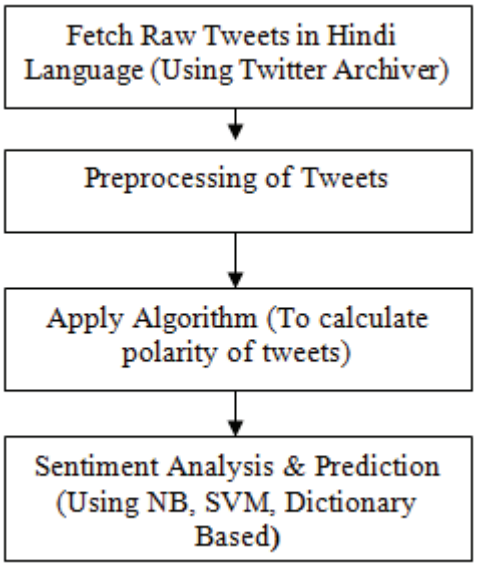

Fig. 1. Steps and techniques used in Experiment.

A. Data Collection

We collected a Hindi 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 into the Spreadsheet. It connected to Twitter every few minutes and fetched the recent tweets. The query was placed in the Twitter archiver using Hindi tweets as a filter. The tweets all discussed different Indian political parties. For each party, we collected all the tweets which used hashtags such as #BJP (भारतीय जनता पार्टी), #Congress (कांग्रेस), #BSP (बहुजन समाज पार्टी), #NCP (राष्ट्रवादी कॉंग्रेस पक्ष), #AAP (आम आदमी पार्टी). Using this information, we were able to identify the number of tweets for and against the different political parties.

B. Preprocessing

Text preprocessing 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 characters and punctuations [17].

C. Negation Handling

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

D. Algorithms Used

We used a supervised approach such as classification algorithms Naive Bayes, Support Vector Machine and unsupervised approach as Dictionary based. We took tweets with the names of Indian political parties such as #BJP, #Congress, #NCP, #AAP. We collected a total of 23,998 tweets relevant to these hashtags.

a) Dictionary Approach

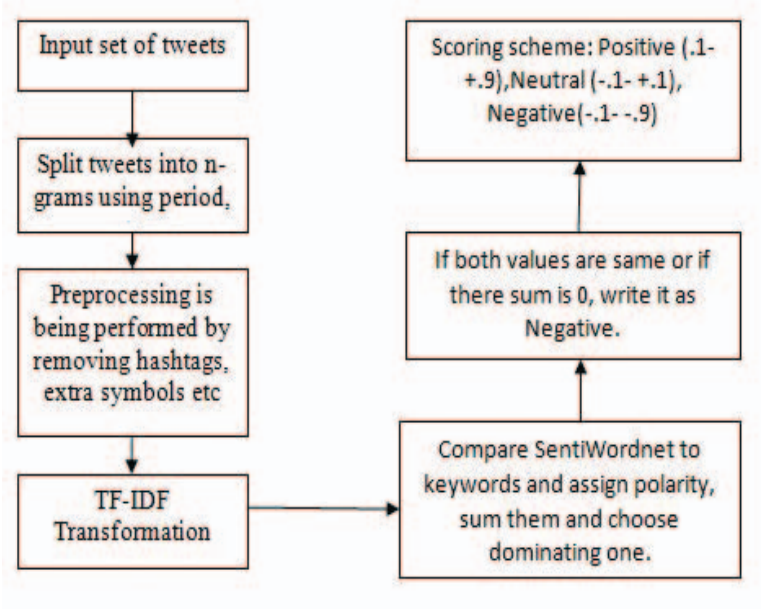

Fig. 2. Flow diagram of using Dictionary Based Approach.

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.

