Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018)
IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7

Sentiment Analysis to Predict Election Results Using Python

Ms. Farha Nausheen
Dept of IT
MJCET
Hyderabad, India
farha@mjcollege.ac.in

Ms. Sayyada Hajera Begum
Dept of IT
MJCET
Hyderabad, India
hajera@mjcollege.ac.in

***Abstract*— Sentiment analysis is an evaluation of the opinion of the speaker, writer or other subject with regard to some topic. In US presidential election 2016, Donald Trump, Hillary Clinton and Bernie Sanders were among the top election candidates. The opinion of the public for a candidate will impact the potential leader of the country. Twitter is used to acquire a large diverse data set representing the current public opinions of the candidates. The collected tweets are analyzed using lexicon based approach to determine the sentiments of public. In this paper, we determine the polarity and subjectivity measures for the collected tweets that help in understanding the user opinion for a particular candidate. Further, a comparison is made among the candidates over the type of sentiment. Also, a word cloud is plotted representing most frequently appearing words in the tweets.**

Keywords—Twitter; Sentiment analysis; Lexicon;

I. INTRODUCTION

With the advent of social media over the last decade, the efforts to determine people's attitudes with respect to a specific topic or event have garnered a wide research interest in natural language processing and introduced "sentiment analysis." Social Networking websites and micro blogging websites in today's world has become the biggest web destinations for people to communicate with each other, to express their thoughts about products or movies, share their daily experience and communicate their opinion about real-time and upcoming events, such as sports or political elections [17], etc.

To achieve a large, diverse dataset of current public opinion or sentiments, Twitter could be used as a valuable resource which allows users to send and read small messages called "Tweets". Basically Twitter is a micro blogging website that allows users to post brief and quick real-time updates regarding different activities and facilitates sharing, forwarding and replying messages quickly [18] which allows the quick spread of news or information. The wide use of hash tags also makes it easy to search for tweets dealing with a specific subject. With 17.1 million tweets, the first presidential debate between Donald Trump and Hillary Clinton was the most tweeted debate ever.

Using sentiment analysis [1] for predicting an election's result is empirically challenging to train a successful model to conduct sentiment analysis on tweet streams [2] for a dynamic event such as an election. Among the key challenges are changes in the topics of conversation and the people about whom social media posts express opinions. For doing this, we first created a supervised multiclass classifier (positive versus negative versus neutral) for analyzing opinions about different election candidates as expressed in the tweets. We then train our model for each candidate separately.

The motivation for this segregation comes from our observation that the same tweet on an issue can be positive for one candidate while negative for another. In fact, a tweet's sentiment is candidate- dependent. We used approx. 10,000 labeled tweets collected for three candidates (Bernie Sanders, Donald Trump, and Hillary Clinton).

The rest of the paper is organized as follows: Section II presents the relevant background work on sentiment analysis. Section III classifies different approaches in sentiment analysis Section IV explains how the proposed model is evolved. Section V describes how proposed approach is implemented to determine the sentiments associated with different candidates. Finally, section VI is based around conclusions.

II. RELATED WORK

Boia et al. [3] and Manuel et al. [4] proposed two approaches that, respectively, rely on emoticons to detect the polarity of tweets and on slang words to assign a sentiment score to online texts. Akcora et al. [5] proposed a method to determine the changes in public opinion over the time, and identify the news that led to breakpoints in public opinion. Gao and Sebastiani [6] presented an approach that focus in the repartition or the frequency of sentiment classes in the set they analyze. P Bhoir et al in [7] implemented method to find the subjectivity of sentences and used rule based system to determine feature-opinion pair and using another technique the orientation of extracted opinion is revealed. In [8], lexicon based classification algorithm is used to analyze and predict a

Muffakham Jah College of Engg. & Technology, Hyderabad - India

user's sentiment polarity. It used degrees of comparison namely, positive, comparative and superlative on words. R. Rezapour et al., in [14] have proposed and evaluated an enhanced model that incorporates informative hash-tags into a lexicon to improve the accuracy of sentiment analysis. Jyothi Ramteke et al., in [15] performed data set creation by first collecting data using twitter streaming API then preprocessing is done to remove special characters and data labeling is done firstly manually using hash-tag labeling and then using VADER tool which is lexicon and rule based sentiment analysis tool. In [16], Bouazizi et al., proposed pattern based approach for sentiment quantification in twitter. They defined two metrics to measure the correctness of sentiment detection and proved that sentiment quantification can be more meaningful task than the regular multi-class classification

III. Approaches In Sentiment Analysis

Sentiment Analysis can be categorized into two approaches: Machine Learning based approach and Lexicon based approach. Machine Learning based approach classifies the text using classification algorithm, whereas Lexicon based approach [12] uses sentiment dictionary with opinion words and match them with data to determine polarity. Sentiment values are assigned to words that describe the positive, negative and neutral attitude of the speaker. There are mainly three schemes in lexicon based approach [13]: Manual Scheme, Dictionary based scheme, Corpus based scheme. Manual scheme is time consuming and has to be combined with other automated approaches.

