Journal of Global Research in Computer Science 2229-371X

Election Prediction Using Twitter Sentiment Analysis

Rashiduzzaman Prodhani*, Atowar Ul Islam and Luit Das

Department of Computer Science and Electronics, University of Science and Technology Meghalaya, Ri-Bhoi, Meghalaya, India

Research Article

Received: 04-Apr-2022, Manuscript No. GRCS-22-59465; **Editor assigned:** 06-Apr-2022, PreQC No. GRCS-22-59465 (PQ); **Reviewed:** 20-Apr-2022, QC No. GRCS-22-59465; **Revised:** 02-Jun-2022, Manuscript No. GRCS-22-59465 (R); **Published:** 14-Jun-2022, DOI: 10.4172/2229-371X.13.3.006

*For Correspondence : Rashiduzzaman Prodhani, Department of Computer Science and Electronics, University of Science and Technology Meghalaya, Ri-Bhoi, Meghalaya, India, **Tel:** 9678611877;

Email: prodhani2008@gmail.com

Keywords: Social media; Twitter; Politics; Sentiment analysis; Naive bayes; Support vector machine; Positive; Negative; Neutral

ABSTRACT

Sentiment analysis is the computational study of opinions, sentiments, evaluations, attitudes, views and emotions expressed in text. It refers to a classification problem where the main focus is to predict the polarity of words and then classify them into positive or negative sentiment. Sentiment analysis over Twitter offers people a fast and effective way to measure the public’s feelings towards their party and politicians. The primary issues in previous sentiment analysis techniques are classification accuracy, as they incorrectly classify most of the tweets with the biasing towards the training data. So I collected live data to predict the accurate election result. Twitter is a place where users posting quick and real-time updates about different activities or events as the spread of information and news are quick enough. We used the python library “Tweepy” for accessing the Twitter API and fetched live data from Twitter. More than 2000 tweets for each political party candidate are fetched by using keywords. Using “TextBlob” library of python, sentiments are applied to each tweet and depending upon more positive tweets for particular candidate and we can visualize a prediction. Text classification algorithms like Naive Bayes, Support Vector Machine (SVM) and Random Forest are used to train model using labelled data. The accuracy of the predicted result is calculated and the result is declared finally, result is represented in the form of pie chart, bar graph for each political candidate representing positive, negative and neutral sentiments.

INTRODUCTION

Sentiment Analysis (SA) is the process of which finds whether a word or sentence or a document is positive, negative or neutral. SA is also known as opinion mining. This innovation is used commonly to discover how different individuals feel about a certain topic. The applications for sentiment analysis are endless. It can be applied to customer reviews, survey responses, competitors, *etc.* Its benefit is popular in business analytics and mostly in situations where text needs to be analysed. Sentiment Analysis aims to discover opinions, identify sentiments and afterwards classify those sentiments into various categories. A well-defined and an accurate system for predicting sentiments could enable us, to extract sentiments from the internet and forecast social behaviour, political drifts, and evolving parties from a particular geographical location [1]. However when we do sentiment analysis various challenges are faced because people don't always express their feelings in a same way, from a particular instance the sentence might seem positive and might seem negative from a different point of view. More over there is possibility of wrong spellings, intensifiers and spams confuses us when we do the analysis and there are millions of ways to join sentences and to treat negations which is even more stimulating. As the growth of existing subjective text on the internet is increasing rapidly, in order to achieve more refined, realistic and subjective opinion on companies and products, people uses internet. Lexicon approach is one of the analysis techniques which give each word a sentiment score. There are three sentiment scores - positive, negative and neutral.

LITERATURE REVIEW

Election forecasting models in political science

In the past decades, political scientists have proposed a series of election forecasting models. Lewis-Beck and Rice (1982) built the first presidential election forecasting model which is rooted in the political science theory. The model treats the job approval rating for the president in the July Gallup poll and the Gross National Product (GNP) growth rate in the first two quarters of the election year as two predictive factors. Based on this model, developed the trial-heat model. The model also consists of two predictive variables, namely the incumbent party's candidate support on Gallup poll in early September and the second-quarter growth rate in the real GDP of the election year. In 1992, the group developed another model named the convention-bump model which considers three predictors including the incumbent party's candidate support of the pre-convention polls, the net change of the incumbent party's candidate support after both conventions are completed, and the second-quarter GDP growth rate in the economy [2].