After performing the steps above, we were left with the following table:

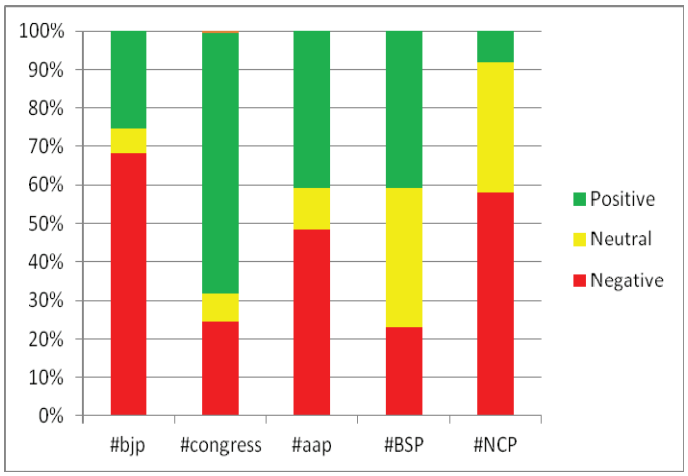

Fig. 3. Percentage of number of positive, negative and neutral tweets using Dictionary Based Approach.

The above chart shows the political parties and their positive, negative and neutral percentages. According to this algorithm and the above figure Congress has 68% positive tweets, giving them more likelihood of winning the elections in the year 2016.

b) Naïve Bayes Classifier

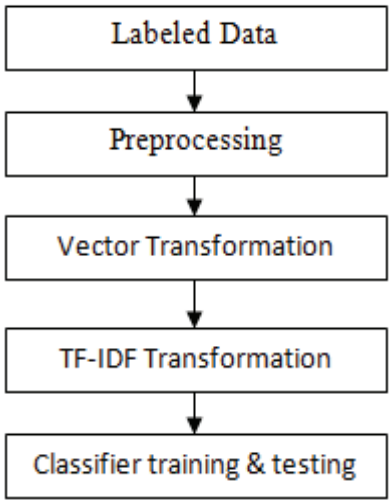

Fig. 4. Steps and techniques used in sentiment classification.

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.

$$P(\text{label} \mid \text{features}) = \frac{\{P(\text{label}) * P(\text{features} \mid \text{label})\}}{P(\text{features})}$$

Label in the above equation shows the polarity or sentiment i.e. positive, neutral and negative, and features are the words which have been extracted from the tweets.

TABLE I. Total number of tweets for each party.

Parties	Total Tweets
BJP	13,612
Congress	6567
AAP	12,975
BSP	2116
NCP	2116

We fetched a total of 42,345 tweets. After preprocessing, we were left with 36,465 tweets. We classified them using NB classifier. For further calculations, we manually labelled the dataset of 36,465 tweets and then performed 5-fold cross validation. In the cross validation method, the process of training and testing was repeated 5 times utilizing 80% of the dataset as training data and the remaining 20% as testing data. After that, the average accuracy of classifiers was obtained.

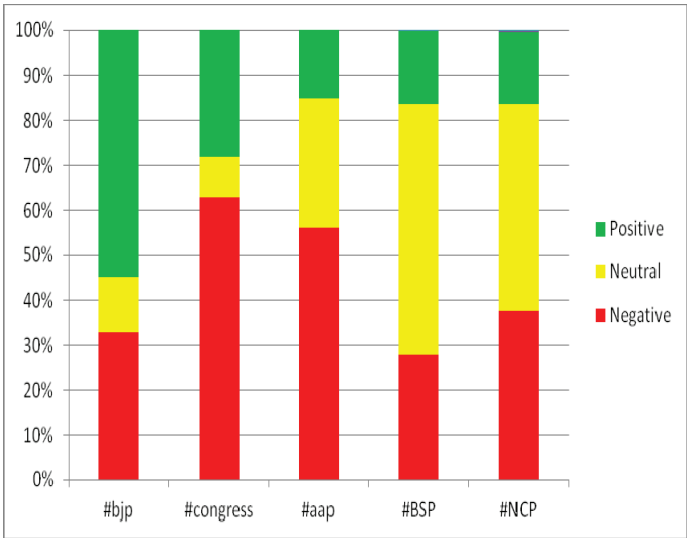

Fig. 5. Percentage of number of positive, negative and neutral tweets using NB classifier.

The above chart shows the political parties and their positive, neutral, and negative percentages. The algorithm gave an accuracy of 62.1%. According to this algorithm and the above figure, BJP with 55% positive tweets had a greater likelihood of winning the elections.

c) Support Vector Machine

It is a learning system utilizes hypothesis space in high dimensional feature space. It is more considered where the number of samples is smaller in number from the number of dimensions.