Dictionary based Scheme is a simple technique which uses small set of seed words and an online dictionary. The strategy here is initial seed set of words with their known orientations are collected and then online dictionaries like WordNet etc are searched to find their probable synonyms and antonyms.

Corpus based scheme basically uses corpus data to identify sentiment words, though it is not as effective as dictionary based scheme. It is helpful in finding the domain and context of specific sentiment words against the corpus data. This characteristic is advantageous in exploring data to determine sentiment words.

IV. Proposed Model

A. Proposed Approach

The proposed system retrieves tweets and performs political sentiment analysis. It consists of four main modules

i. Retrieval module
ii. Preprocessing module
iii. Analysis module
iv. Result visualization module

i. Retrieval module

Twitter allows its users to retrieve tweets by using twitter API Twython. User tweets are collected for Trump, Hilary, and Bernie by executing python code. These tweets are then filtered and classified based on the sentiments.

ii. Preprocessing module

The collected tweets may contain hashtags, URLs, @tags, emoticons, trailing whitespaces and newline characters. The preprocessing module performs the filtration process using python libraries for retrieving the meaningful parts of the tweet by removing the unnecessary content.

iii. Analysis module

The original sentiments are used to track two trends over time:

- Polarity: Is the sentiment associated with the entity positive, negative or neutral?
- Subjectivity: How much sentiment (of any polarity) does the entity gather?

Subjectivity indicates proportion of sentiment to frequency of occurrence, while polarity indicates percentage of positive sentiment references among total sentiment references. The Average polarity and subjectivity measures are calculated as in the proposed algorithm from Fig. 1

i. Initialize positive, negative and neutral tweet count to zero.
positive_tweet_count=0; negative_tweet_count=0;
neutral_tweet_count=0;
ii. Also initialize Avg_polarity=0, Avg_subjectivity=0;
iii. Fetch the stored tweets from the file
a. For each line in the file do
b. Sent1 = TextBlob(line);
c. Calculate polarity and subjectivity as
Polarity = Sent1.sentiment.polarity;
Subjectivity=Sent1.sentiment.subjectivity;
iv. If Polarity > 0, increment positive_tweet_count;
v. If Polarity < 0, increment negative_tweet_count;
vi. If Polarity = 0, increment neutral_tweet_count;
vii. Calculate Average_polarity and Average_subjectivity as
a. Avg_polarity=Sum (Polarity)/len(Polarity);
b. Avg_subjectivity=Sum (subjectivity)/len(subjectivity);
viii. Calculate percentage of positive, negative and neutral tweets.

Fig. 1: Proposed Algorithm

iv. Result Visualization Module:

The final step of this process is to take in the filtered tweets and generate word cloud, group bar graph and table to visualize the results.

Block diagram of proposed architecture for the sentiment classification is shown in Fig 2. This classifier is used to determine the polarity and subjectivity scores.

B. Tools Used

i. Twython

Twython is the prime Python library which provides an easy means for accessing Twitter data. Actively maintained and featuring support for Python 2.6+ and Python 3. It can be used to query data [9] for the purpose of User information, Twitter lists, Timelines and things found in Twitter API Docs. It offers support for User authenticated calls and application authenticated calls.

