International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-2, July 2020

Multilingual Lexicon based Approach for Real-Time Sentiment Analysis

Swati Sharma, Mamta Bansal

Abstract: ***The information on WWW has mounted to a greater height, overriding to fledgling analysis in the direction of sentiments using Artificial Intelligence. Sentiment Analysis deals with the calculus exploration of sentiments, opinions and subjectivity. In this paper, multilingual tweets are analyzed for identifying the polarities of various political parties like AAP, BJP, Samajwadi, BSP and Congress; so that the users will get an idea that to which party they should give their vote. The data is being analyzed using Natural Language Processing. Using different smoothening techniques, noise is removed from data, classified by using Machine learning algorithms and then the accuracy of the system is gauged using various evaluation precision measures. The central premise of this research is to benignant common people and politicians both. For common people; is for deciding their precious vote, to which party to give will be good for themselves and nation too. For politicians; they will have an idea about themselves i.e. after seeking the polarities of different parties, the politicians will have an idea which party is preferable and which is not preferable, so that the politicians can work accordingly. The system shows comparison among VADER and SVM algorithm; and SVM algorithm showed 90% accuracy.***

Keywords: Lexicon, NLP, SVM, VADER

I. INTRODUCTION

In this work, a framework is proposed for identifying the polarity of election parties like Bhartiya Janta Party (BJP), Smajwadi, AAP (Aam Aadmi Party), BSP (Bahujan Samaj Party) and Indian National Congress Party to analyze which party is more preferable using Natural Language Processing Lexicon based approach by collecting tweets of multilingual; this is mainly concerned with communication between machines and humans. Natural Language Processing and Machine Learning are uproots of Artificial Intelligence. Artificial Intelligence; these are its two forks. They work conjunctively for solving data problems of analysis [1]. Natural Language Processing is a field of Artificial Intelligence and Computer Science. It accord with computational linguistics; deals with the reciprocation among natural language and machine language and further perturbed with training machines to successful analysis of large amount of data. Machine Learning is central to Artificial Intelligence as it entitles machines for getting them self train without blunt programming. When new data set is being provided to the machines, they learn and get train by themselves [2]. Sentiment Analysis is procedure of computing, recognizing and equating opinions manifested in text, for identifying the subjective information of text and polarity of the text as negative or positive. Lexicon based NLP approach; Machine Learning approach and combination of Machine Leaning plus Natural Language Processing are perspectives toward execution of Sentiment Analysis. Attribute Selection, Filtering, Stemming, Lemmatization, Feature Generation and applying Statistical classifier are the prime steps for analysis of sentiments [3]. Sentiment Analysis, also referred as OM (Opinion Mining) is a category of text mining which measures the inclination of an individual's opinion. Sentiment Analysis gives attention to identify the individual behavior in respect to any topic whether he/ she is feeling positive or negative or neutral [4]. To classify individual's opinion different supervised learning techniques and unsupervised learning approaches are used [5]. Multilingualism is an important challenge to handle in the arena of Sentiment Analysis (SA). People from different state, different religion and different country use different languages and while posting an individual uses the language the way they talk. So, the tweets are of variant languages and it is a matter of concern to consider all the languages. If the tweets of a single language say, English are considered as an input and on those input complete methodology is applied of Sentiment Analysis (SA), then the output received will not give a surety that it is completely correct [6]. Until unless, we will not consider the tweets of different languages, the result will not going to be accurate. So, to have an accurate and correct output, multilingualism needs to be considered. In this, the humans program the machines to process the huge size of data and on the dependency of the programming done, the system responds [7]. The data being collected from twitter via Twitter Streaming API. The real significance of work is that it works on real time tweets; real time tweets means the tweets posted just at the exact moment of time and anywhere in the whole world. If we wish to run an application for data analysis, the real time Data analysis is the best course of action to engulf all the needs. As if we are working on real time data, the output we receive out of it is going to be most accurate and most updated [8].

