International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-3, February 2019

Sentiment Analysis of Twitter Accounts using Natural Language Processing

Nafees Akhter Farooqui, Ritika, Aayush Saini

Abstract*: *Every hour heaps of data are generated by blogs, social websites, and web pages. Many business houses gather all of this data to understand consumers, marketing strategies and their desires better and make appropriate changes to reshape the way businesses work. To extract information from this content, we need to rely on natural language processing (NLP) techniques. Many organizations want to get an overview of any policy or any product launched in the market. The overview of human sentiment can be calculated by using natural language processing through Python as it is a strong and easy language which is spreading across the globe covering its track in every sphere of modern technology.

Index Terms: Modern Technology, Natural Language Processing, Python, Sentiments.

I. INTRODUCTION

At last few years, an enormous number of people are involved in the social networking sites like Twitter, Facebook, Instagram. These sites express their belief, emotions and the opinions about the personalities and the places. There are various methods used for sentiment analysis. These methods are categorized predominantly as Artificial intelligence, natural language processing, statistical and knowledge-base are based on different methods. It is challenging research to analyze the sentiments and opinions computationally [1]. Therefore, it extracts the information from the available data through the twitter account for the prediction of marketing, political elections, business analysis, communication, research, and educational solutions. Sentiment analysis can be obtained through the behavioral analysis of social and commercial tweets on the tweeter accounts [2]. Current research had proved [3,4] the people's vision, perceptions and choices get from the Twitter accounts and some other social networking sites. An algorithm [5] had been proposed to manipulate the emotions from tweets. It considered a huge amount of data for sentiment analysis. Kanavos proposed a method to identify the social communities with behavioral factors [6] and assigned a metric value to each user's sentiments posts. In this paper, we analyze the sentiments and emotions of users in different aspects like an election, business, education, etc. Thesesentiments are collected from different Twitter profiles. The analysis of the emotions of the different users based on the different aspects. We also validate the results of the sentiments by different classifiers. The experimental results show that the polarity score of the different sentiments. In this paper, we build a model to classify the sentiments of the most popular blogging sites like Twitter into positive, negative and neutral sentiment.

Manuscript published on 28 February 2019.
***** Correspondence Author (s)
Nafees Akhter Farooqui*, Department of Computer Applications, DIT University, Dehradun, India,
Ritika, Department of Computer Applications, DIT University, Dehradun, India,
Aayush Saini, Department of Computer Applications, DIT University, Dehradun, India,

II. LITERATURE SURVEY

In general sentiment analysis is applied to the Twitter data that can be handled with the Natural Language Processing. The analysis of Twitter data is based upon the classification level to the learning of the words and phrases. The classification of Twitter messages is similar to the analysis of sentiments at sentence level [12]. However, the casual and informal languages used in tweets, the Twitter sentiment analysis is a unique task in microblogging domains. The problem in microblogging domain is how one can work with sentiment analysis techniques on the well-formed data [13,14,15,16]. Many researchers include the part-of-speech features, but results remain diverse. They investigate in several ways like automatically collecting training data from tweepy API. The mining sentiment is based on two main approaches namely Dictionary Based (DB) and Machine Learning Based (MLB) that is shown in Table I. The DB technique uses the predefined dictionary for the classification of the sentiments but has limitation to classification. In the DB based system, there are lack of the domain-based semantics due to the use of the Bag-of-words concept. In contrast, MLB systems have the domain-specific training data for the sentiment classifications. The linguistic dissimilarities and class imbalance problem in the text can be solved by the bootstrapping technique [17]. Coletta et al. [18] explain the combination of SVM and cluster classification of Twitter data. Bollen et al. demonstrated societies attitude and emotion on social and commercial news-based tweets on Twitter accounts. Kouloumpis et al. [19] suggested a framework for Twitter sentiment analysis by manipulating the irregularities in casual performs. Thus, he uses the hashtags to classify the tweets as like and dislike.

Table I: Assessment of Current Methodologies

	Dictionary Based	Machine Learning Based
Approach	classifying individual words are Dictionary based.	Probabilistic classification is implemented through Supervised or Unsupervised learning.
Advantage	Computational overhead will be minimum.	Customization of the work by more suitability for specific domains.
Limitation	Lack of domain-based classification capability.	There are class imbalance and linguistic dissimilarities problems arises.