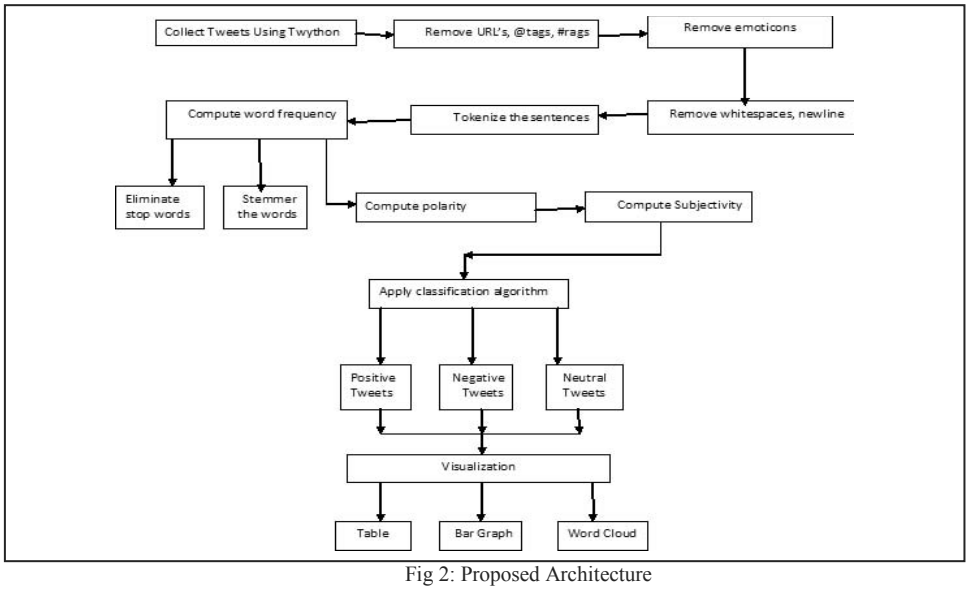

Fig 2: Proposed Architecture

ii. Natural Language Toolkit

NLTK [10] is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries[4], and an active discussion forum.

iii. TextBlob

TextBlob[11] is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

V. Experimental Results

Twython is a library of python providing an easy way to access twitter data in python. With the help of twython we collect tweets during election and these tweets are filtered using NTLK (Natural Language Toolkit) where unnecessary common words are removed. These filtered tweets are stored in text files which are then imported for sentiment analysis, to determine the opinion of the subject with respect to the overall contextual polarity.

The proposed algorithm is implemented on the filtered tweets to determine the counts of positive, negative and neutral tweets for Trump, Hilary and Bernie. The sentiment of the tweet is classified depending on the polarity of the individual words present in the tweet. A score of +1 is given to a positive word (like good, great etc) and -1 is given to a negative word (like bad, worse etc) and 0 for neutral words (like quite, average etc). The polarity ranges between [-1,+1]. The total polarity of a tweet is then calculated by adding the scores of all the individual words. These counts are then used to determine the percentage of positive, negative and neutral tweets. Subjectivity reflects users own views about the candidate and ranges between [0, 1] where 0 specifies objective content and 1 specifies subjective content. The calculated values are depicted in the tabular form as in Table I. From the table it can be inferred that Hillary has got more number of positive tweets thus giving a good average polarity score over Bernie and Trump. It can also be observed that average subjectivity score of Bernie is better compared to Hillary and Trump. To generate graph depicting sentiment analysis among Trump, Hilary and Bernie, plotly is used which is a visualization tool.

TABLE I. Evaluation of Polarity and Subjectivity Measures

Name	**Donald Trump**	**Hillary Clinton**	**Bernie Sanders**
No. of positive comments	17812	23716	20748
No. of Neutral comments	7853	12475	5741
No. of Negative comments	24335	13809	23511
% Positive	35.624	47.432	41.496
% Negative	15.706	24.95	11.48
% Negative	48.67	27.618	47.022
Average Polarity	-0.0073	0.0482	0.0282
Average Subjectivity	0.278	0.2809	0.3138

The graph in Fig. 3 depicts that Hilary got the highest number of positive and neutral tweets whereas Trump got the highest number of negative and lowest number of positive tweets.

Word Cloud generated using Plotly gives a graphical representation of the most frequently appearing words in the tweets. The frequent word will have prominent appearance in the Word Cloud. Fig. 4 represents the word cloud which is plotted for Hillary by inputting the processed tweets file of Hillary. Similarly, word clouds for trump and Bernie can also be obtained.

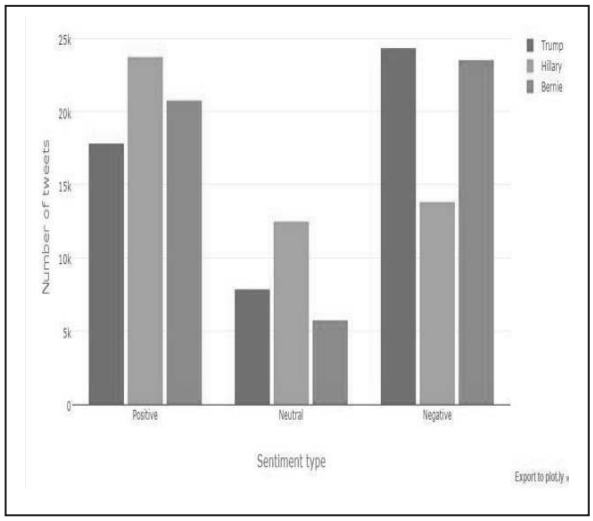

Fig 3: Comparison of sentiment type for different candidates.