Manuscript received on May 25, 2020.
Revised Manuscript received on June 29, 2020.
Manuscript published on July 30, 2020.
***** Correspondence Author
Swati Sharma*, PhD Scholar, Shobhit University, AP at MIET, Meerut, India. E-mail: swati.sharma.it@miet.ac.in
Mamta Bansal, Professor at Shobhit University, Meerut, India. E-mail: mamta.links@gmail.com

If we analyze the data by collecting the database, the output will be on the basis of the database collected but as we Indians changes our choice, our mindset very frequently by going with the flow so our opinions vary from time to the time and so our tweets vary. So, the best option is to analyze real time tweets.

We live in multilingual country and generally people post their opinions in their native language so to consider multilingualism is a major concern.

II. METHODOLOGY

A. Data Collection

Tweepy helps to mine data of Twitter user. It is installed using pip. For real time downloading of tweets, Streaming API is being used. It is basically used for collecting huge quantity of Twitter data and for making the live feed by user stream. There is a difference among Rest API and Streaming API. By using Rest API, the data is pulled from the twitter whereas by using the Streaming API, the data is pushed to a steady session [9]. Thus, Rest API downloads less quantity of data in comparison to Streaming API [10]. There is an instance of Twitter, named as tweepy.Stream, which is used to substantiate a streaming network and further forward the tweets to an instance, named as StreamListener instance. on_data method of the instance StreamListener collects all the messages and further make a call to function, depending on the type of messages [11]. These methods act as the stubs.

B. Data Cleaning

The initial step of transforming unstructured data to structured data is the data cleaning. Following points needs to cover in process of data cleaning: First is removal of uninformative tweets i.e. the tweets which are very short in length that it seems it doesn't contain any useful information should be removed. Second is removal of redundant tweets i.e. after contrasting the tweets with other ones; the tweets having 90% – 95% having same content are eradicated. Third is removal of punctuations and emoticons, as they do not play vital role in process of analysis. Moreover, when the tweets are extricated, the emoticons display as a square box instead of actual emoticons. Thus, the square boxes are treated as a trash. By using 'replace' keyword the emoticons as square boxes are removed. Fourth is removal of retweets i.e. some tweets contains the keyword 'RT' which means retweet or posting the same tweet again, which does not play significance in analysis purpose. Thus, retweets should be removed. Fifth is removal of URL's. URL is uniform Resource Locator which contains the link to some other webpage or website. While analysis, the URL provides no useful information and thus they are removed. Last but not the least is transforming the complete database in to lower case. Already data is inconsistent i.e. randomly arranged with no sequence. So to make it suitable for analysis purpose, case sensitivity plays a vital role.

C. Training and Testing

After cleaning the data, it is trained and tested around its axis. By using the crowd sourcing data, the hand tagged subset of data is being prepared. It is technique of developing the dataset with the aid of large number of people.

By assembling data using crowd sourcing technique, practitioners can estimate valuable, disperse and plenty amount of data in low cost. It permits to assemble real time data and can get a lot more innumerable and extensive observations. It permits researchers to easily reach people and different places and providing insight to researchers for various events. It guarantees large number of contributors; for successful crowd sourcing it is essential. It keeps the network growth with a good rate.

III. RESULTS AND DISCUSSION

The positive polarity and negative polarity indicating how likely the party is, has been shown by analyzing the tweets in the below presented graphs. Fig. 1 shows the polarity of tweets in related to BJP. The x-axis in the graph represents the number of tweets identified for BJP and y-axis represents the reviews detected by sentiment analysis process In Fig. 2, the x-axis shows the number of tweets in favor of Congress party, the y-axis shows the reviews based on tweets i.e. 4500 tweets refer positive tweets and 2000 refers negative polarity. We can notice that the Congress is not as likely in Canada as it is shows negative polarity. In the output graph Fig. 3, the x-axis shows the number of tweets in favor of AAP party, the y-axis shows the reviews based on tweets i.e. 9500 tweets refer positive tweets and 2000 refers negative polarity. We can notice that the AAP is not as likely in Chicago as it is shows negative polarity In the output graph Fig. 4, the x-axis shows the number of tweets in favor of BSP party, the y-axis shows the reviews based on tweets i.e. 1200 tweets refer positive tweets and 200 refers negative polarity. In the output graph Fig. 5, the x-axis shows the number of tweets in favor of Samajwadi party, the y-axis shows the reviews based on tweets i.e. 7500 tweets refer positive tweets and 2000 refers negative polarity.