Phan Ngoc and Myungsik Yoo [20] suggested an Artificial intelligent Neural Network technique, which is based on the ranking of content. For example, the Facebook page of the user has different polarity content. Ana Minanovic [21] has proposed a data collection method that is used for sentiment analysis. It uses the KNIME for online reviews and tweets analysis. Alexander Porshnev et al. [22] analyze the tweets of stock market data by a combination of the Neural Networks and Support Vector Machine. Christos Troussas [23] identify the emotions of specific status on Facebook by using the Naive Bayes algorithm. Therefore, this type of analysis is based on content. Comparison of the advantages and limitations of the existing mechanism are given in Table II.

This mechanism is based on the feature vector and plain sentiment text mining. The emotions and sentiments can be measured by the polarity report of the comments.

Table II: Evaluation of Existing Mechanism

Mechanism	Advantage	Limitation
[20]	Lexicons & Emoticons are taken for the analysis.	The absence of a Domain Context
[21]	Comparative analysis of Online polling and social reviews.	Shortage of rich data
[22]	Uses the Combination of SVM and Neural Network.	Only work on the stock market data.
[23]	Naive Bayes classifier	Starting point procedure

To overcome these limitations, we use the feature vectors for the sentiment analysis. The imbalance problem can be solved by multiple training models over different subsets of the same dataset. Comparative analysis of the two political parties is based on the actual Twitter data [7], mined from Twitter accounts by using Tweeter API [8]. They were used Senti WordNet [9] and WordNet [10] sentiment analyzers to find positive and negative scores. Data collection and mining process had been performed by the Twitter streaming API, which is used for the prediction of presidential elections. This API is used for the understanding of the public opinion. In this activity firstly collect the data from the tweets and remove the retweet after the automatic buzzer detect the repetition of tweet then breaking them into several sub-tweets and measure the sentiment polarity of the election. We also find the result of the election in the form of the positive and negative scores, but there are some errors which can be measured regarding the mean absolute error (MAE) [11].

III. DESIGN AND IMPLEMENTATION

We have proposed a system that explained the process of the gathering of data, sentiment analysis, and classification of Twitter opinions. Great works and tools are focusing on text mining on twitter. In this paper, the wealth of available libraries has been used. We consider the opinion of the current political views by the posted tweet of users in the form of hashtags. Then we store the tweets in the database and pre-process these datasets (set of tweets of users). After that divide the datasets into the training and testing samples. Here the 1000 tweets are taken as training samples, and 400 tweets are test samples. Content Polarity and subjectivity are calculating using Text Blob. Then apply the Natural Language processing method to build a score checking module. This module is used to assign and check the sentiment score for each tweet. Then visualize and test the module. The framework of sentiment-analysis is shown below in Figure 1.

The Framework contains the following modules:

- Retrieval Module.
- Pre-Processing Module.
- Polarity Calculation Module.
- Score Checking Module.

A. Retrieval module

Retrieval module collects the public opinion in the form of the hashtags which represents the views about political parties. To retrieve views Tweepy API [24,25] is used.

B. Pre-Processing Module

In this step, irrelevant Twitter opinions are removed. Also, redundancy of the tweets are checked and removed before analyzing the sentiments. The data undergo the following processes:

- Removal of retweet.
- Noise data removal.
- Emotion tagging.
- POS tagging.
- Feature vector creation.