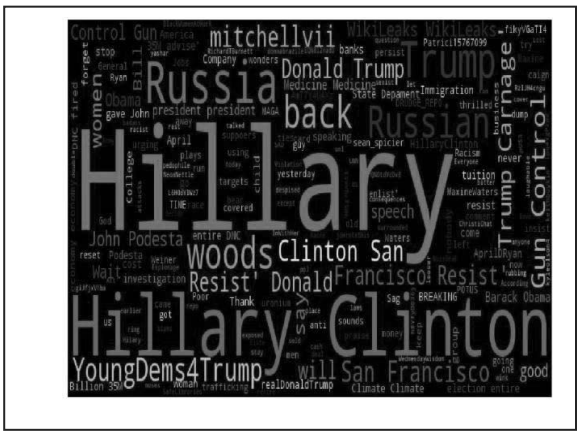

Fig 4: Word Cloud for tweets of Hillary

VI. Conclusion

In this paper, we propose lexicon based sentiment analyzer which classifies the tweets based on its sentiment value. The tweets considered are from US presidential elections 2016. The sentiment classification is done based on polarity and subjectivity measures. These measures signify the positive, negative or neutral attitude of users towards a particular election candidate, thereby enabling us to present the comparison between the top candidates for presidential elections 2016. In the future, a multi lingual based sentiment classification can be built to acquire tweets in different languages.

ACKNOWLEDGMENT

We would like to thank MJCET for the grants and literary material.

REFERENCES

[1] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." ICWSM 7.21 (2007): 219-222.

[2] Mondher Bouazizi, Tomoaki Ohtsuki, "A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter", *Access IEEE*, vol. 5, pp. 20617-20639, 2017, ISSN 2169-3536.

[3] Boia, Marina, et al. "A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets." Social computing (socialcom), 2013 international conference on. IEEE, 2013.

[4] Manuel, K., Kishore Varma Indukuri, and P. Radha Krishna. "Analyzing internet slang for sentiment mining." 2010 Second Vaagdevi International Conference on Information Technology for Real World Problems. 2010.

[5] Akcora, Cuneyt Gurcan, et al. "Identifying breakpoints in public opinion." Proceedings of the first workshop on social media analytics. ACM, 2010.

[6] Gao, Wei, and Fabrizio Sebastiani. "Tweet sentiment: From classification to quantification." Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on. IEEE, 2015.

[7] Bhoir, Purtata, and Shilpa Kolte. "Sentiment analysis of movie reviews using lexicon approach." Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on. IEEE, 2015.

[8] Mandal, Santanu, and Sumit Gupta. "A Lexicon-based text classification model to analyse and predict sentiments from online reviews." Computer, Electrical & Communication Engineering (ICCECE), 2016 International Conference on. IEEE, 2016.

[9] https://twython.readthedocs.io/en/latest/

[10] http://www.nltk.org/

[11] https://pypi.python.org/pypi/textblob

[12] Aung, Khin Zezawar, and Nyein Nyein Myo. "Sentiment analysis of students' comment using lexicon based approach." Computer and Information Science (ICIS), 2017 IEEE/ACIS 16th International Conference on. IEEE, 2017.

[13] Hailong, Zhang, Gan Wenyan, and Jiang Bo. "Machine learning and lexicon based methods for sentiment classification: A survey." Web Information System and Application Conference (WISA), 2014 11th. IEEE, 2014.

[14] Rezapour, Rezvaneh, et al. "Identifying the Overlap between Election Result and Candidates' Ranking Based on Hashtag-Enhanced, Lexicon-Based Sentiment Analysis." *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on*. IEEE, 2017.

[15] Ramteke, Jyoti, et al. "Election result prediction using Twitter sentiment analysis." *Inventive Computation Technologies (ICICT), International Conference on*. Vol. 1. IEEE, 2016.

[16] Bouazizi, Mondher, and Tomoaki Ohtsuki. "Sentiment analysis in twitter: From classification to quantification of sentiments within tweets." *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016.

[17] J. M. Soler, F. Cuartero, and M. Roblizo, ``Twitter as a tool for predicting elections results," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2012, pp. 1194_1200.

[18] A. Java, X. Song, T. Finin, and B. Tseng, ``Why we Twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD 1st SNA-KDD Workshop Web Mining Social Netw. Anal.*, Aug. 2007, pp. 56_65.