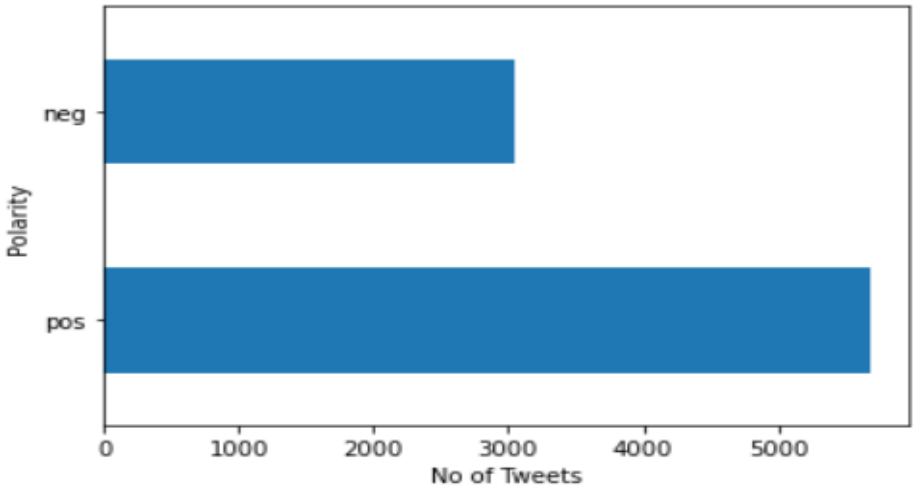

Fig. 1 Polarity of BJP Party

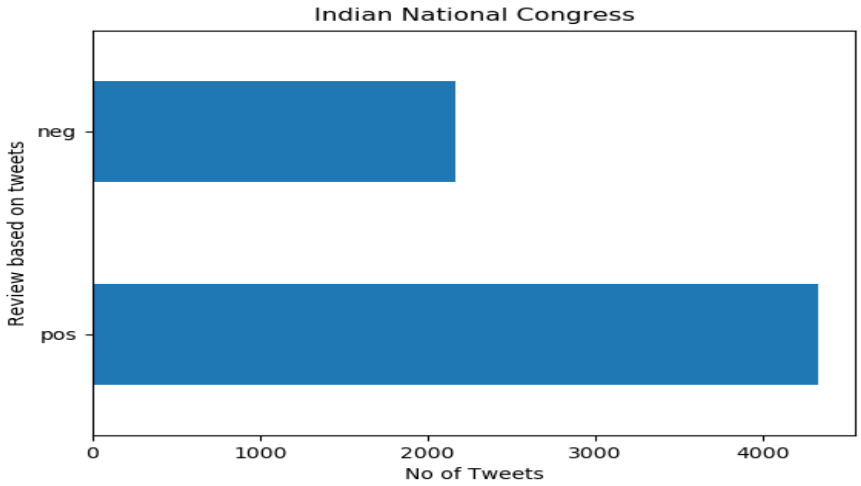

Fig. 2 Polarity of Congress Party

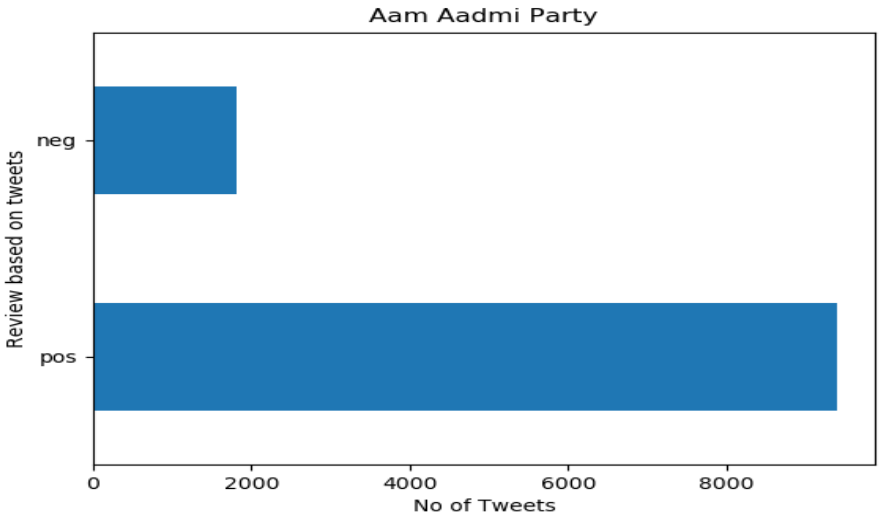

Fig. 3 Polarity of AAP

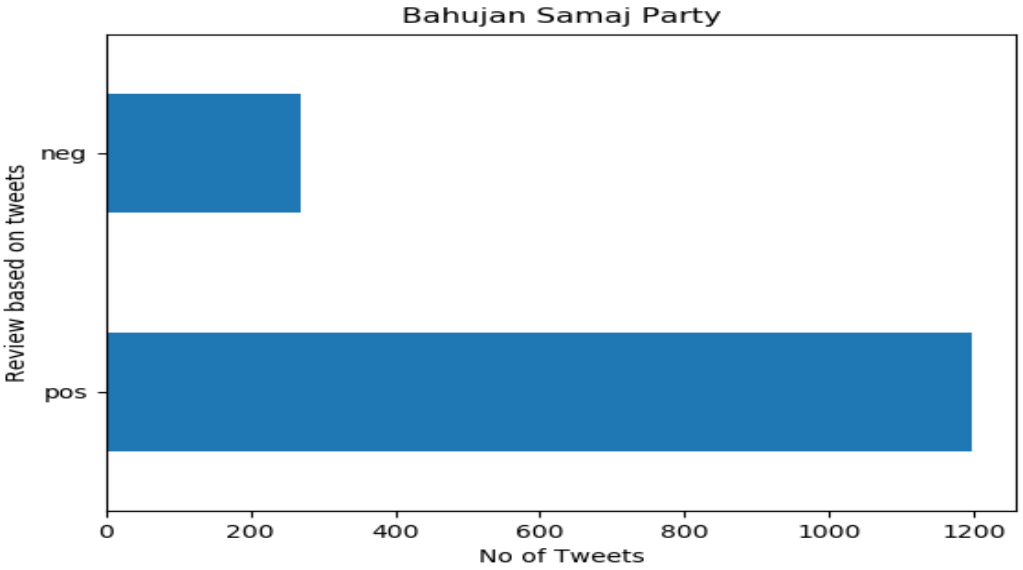

Figure 4 Polarity of BSP

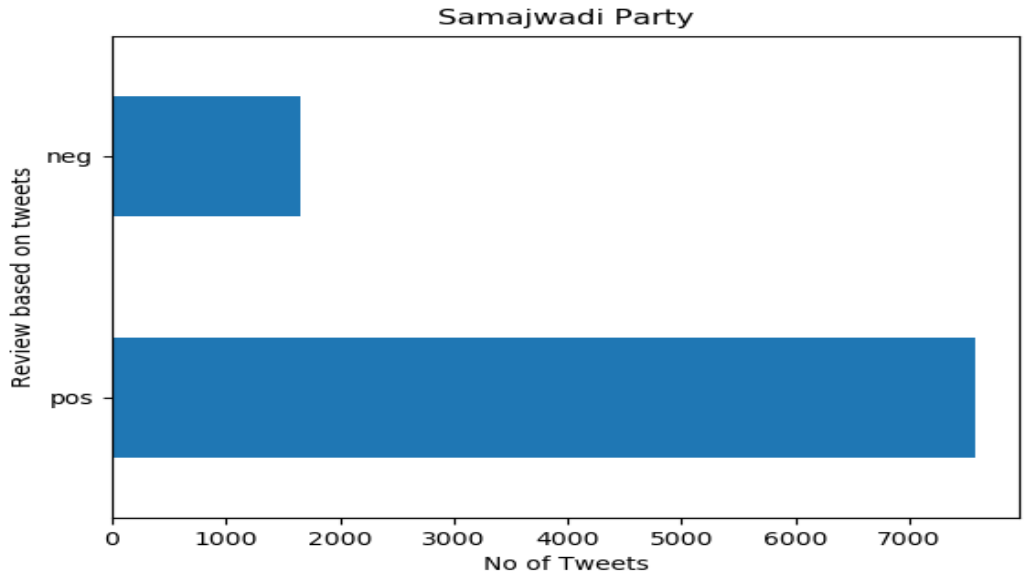

Figure 5 Polarity of Samajwadi Party

Support Vector Machine algorithm, it produces an optimal output for identifying the maximum margin hyper plane classes. Over fitting problem is specially taken care by SVM method. The classes which are having single feature are classified using linear demission surface.