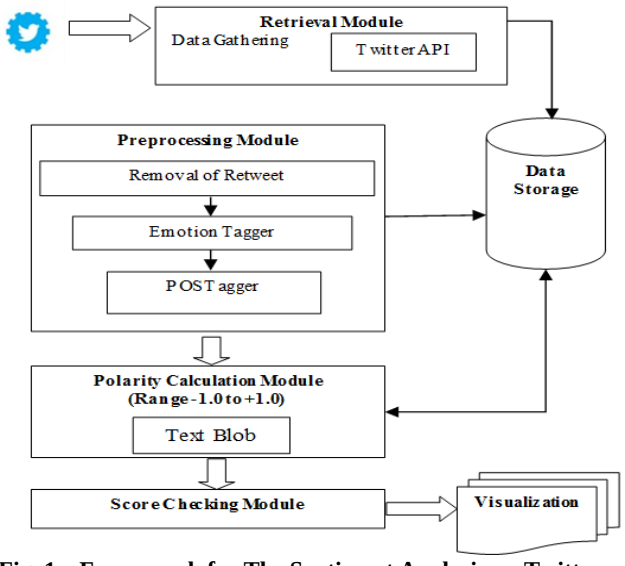

Fig. 1. Framework for The Sentiment Analysis on Twitter Accounts

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

C. Polarity calculation module

This Module detects the emotions or opinions from a large unstructured formatted data. Three polarity classes are categorized as negative, neutral and positive. The polarity score is ranging from -1 to +1 based on the tweets, where -1 score means a negative sentiment and a +1 score means a positive sentiment while the zero value is considered as a neutral sentiment. The partiality is also measured by assigning a score from 0 to 1 where a value near to 0 represents impartial and near to 1 is partial. TextBlob [26] have the simple features of natural-language processing. That is used for the measurement of polarity and partiality calculation of tweets.

D. Score checking Module

To validate and visualize the result in the form the graph and histogram obtained from TextBlob, a program is written to analyze the Twitter data.

IV. EXPERIMENTAL SETUP

The code is written in python by using the following algorithm:

Algorithm:

1. Import all the necessary libraries, i.e., tweepy, textblob, sys, os, matplotlib.
2. Establish and authenticate the twitter developer account for accessing twitter's inbuilt functionality using tweepy API.
3. Input: search_term ← string (take input a string from user to be searched for, can be any word/hashtag).
4. Input: cnt ← int value (number of tweets to be fetched).
5. Call the search method of tweepy API by passing the values of search_term and cnt.
6. Set api.wait_on_rate_limit := True to wait if the fetching limit is exceeded.
7. Set tweets: = total_tweets: = retweets := 0
8. Set positive_counter = negative_counter = neutral_counter = 0
9. Create a list tweets_grabbed for catching duplicate retweets.
10. Create files:
    1. data_file for storing the polarity data from tweets.
    2. temp_tweets_file for temporarily storing tweets for checking of duplicate tweets (not retweet).
    3. main_tweets_file for storing tweets in a structured way for later reading and understanding.
11. While api_wait_on_rate_limit != False:
    for tweet in public_tweets:
    if tweets == cnt:
    set api.wait_on_rate_limit := False
    end if
    else: if tweet.text in temp_tweet_files:
    print "Duplicate tweet"
    retweets += 1
    continue the loop
    end if
    else if 'RT' in first 2 indexes of tweet.text:
    if retweet already in tweets_grabbed:
    break
    end if
    else:
    add that retweet to tweets_grabbed
    end else
    end else if
    else:
    convert tweet.text → TextBlob(twee.text) and set to analysis_data
    if analysis_data has polarity>0 and subjectivity>0.6:
    increment positive_counter and tweet
    append that tweet to tweets_grabbed
    write polarity on data_file and tweet on temp_tweets_files
    end if
    else if analysis_data has polarity<0 and subjectivity>0.6:
    increment negative_counter and tweets append that tweet to tweets_grabbed write polarity on data_file and tweet on temp_tweets_files
    end else if
    else if analysis_data has polarity>0 and subjectivity>0.6:
    increment neutral_counter and tweets append that tweet to tweets_grabbed write polarity on data_file and tweet on temp_tweets_files
    end else if
    end else
    increment total_tweets
    end else
    end for
    end while
12. Close all the files and delete temp_tweets_file.
13. calculate the percentage of positive, negative and neutral tweets.
14. plot the data using matplotlib.

V. RESULTS AND DISCUSSION

After preprocessing of the datasets taken from Tweepy API apply, the above algorithm has evaluated the 20 years of classic KKHH datasets by using the Python after 1st run and 2nd run. Kejriwal Calls for donation data sets for their sentiment analysis is analyzed, and the results are visualized. All the visualization results show that up trend and low trend patterns that represents the positive and negative sentiments of the users.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

20Years Of Classic KKHH – 1st Run

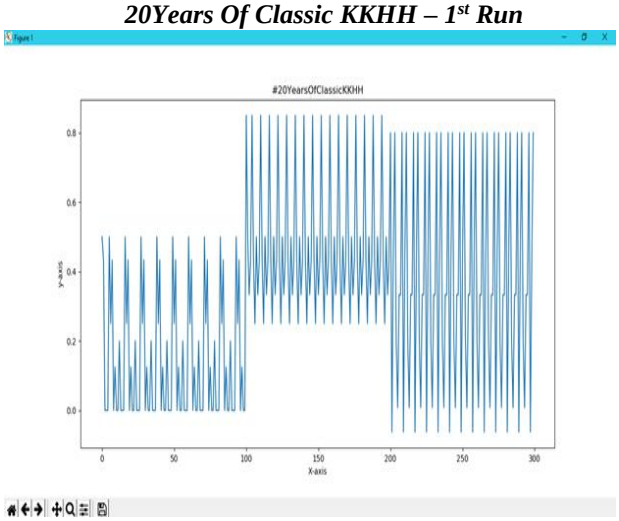

Fig. 2. 1st Run: Output(raw)