Election prediction based on Twitter sentiments

With the advent and increasing popularity of big data in the current century, researchers started to incorporate Twitter data as the assistance in election predictions [3]. A stream of studies has suggested that Twitter data are powerful for the prediction of political election outcomes by using sentiments extracted from tweets in the U.S. and other countries [4]. Tweets are typically collected using keywords related to certain elections through the Twitter Application Programming Interface (API). Sentiment analysis is conducted on these data to extract sentiments towards a candidate or party. "Sentiment analysis, also called opinion mining, aims to analyse people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions from written language towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes" [5]. The objective of sentiment analysis is to identify or categorize the attitude expressed in a piece of text. Specifically, the attitude may be positive (favourable), negative (unfavourable), or neutral towards a subject [6]. There are mainly two types of approaches for sentiment analysis on election-related Twitter data. One is the lexicon-based approach and the other is the machine learning approach. The lexicon-based approach for sentiment analysis relies on a pre-defined sentiment lexicon and compares the presence or frequency of words in the given text with the words in the lexicon. For example, Ahmed, Jaidka, and Skoric predicted elections in four countries and compare the quality of predictions and the role of different technological infrastructures and democracies setups in these countries. For sentiment analysis, they applied a sentiment lexicon called SentiStrength to assign a positive score and a negative score to all the tweets relevant to a party. Because of the different internet connectivity and political environment, the prediction accuracy is different in the four countries. The machine learning approach for sentiment analysis is generally divided into two sub-categories, supervised learning methods and unsupervised learning methods. Past work usually employed supervised learning methods that need good pre-labelled training datasets. For instance, Wang et al. developed a system for real-time analysis of tweet sentiment towards presidential candidates in the 2012 U.S. election. They trained a Naïve Bayes model on unigram features to classify the sentiment. Paul et al. collected geotagged tweets from a period of 6 months leading up to the U.S. presidential election in 2016 and classified the tweets towards either democratic or republic based on their sentiment at the county level. They trained the Support Vector Machine (SVM), Multinomial Naïve Bayes (MNB), Recurrent Neural Network (RNN), and fast text models by 1.6 million tweets from the Stanford Twitter Sentiment (STS) corpus.

Twitter sentiments and poll data comparison

Past studies have shown the feasibility of substituting poll with the Twitter sentiment. O'Connor et al. Found that the Twitter sentiment had correlations with public opinion by analyzing several surveys on consumer confidence and political opinion from 2008 to 2009. Beauchamp modeled state-level polls during the 2012 presidential election as a function of political tweets and found that Twitter-based measures can predict opinion polls. Anuta, Churchin, and Luo found that both the polls and Twitter were biased in the 2016 U.S. election. The poll had a small bias towards Hillary Clinton while

Twitter had a slightly larger bias towards Donald Trump. Bovet, Morone, and Makse showed that the Twitter opinion trends in the 2016 U.S. presidential election followed the aggregated New York Times polls with remarkable accuracy by comparing them based on their proposed method.

MATERIALS AND METHODS

Extracting tweets from Twitter API Tweepy

Twitter is a popular social network where users share messages called tweets. Twitter allows us to mine the data of any user using Twitter API or Tweepy. The data will be tweets extracted from the user. The first thing to do is get the consumer key, consumer secret, access key and access secret from twitter developer available easily for each user. These keys will help the API for authentication.

Steps to obtain keys:

- Login to twitter developer section
- Create an App
- Create your Twitter Application
- Details of the new app will be shown along with consumer key and consumer secret.
- For access token, click Create my access token". The page will refresh and generate access token.

Tweepy is one of the libraries that should be installed using pip. Now in order to authorize our app to access Twitter on our behalf, we need to use the authorization interface. Tweepy provides the convenient Cursor interface to iterate through different types of objects. Twitter allows a maximum of 3200 tweets for extraction.

Sentimental analysis

Sentiment analysis is the process of which finds whether a word or sentence or a document is positive, negative or neutral. SA is also known as opinion mining. This innovation is exploited to decide how individuals feel about a topic like election, film or a book and so forth. The applications for sentiment analysis are endless. It can be used to track customer reviews, survey responses, competitors, *etc.* The need for sentiment analysis is increasing rapidly because SA is efficient. To extract sentiments and opinion, sentiment analysis evaluates thousands of text documents within few seconds when equated to groups of people who take hours of time to manually analyse each word or sentence. Text and sentiment analysis is embraced by many businesses and tries to use it into their processes because of its efficiency. Sentiment analysis of social media can be an impressive foundation of information and can deliver insights that can:

- Define "marketing strategy"
- Enhance "campaign success"
- Enhance "product messaging"
- Enhance "customer service"

In a nutshell, sentiment analysis of social media can improve prediction result if done properly. However an inaccurate sentiment analysis data while deciding can give unsatisfactory result.