The classes which are having binary features are classified by drawing line among classes and the class which are having multiple features; they are classified by drawing hyper planes. The classification rate of SVM is 90%. The confusion matrix using SVM is shown in Fig. 6 and evaluation parameters in Table I.

VADER, sentiment analysis tool is preferred for analyzing social media sentiments. The confusion matrix using VADER is shown in Fig. 7 and evaluation parameters in Table II.

8623 (FN)	**555 (TN)**
949 (TP)	**4513 (FP)**

Fig. 6 Confusion matrix using SVM

The graph shown in Fig. 8 represents the accuracy of SVM and VADER. The classification rate of SVM is 90% whereas that of VADER is 65%.

Table-I: Evaluation measures of classifier using SVM

	Precision	Recall	F1- measure	Support
Neg	0.90	0.93	0.91	9178
Pos	0.89	0.82	0.85	5462
Macro avg	0.89	0.88	0.88	14640
Weighted avg	0.89	0.89	0.89	14640
Classification Rate	0.90			

Table-II: Evaluation measures of classifier using VADER

	Precision	Recall	F1- measure	Support
Neg	**0.90**	**0.50**	**0.65**	9178
Pos	**0.52**	**0.90**	**0.66**	5462
Macro avg	**0.71**	**0.70**	**0.65**	14640
Weighted avg	**0.76**	**0.65**	**0.65**	14640
Classification Rate	**0.65**			

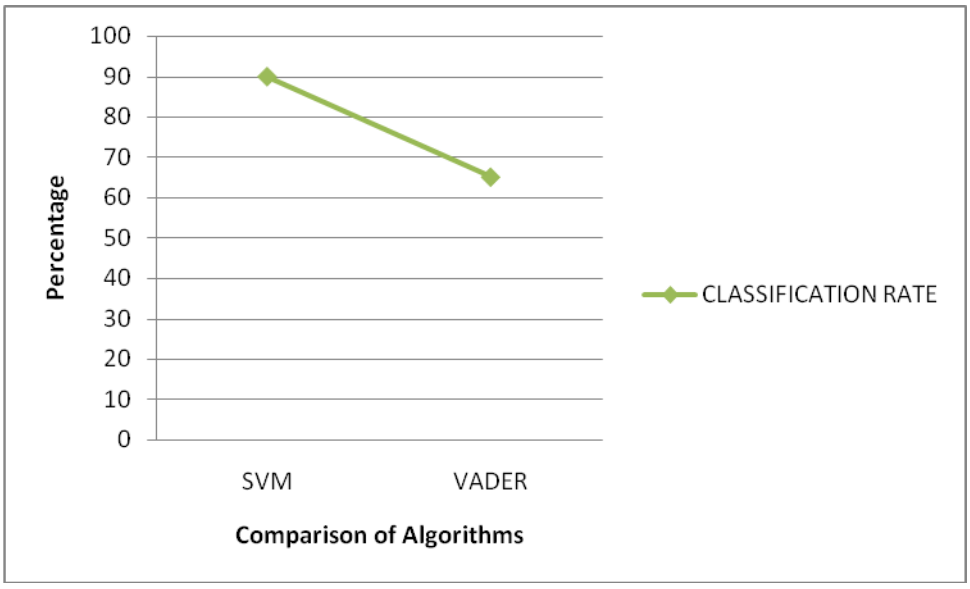

Fig. 8 Comparison of classification rate among SVM and VADER

4627 (FN)	**4551 (TN)**
526 (TP)	**4936 (FP)**

Fig. 7 Confusion matrix using VADER

The classification rate of SVM is 90% whereas that of VADER is 65%.