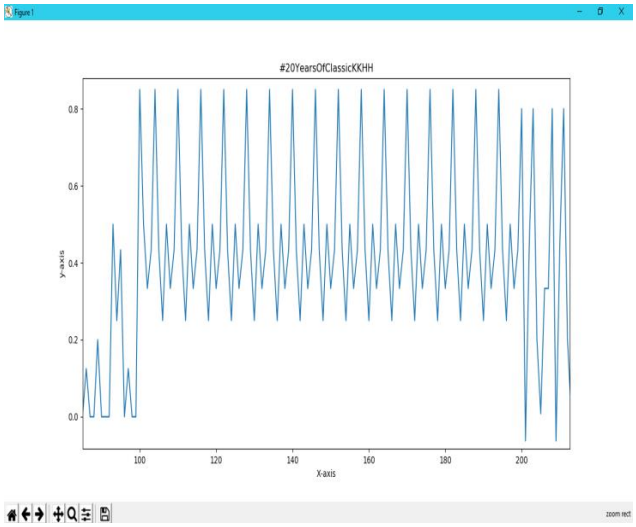

Fig. 3. 1st Run: Output (Curated distinct data and trend)

20 Years of Classic KKHH – 2nd Run

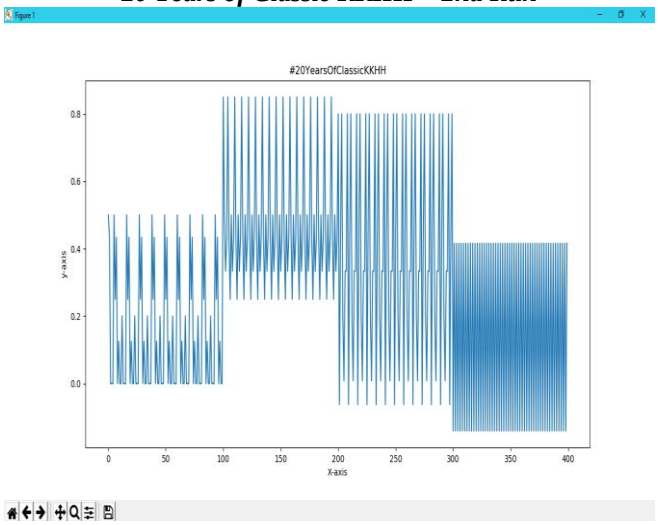

Fig. 4. 2nd Run: Output(raw)

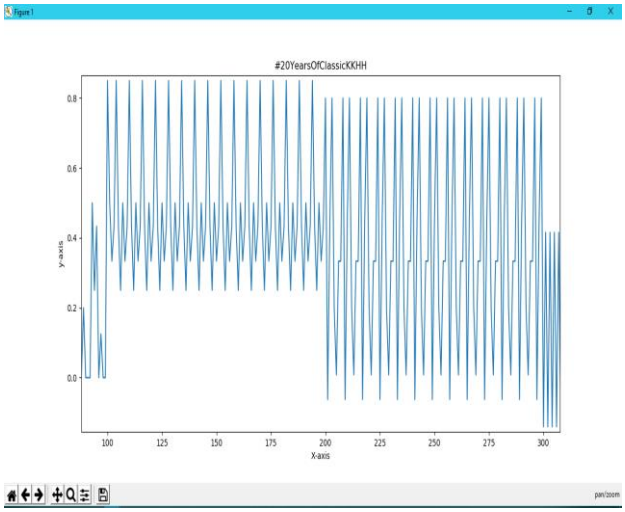

Fig. 5. 2nd Run: Output (Curated distinct data and trend)