RESULTS AND DISCUSSIONS

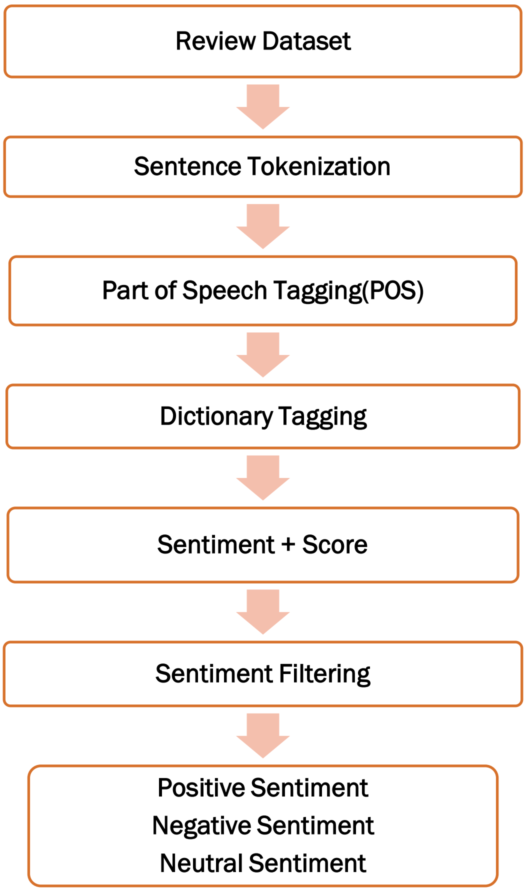
Lexical-based Approach

Lexicon approach is a method in which each word in a sentence is given a sentiment polarity. This approach works on a bag of words. It has three sentiment scores which unfolds how positive, negative and neutral the words present in the sentence are [7]. The value of positive, negative and neutral score varies between '-1 to 1', sum of the scores is 1 for each sentence.

After the tweets are pulled out and cleaned, sentence tokenization is done. In tokenization every sentence is separated into various components each representing an instantaneous token and afterward this token is sent to a part of speech tagger. It is also recognized as lexical classifications or word classes. The process in which we categorize words into their parts of speech and to tag them consequently is known as "part-of-speech tagging", "POS-tagging," or simply "tagging". In Part of Speech (POS) tagging the tagger associates "the significance and role of each word in syntactic setting". The tagger is useful once we obtain sentiments because it tells the classifiers the doable elements of content where the sentiments might exist. Natural Language Tool-Kit (NLTK) is employed to train various a Part Of Speech (POS) taggers.

POS tagging is followed by the dictionary tagging in which it matches words with dictionary objects and if any matching conditions become true then the word is given a sentiment score (which is "positive score – negative score").Dictionary tagging is a crucial step as it contains words carrying sentiments which will be matched with the dataset. The dictionary has "the word", "an opinion score" and "a synset score". This process is repeated for every sentence in the whole document. Polarity score is number arranging between '-1' to '1' .If a word shows strong sentiment and is very intense then as a result a number nearby to 1 will be allotted to that word and if a word is less intense having weak sentiment then accordingly a number near to -1 is allotted to that. Sentiment score "0" signifies that word is neutral that it doesn't have any sentiment. After every sentence a sentiment score is assigned which is used to sieve these sentences into four unique classes as Positive, Negative, Neutral and Spam. If the sentences have either values greater than '-1' or '1', then these sentences are considered spam and the main purpose to unfold these spam is to intensely support or intensely oppose any issue and their motive is to unfold rumor at intervals the social media. We computed the sentiment for every tweet, initially the tweets were tokenized and every tokenized word is matched with dictionary words and then an appropriate sentiment score is given to the tokens (from dictionary). We summed the score for every tweet to search for (abundant what proportion what quantity) positive and negative tweets are there for every candidates collaborating within the election. To exemplify sentiments in an effective way we calculated percentage polarity.