IV. CONCLUSION AND FUTURE SCOPE

In this work, lexicon based polarity identification technique has been presented for identifying the polarities of few parties such as AAP, BSP, BJP, INC and Samajwadi party as positive and negative using NLP (Natural Language Processing) algorithms. This experiment was conducted on real time and multilingual tweets for identifying the positive polarity and the negative polarity of the parties. The result shows that comparatively which party is suitable and which party is not suitable. According to the results, SVM showed more accuracy than VADER. In future work, all the election parties can be included and much of the work can be done considering sarcasm, satire and irony.

REFERENCES

1. Bing Liu, Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies Morgan and Claypool Publishers May (2012).
2. B.J. Jansen, M. Zhang, K. Sobel and A. Chowdury, Micro-blogging as online word of mouth branding., CHI'09 Extended Abstracts on Human Factors in Computing Systems, Boston, MA, USA, (2009) : 3859-3864.
3. Arti Buche, Dr.M.B.Chandak, Akshay Zadgoanakar, Opinion Mining and Analysis: A Survey, International Journal on Natural Language Computing (IJNLC) Vol 2 No 3 June (2013) : 39-48.
4. S. Asur and B.A. Huberman, Predicting the future with social media, Proc. of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Washington, DC, USA: IEEE Computer Society, vol. 1 (2010) : 492-499.
5. Bucur Cristian, Aspects regarding detection of sentiment in web content, International Journal of Sustainable Economies Management (IJSEM), Volume 3, issue 4, p.24-32, ISSN: 2160-9659 (2014).
6. S Chandrakala, C M Sindh, Opinion Mining and Sentiment Classification: A Survey, SOCO DOI: 10.21917/ijsc.2012.0065 (2012).
7. Amrita Kaur, Neelam Duhan, A Survey on Sentiment Analysis and Opinion Mining, International Journal of Innovations & Advancement in Computer Science, IJIACS USSN 2347 – 8616 Volume 4, May (2015).
8. B.O. Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, Proc. of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 122–129 (2010) : 122-129.
9. B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, Proc. of the Conference on Empirical Methods on Natural Language Processing, (2002) : 79-86. Huizhi Liang , Umarani Ganeshbabu , Thomas Thorne. A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution, Page(s): 54164 – 54174, 06 March 2020
10. Huyen Trang Phan , Van Cuong Tran , Ngoc Thanh Nguyen , Dosam Hwang . Improving the Performance of Sentiment Analysis of Tweets Containing Fuzzy Sentiment Using the Feature Ensemble Model, Page: 14630 – 14641, 03 January 2020, ISSN: 2169-3536, INSPEC Accession Number: 19313387
11. L. Belcastro, R. Cantini, F. Marozzo, D. Talia, and P. Trun_o. Discovering political polarization on social media: A case study, in Proc. 15th Int. Conf. Semantics, Knowl. Grids, Guangzhou, China, 2019, pp. 1_8.
12. K. Jaidka, S. Ahmed, M. Skoric, and M. Hilbert. Predicting elections from social media: A three-country, three-method comparative study, Asian J. Commun., vol. 29, no. 3, pp. 252_273, May 2019.

AUTHORS PROFILE

Swati Sharma, B.tech(Honors.),M.Tech (Honors.), Ph.D. pursuing from Shobhit university. Assistant professor in MIET, Meerut since 2010.

Dr. Mamta Bansal, Ph.d from Shobhit University, Meeut, Professor in Shobhit University, Meerut.