Kejriwalcalls-4Donation – 1st Run

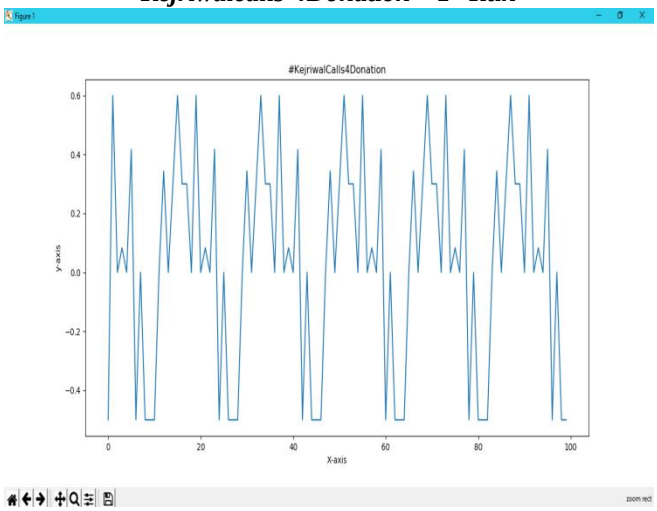

Fig. 6. 1st Run: Output(raw)

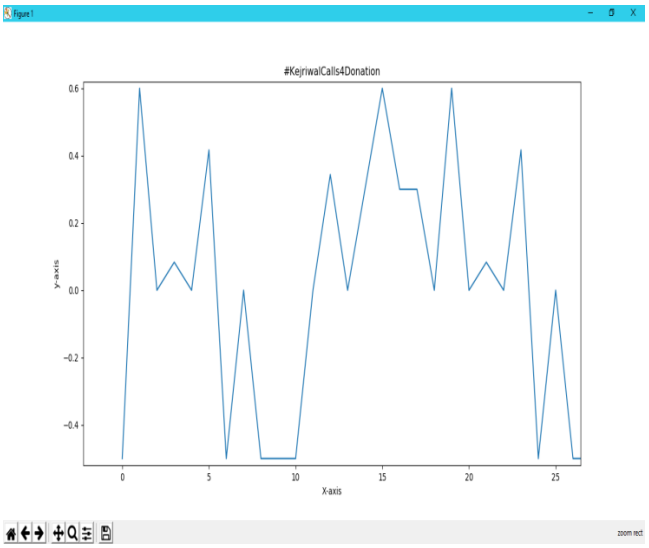

Fig. 7. 1st Run: Output (Curated distinct data and trend)

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Kejriwal Calls-4 Donation – 2nd Run

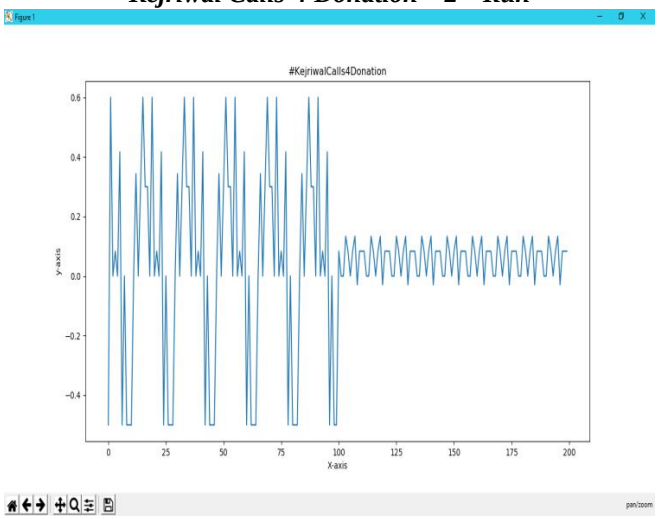

Fig. 8. 2nd Run: Output (Raw)

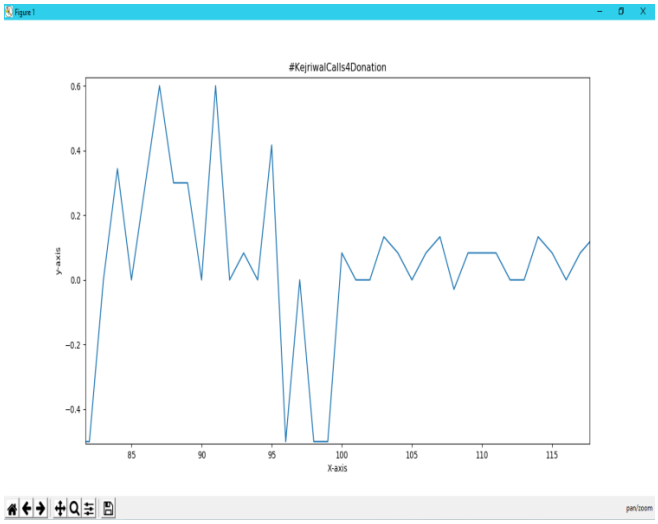

Fig. 9. 2nd Run: Output (Curated distinct data and trends)