Figure 1. Lexical orientation.



$$\text{Positive Sentiment\%} = \frac{\text{No of Positive Tweets}}{\text{Total No of Tweets – No of Spam Tweets}} * 100$$

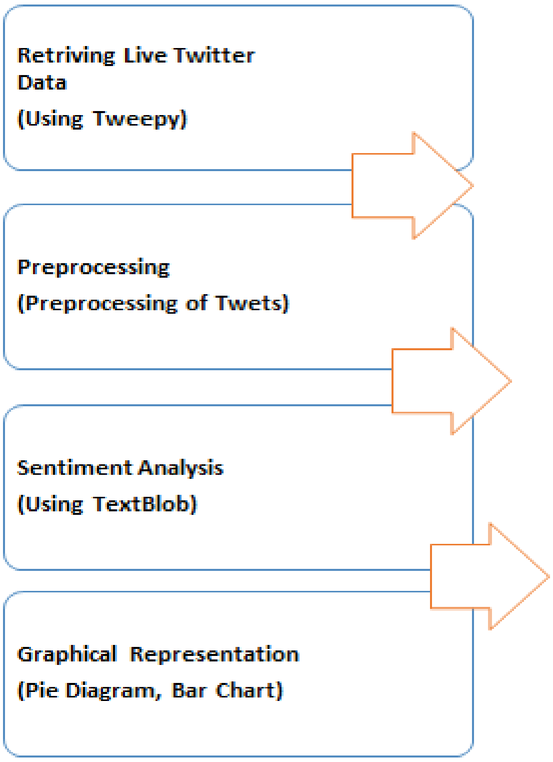
$$\text{Negative Sentiment\%} = \frac{\text{No of Negative Tweets}}{\text{Total No of Tweets – No of Spam Tweets}} * 100$$

$$\text{Neutral Sentiment\%} = \frac{\text{No of Neutral Tweets}}{\text{Total No of Tweets – No of Spam Tweets}} * 100$$

Methodology and classification algorithm

The data is retrieved from different social sites so that opinion of maximum people can be considered and the result would be more accurate. API is basically used for collecting tweets provided by the twitter through streaming API. Twitter data is unique from data shared by most other social platforms, because it reflects information that users choose to share. If someone wants to access APIs, they are required to register in an application. By default, applications can only access public information on Twitter (Figure 2).

Figure 2. Methodology block diagram.



Pre-processing

Pre-processing is to analyse data and do some cleaning on the text which isn't returning any meanings and apply our algorithm for classifying text into either positive sentiments or negative sentiments. Also it contains Hashtags, URLs, Punctuation, common stop words that has no meaning. It is an integral step as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn. In this stage, special characters like '@' and URLs are stripped off to overcome noise. One of the most important goals of pre-processing is to enhance the quality of the data by removing noise. It is a technique which is used to transform the raw data in a useful and efficient format. Some important steps taken towards data pre-processing are as explained below.

Removing Hashtags and urls: As URLs contains no information it is better to remove it and clean the data. Also Hashtags can be really important for us. As almost everyone is spending more time to choose correct hashtags while writing tweets, it may provide really important words to our word pool. So, no need to lose them, because of this we delete the # character and keep the rest.

Lower case conversion: There are many ways in which people can write the same thing, character data can be difficult to process. It has to be done properly or else can lead to huge errors which will ultimately affect the model while training. For example, "Election" and "election" can be considered as two separate words. Hence, for accurate string matching we are converting our complete text into lower case, so that it will become easy to segregate data having the same meaning.

Removing punctuation and numbers: All punctuation, numbers are also needed to be removed from reviews to make data clean and neat. Unnecessary commas, question marks, other special symbols should also be removed. Here, we are not removing dot (.) symbol from our reviews because it is splitting our text into sentences.

Removing stop words: Word like "the", "a", "on", "is", "all" can be removed by comparing text to a list of stop words. "Stop words" are the most common words in a language. These words do not carry important meaning and are usually removed from texts.

Sentiment analysis using text blob

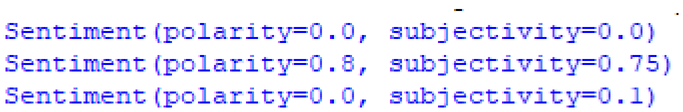
Text blob is a python library and offers a simple API to access its methods and perform basic NLP tasks. Here, we have used this library to perform text classification in either positive or negative on the basis of sentiment analysis. This library is just like a Python string with the functionality of that we can easily use its functions. It provides a really important functionality that can easily summarize the text, provide you with sentiments of the text, spelling correction, translation, and language detection and so more.

• Polarity ranges from -1 to +1 (negative to positive) and tells whether the text has negative sentiments or positive sentiments. Polarity tells about factual information.

• Subjectivity also ranges from -1 to+1 (negative to positive). So more +ve subjectivity means less factual data and mostly public opinion.

There are many cases where polarity is zero because there is some data which either doesn't contain any text or simply have links or hash tags only (Figure 3).