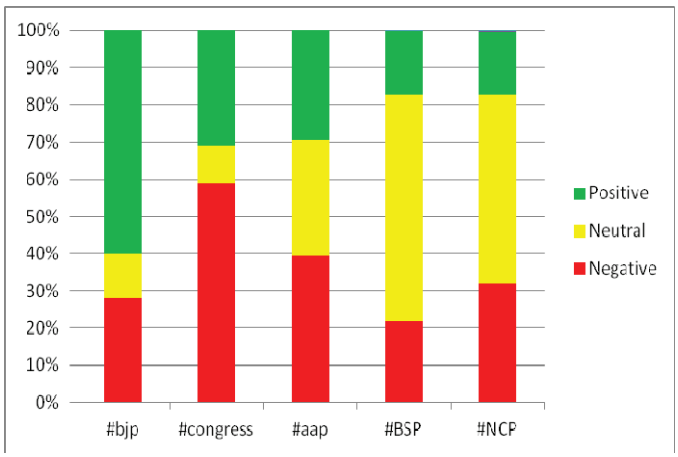

Fig. 6. Percentage of number of positive, negative and neutral tweets using the SVM classifier.

The above chart shows the political parties and their positive percentages, neutral percentages, and negative percentages. The accuracy of this algorithm was 78.4%. According to this algorithm and the above figure, BJP with 60% positive tweets had a greater likelihood of winning the elections in the year 2016.

IV. Result and Conclusion

As it is very difficult to predict the results of elections using other methods, including public opinion polls, and with the growing prevalence of social media, such as Facebook and Twitter, the authors decided to utilize sentiment analysis of Twitter tweets to predict the results of the Indian general election.

TABLE II. Accuracy of algorithm.

Algorithm	Accuracy
Naive Bayes	62.1%
SVM	78.4%
Dictionary Based	34%

As shown in the above table the accuracy of the Naive Baye's algorithm was 62.1% and the accuracy of Support Vector Machine was 78.4%. We made our final prediction utilizing SVM, since the accuracy of the algorithm is higher. We predicted that the party that had a better chance of winning the 2016 general election is BJP.

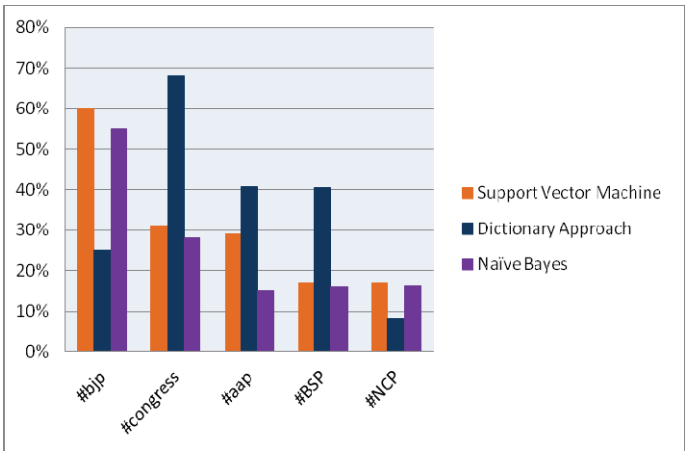

Fig. 7. Comparison of Positive polarity of all the three algorithms.

TABLE III. Precision Recall of algorithm

Algorithm	Precision	Recall
Naive Bayes	.71	.61
SVM	.75	.78

We also calculated the precision and recall as shown in the above table. The result of the Naive Baye's algorithm was .71 and .61. For Support Vector Machine we obtained .75 as precision and .78 as recall. As it turned out, the BJP did win 60 out of 126 of the constituencies in India in 2016.

V. Limitation

The limitation of our research is that we did not considered the emoticons which are also a relevant aspect when defining the polarity of a tweet. Since the data was labelled manually the amount i.e. 36,465 was not large enough to provide more accurate results, so we can fetch more tweets and then label them. In the future, we can also increase the size of the Hindi SentiWordnet.

VI. 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] 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] Y. Sharma, V. Mangat and M. Kaur, "A Practical Approach to Sentiment Analysis of Hindi Tweets," Preceedings of the 1st International Conference on Next Generations Computing Technologies (NGCT), Sept. 2015.