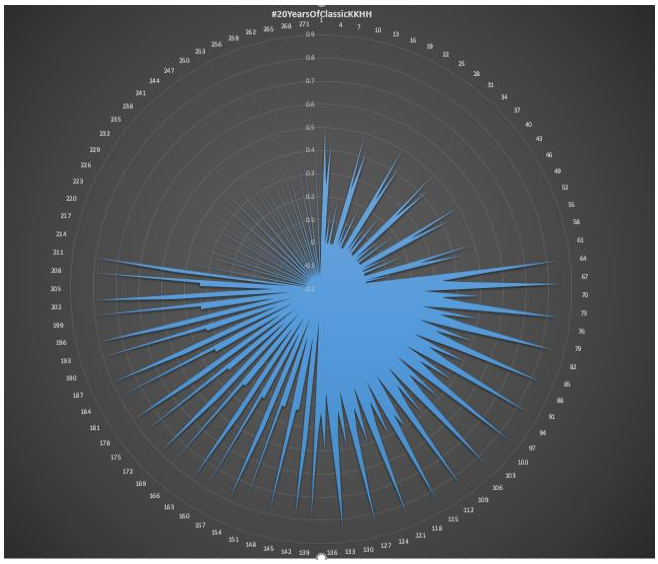

Fig 10. Visualization of the Score of 20 Years Classic KKHH

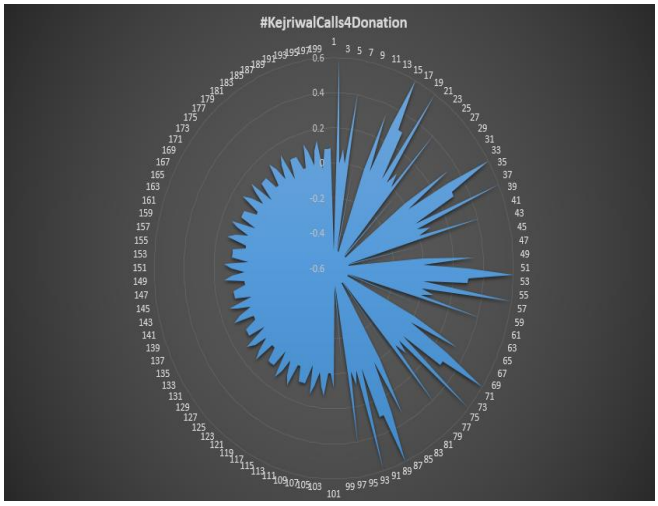

Fig 11. Visualization of the Score of Kejriwal Calls for Donation

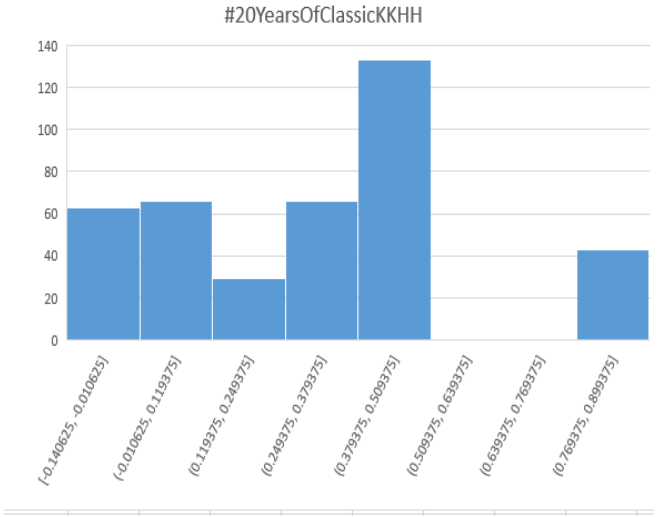

Fig 12. Histogram Chart of 20 Years Classic KKHH

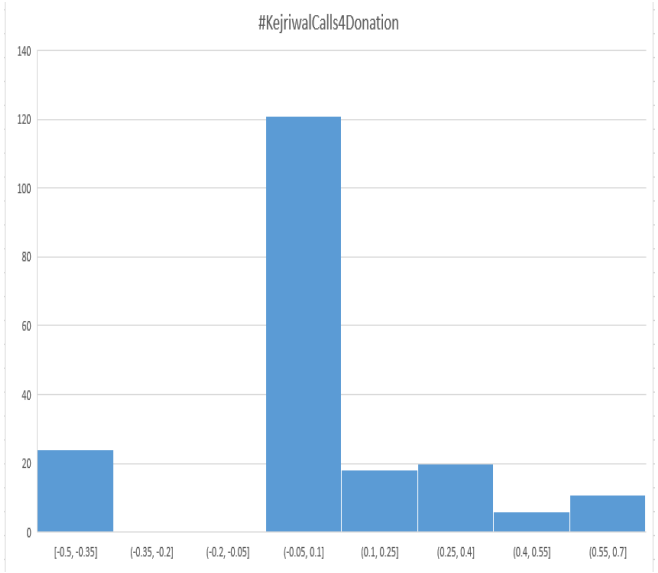

Fig 13. Histogram Chart of Kejriwal Calls for Donation

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

VI. CONCLUSION

The projected framework gathers data from the twitter and uses natural language processing techniques to extract features. Then natural language processing is applied to the data to classify the sentiments as Positive, Negative and Neutral. Polarity and partiality are also calculated by the dictionary, that consists of a semantic score of a tweet. It is observed that natural language processing is a better method for sentiment analysis as compared to traditional methods. There are some limitations in natural language processing, so in future, another machine learning and data mining techniques may be used to eliminate the limitations of the given feature vectors and their selections. The future work will be focused on the Multilingual Machine learning algorithm that handles the different types of task and easily classifies the data in groups and score which is based on sentiments on real-time data.

ACKNOWLEDGMENT

The author thanks the DIT University, Dehradun for providing the research grant to support this research work. The corresponding author wishes to thanks Prof K.K. Raina and Prof. B. S. Panwar for the great cooperation and motivation for this research.

REFERENCES

1. Jagdale, O.; Harmalkar, V.; Chavan, S.; Sharma, N. Twitter mining using R. Int. J. Eng. Res. Adv. Tech. 2017, 3,252–256.
2. Anjaria, M.; Guddeti, R.M.R. Influence factor-based opinion mining of Twitter data using supervised learning. In Proceedings of the 2014 Sixth International Conference on Communication Systems and Networks, Bangalore, India, 6–10 January 2014; pp.1–8.
3. Miranda Filho, R.; Almeida, J.M.; Pappa, G.L. Twitter population sample bias and its impact on predictive outcomes: A case study on elections. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015; pp. 1254–1261.
4. Castro, R.; Kuffó, L.; Vaca, C. Back to# 6d: Predicting venezuelan states political election results through twitter. In Proceedings of the 2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG), Quito, Ecuador, 19–21 April 2017; pp. 148–153.
5. Kanavos, A.; Nodarakis, N.; Sioutas, S.; Tsakalidis, A.; Tsolis, D.; Tzimas, G. Large scale implementations for twitter sentiment classification. Algorithms 2017, 10, 33. [CrossRef]
6. Kanavos, A.; Perikos, I.; Hatzilygeroudis, I.; Tsakalidis, A. Emotional community detection in social networks. Comput. Electr. Eng. 2017, 65, 449–460. [CrossRef].