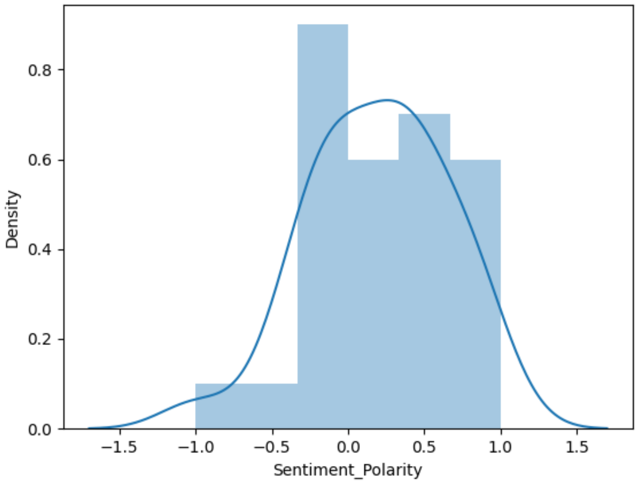
Figure 3. Sentiment polarity.



```
Sentiment(polarity=0.0, subjectivity=0.0)
Sentiment(polarity=0.8, subjectivity=0.75)
Sentiment(polarity=0.0, subjectivity=0.1)
```

In the above three outputs in the 2nd statement, we can see that subjectivity is 0.75 which is indicating text contained at nth row is barely a personal opinion (Figure 4).

Figure 4. Sentiment polarity graph.



From above sentiment polarity graph we can easily interpret that polarity ranges from -1 to +1 and a larger no. of people have positive reviews because it is mostly concentrated between 0 and 0.5.

Data visualization or graphical representation

Visualizing data gives us a clearer picture of what are we actually doing. So it's the most important step to include in our projects for making it understand better in the simplest and easy way possible in our presentations. And also, it frames a clear picture in front of us that which attribute is contributing better to our output. It is an important step before applying any analysis and modeling [8].

CONCLUSION

Twitter is a micro blogging facility that has more than 500 million messages on a daily basis. Scholars have been utilizing Twitter to monitor people reactions in political activities, such as debates and campaigns. By doing so, some of them claim that a forecast or prediction to an election can be made. I have built a model to sentiment analysis of political tweets collected from Twitter. I implemented the model on The 2021 Legislative Assembly election of one state and collected data based on the two popular candidates competing for each other. Although there were many limitations of the data, we believe that this data, if analyzed thoroughly, can potentially tell us a lot, and importantly can empirically test some of the assumptions made when building the model.

The proliferation of social media in the recent past has provided end users a powerful platform to voice their opinions. Businesses (or similar entities) need to identify the polarity of these opinions in order to understand user orientation and thereby make smarter decisions. One such application is in the field of politics, where political entities need to understand public opinion and thus determine their campaigning strategy. Sentiment analysis on social media data has been seen by many as an effective tool to monitor user preferences and inclination. Popular text classification algorithms like Naive Bayes and SVM are Supervised Learning Algorithms which require a training data set to perform Sentiment analysis. The accuracy of these algorithms is contingent upon the quantity as well as the quality of the labelled training data.

REFERENCES

1. Indonesian Presidential election on 2009. The Kompas Newspaper, Apr 26 2017.
2. Sharma Y, et al. Sentiment analysis and opinion mining. Int J Soft Comput Artif Intell. 2015;3:59–62. [Crossref] [GoogleScholar] [Indexed]
3. Heredia B, et al. (2017) Exploring the effectiveness of Twitter at polling the United States 2016 presidential election. In: IEEE 3rd International Conference on Collaboration and Internet Computing (CIC). 283–290. [Crossref] [GoogleScholar] [Indexed]
4. Joyce B, et al. (2017) Sentiment analysis of tweets for the 2016 US presidential election. In: Undergraduate Research Technology Conference (URTC), IEEE MIT. 1–4. [Crossref] [GoogleScholar] [Indexed]
5. Kao A, et al. (2007) Natural language processing and text mining. Berlin: Springer.
6. Vikash Nandi, et al. “Political sentiment analysis using hybrid approach”. Int Res J Eng Technol. 2016;3:1-8. [GoogleScholar] [Indexed]
7. Nausheen F, et al. (2018) “sentiment analysis to predict election results using python”. Proceedings of the Second International Conference on Inventive Systems and Control. 1259-1262. [Crossref] [GoogleScholar] [Indexed]
8. Andranik T, et al. (2010) “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”. Proceedings of the 4rth International AAAI Conference on Weblogs and Social Media. 178-185. [GoogleScholar] [Indexed]