7. Jose, R.; Chooralil, V.S. Prediction of election result by enhanced sentiment analysis on twitter data using word sense disambiguation. In Proceedings of the 2015 International Conference on Control Communication & Computing India (ICCC), Trivandrum, India, 19–21 November 2015; pp. 638–641.
8. Twitter Apps. Available online: http://www.tweepy.org/ (accessed on 26 February 2018).
9. Esuli, A.; Sebastiani, F. Sentiwordnet: A High-Coverage Lexical Resource for Opinion Mining; Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR): Pisa, Italy, 2006.
10. Miller, G.A.Wordnet: A lexical database for English. Commun. ACM 1995, 38, 39–41. [CrossRef].
11. Dokoohaki, N.; Zikou, F.; Gillblad, D.; Matskin, M. Predicting Swedish elections with Twitter: A case for stochastic link structure analysis. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Paris, France, 25–28 August 2015; pp. 1269–1276.
12. Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135.
13. Tumasjan, A.; Sprenger, T. O.; Sandner, P.; and Welpe, I. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In Proceedings of ICWSM.
14. O'Connor, B.; Balasubramanyan, R.; Routledge, B.; and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of ICWSM.
15. Barbosa, L., and Feng, J. 2010. Robust sentiment detection on twitter from biased and noisy data. In Proc. of Coling.
16. Bifet, A., and Frank, E. 2010. Sentiment knowledge discovery in Twitter streaming data. In Proc. of 13th International Conference on Discovery Science.
17. A. Hassan, A. Abbasi, and D. Zeng, "Twitter sentiment analysis: A bootstrap ensemble framework," in Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013, pp. 357–364.
18. F. Coletta, N. F. F. d. Sommaggio Silva, E. R. Hruschka, and E. R. Hruschka, "Combining classification and clustering for tweet sentiment analysis," in Intelligent Systems, 2014 Brazilian Conference on. IEEE, 2014, pp. 210–215.
19. E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!" ICWSM, vol. 11, pp. 538–541, 2011.
20. P. T. Ngoc and M. Yoo, "The lexicon-based sentiment analysis for fan page ranking in Facebook," in Information Networking (ICOIN), 2014 International Conference on. IEEE, 2014, pp. 444–448.
21. A. Minanovic, H. Gabelica, and Z. Krstic, "Big data and sentiment analysis using knime: Online reviews vs. social media," in Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on. IEEE, 2014, pp. 1464–1468.
22. A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from Twitter sentiment analysis," in Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE, 2013, pp. 440–444.
23. C. Troussas, M. Virvou, K. J. Espinosa, K. Llaguno, and J. Caro, "Sentiment analysis of Facebook statuses using naive bayes classifier for language learning," in Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on. IEEE, 2013, pp. 1–6.
24. Twitter Apps. Available online: http://www.tweepy.org/ (accessed on 28 February 2018).
25. Roesslein, J. Tweepy Documentation. 2009. Available online: http://docs.tweepy.org/en/v3.5.0/ (accessed on 26 February 2018).
26. Available online: https://textblob.readthedocs.org/en/dev/ (accessed on 26 February 2018.

AUTHORS PROFILE

Nafees Akhter Farooqui received Bachelor of Science (Hons) from Aligarh Muslim University, Aligarh in 2005 and Master of Computer Application from Integral University, Lucknow in year 2010. He is currently pursuing Ph.D. and working as Research Associate in Department of Computer Applications, DIT University, Dehradun since 2017. He is a member of International Association of Engineers (IAENG) since 2017, ACM since 2011. He has published more than 18 research papers in reputed international journals including conferences proceeding and attended more than 10 Workshops and FDP's during the teaching and research. His area of interest are Machine Learning, Deep Learning, Data Mining and Artificial Intelligence. He has 9 years of teaching experience and guided many projects of UG and PG level students during teaching.

Dr. Ritika is working as an Associate Professor of Computer Science and Applications (Head, Department of Computer Applications) at DIT University. She received her Ph.D. degree in Computer Science from Gurukul Kangri University, Haridwar in the year 2010, M.Tech. degree in Computer Science and Engineering from Uttarakhand Technical University, Dehradun and MCA from Gurukul Kangri University Dehradun. She specializes in core areas of computer science and holds experience of more than 17 years. Her area of Interest are Machine Learning, Data Mining and Data ware housing, Mobile and Adhoc Networks. She has life time membership of ISCA, CSI, IEEE, IETE, ACM, IAENG, Active member of CSI Dehradun Chapter. She is SPOC of NPTEL Local Chapter DIT University Dehradun. Received certificate of appreciation for instrumental role as SPOC NPTEL DIT Chapter. She received "Active Participation - Woman-CSI Dehradun chapter" Award at CSI Annual Convention 2015, New Delhi.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

She has published more than 45 papers in reputed International Journals and conferences related to computer science. She has participated in more than 40 Workshops / Seminars / Faculty Development Program (FDP) organized at National and International level. She has supervised number of research scholars, and many others are presently under her guidance. She is involved in various other activities to give students full exposure to recent developments and various advance industrial trends. She is editorial board member of International Journal of Computers & Technology (IJCT). ISSN: 2277-3061, International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319 – 2682, International Organization of Science Research ISSN 22788219 e-ISSN 22503021.

Aayush Saini currently pursuing Bachelor of Computer Application from DIT University, Dehradun. His area of interest are Machine Learning, Deep Learning, Web technology and Artificial Intelligence**.** He is also worked on many commercial and academic projects during his study. At current he is working on DIT University Central Library Website.