IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 4.819

Volume 4, Issue 3, April 2021

Election Result Prediction Using Sentiment Analysis

Ankit Maurya[1], Satish Lodh[2], Mayur Joshi[3], Prof. Vinaykumar Singh[4]

Students, Department of Electronics[1,2,3]
Assistant Professor, Department of Electronics[4]
Shree L. R. Tiwari College of Engineering, Mumbai, India

Abstract: *Social media today has become a very popular communication tool for users. Millions of users share their opinions on different aspects on daily basis. Sentiment Analysis determines the polarity and inclination towards any specific topic, idea or entity. Applications of such analysis can be seen during elections, movie promotions, and many other fields. In our project, we aim to predict the winning probability of any political party by using both labelled as well as unlabeled data. Labelled data can be collected by using polling method but the result may not provide better accuracy. Hence, it is necessary to fetch live data to predict the accurate election result. Twitter is a microblogging site which allows the users in posting quick and real-time updates about different activities or events as the spread of information and news is quick enough. With the help of hashtags, the needed data can be easily generated and put to use. We exploited the python library "Tweepy" for accessing the Twitter API and fetched live data from Twitter. 350 tweets for each political party are fetched by using keywords. Using "TextBlob" library of python, sentiments are applied to each tweet and depending upon more positive tweets for particular party, winning party is declared. Also, popular text classification algorithms like Na¨ive Bayes, 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 bargraph for labelled data according to the number of voted for each political party and for unlabeled data using pie chart for each political party representing positive, negative and neutral sentiments.*

Keywords: Social Media, Twitter, Politics, Sentiment Analysis, Naive Bayes, Support Vector Machine.

I. Introduction

An election is a most important part in the democracy. It is the instrument of democracy where the voter communicate with the representatives. One vital component in an election is the election polls/survey. opinion poll has existed since the early 19th century, based on this data source prediction of the outcome of an election is done. Traditional polls are too costly and still accuracy won't be achieved. Social media has become the most popular communication tool on the Internet. Social media sites have become valuable sources for opinion mining because people post everything right from the details of their life to the products and services they use.

Twitter is an online social networking service that enables users to send and read short 240-character messages called "tweets". The use of hash tags makes the problem of text classification relatively easier since the hash tag itself can convey an emotion or opinion. Currently around 6500 tweets are published per second, which results in approximately 561.6 million tweets per day. It allows users to post briefand quick real-time updates regarding different activities and facilitates sharing, forwarding and replying messages quickly which allows the quick spread of news or information.

This is an interesting research area that combines politics and social media which both concern today's society. The introduction proposes a new framework to predict the election result and sentiment analysis from Twitter data that focuses on Election.

DOI: 10.48175/IJARSCT-V4-I3-018

II. Proposed Method

2.1 Method 1

It is based on the Labelled data. Labelled data is a group of sample data that have been tagged with one or more labels. Labelling typically takes raw data and adds a label to each piece of that unlabelled data with meaningful tags that are useful tags related to the concept. Labelled data is used by Supervised learning which adds meaningful tags or labels or class to the rows. The data can come from social media, polling i.e. asking people about their opinion on elections and tags can be added based on requirement of the data from a raw data which can be done by asking the specialists about the data. Classification can be done using machine learning algorithms such as SVM, Naive Bayes and Random forest and Regression is applied to labelled datasets for Supervised learning. Different modules that are applied on the labelled data are:

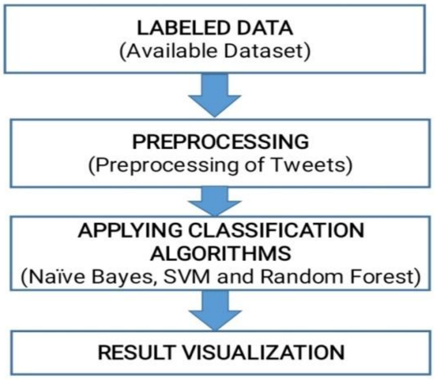
LABELED DATA
(Available Dataset)
PREPROCESSING
(Preprocessing of Tweets)
APPLYING CLASSIFICATION ALGORITHMS
(Naïve Bayes, SVM and Random Forest)
RESULT VISUALIZATION

Figure 1: Block diagram first method

1. Data Preprocessing
2. Applying Classification Algorithms
3. Result Visualization

A. Data Preprocessing

The size of the dataset is 4733 X 109. It may contain some of the missing values. Before applying any algorithms, we have to take care of missing values also the categorical data must be converted into numeric data. Label Encoder encode labels with a value between 0 and n classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier. Then algorithms are applied by splitting data into training and testing dataset.

B. Applying Classification Algorithms

Machine learning is a subset of Artificial intelligence which provides machine the ability to learn automatically and improve from experience without being explicitly programmed. This is why the data set is divided into the ratio of 80:20 i.e. 80% of the data is used for training the model and 20.

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).

Na¨ıve Bayes is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble.

C. Result Visualization

This is the final step in which votes for each political party is represented in the form of Bar Graph. As Data Visualization graphical format makes it easier to understand.

2.2 Method 2

In this method, 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.

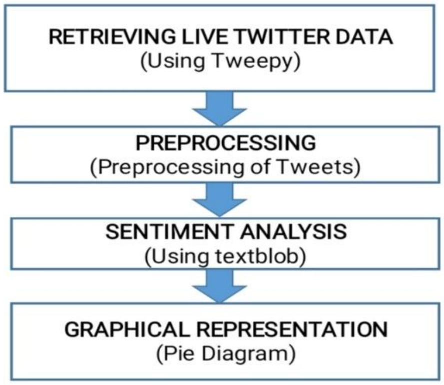

Figure 2: Block diagram second method

1. PreProcessing
2. Sentiment Analysis
3. Graphical Representation

A. PreProcessing

As we are using various Social Media as a source of information, nobody checks the spelling rules before publishing tweets or writing comments or people use a lot of slang words and repetitive letters in the sentence. 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 preprocessing 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 preprocessing 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 Punctuations and Numbers:

All punctuations, 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 stopwords. “Stop words” are the most common words in a language. These words do not carry important meaning and are usually removed from texts.

B. Sentiment Analysis

Sentiment is the prediction of emotions in a word, sentence or corpus of documents. It is a categorizing conversations into positive, negative or neutral labels. The original sentiments are used to track two things at a time that is the count and the average polarity for each party. The analysis of the tweets are done based on the polarity of the tweets and the conditions applied are:

- Polarity¿0=Increment positive tweet count;
- Polarity¡0=Increment negative tweet count;
- Polarity=0 is Increment neutral tweet count

All these conditions are used to categorize the data using python library textblob.

C. Graphical Representation

Result visualization is the act of taking information (data) and placing it into a visual context, such as a map or graph. In this module, result is shown in the form of chart, bar graph or pie diagram. It shows the political parties and their positive, negative and neutral percentages.

III. Conclusion

Na¨ıve Bayes, SVM and Random Forest are supervised learning algorithms used to train model by providing labelled data in the form of words. Here we have processed our data using various libraries of sklearn and applied these algorithms to predict election result. We have also created confusion matrix to know the accuracy of the result obtained. the Unlabelled Data we have searched tweets related to each party is fetched and sentiment for each sentence is calculated using textblob library.

Acknowledgment

We would like to thank our project guide Assistant Prof. Abhijit Somnathe and Assistant Prof. VinayKumar Singh who initiated us into the Election Result Prediction Using Sentiment Analysis. He has been a source of inspiration and her insight and vision has made it possible for us to pursue and understand the developments in these areas. Her patience, encouragement, critique and availability made this dissertation possible.

References

[1]. Ms. Farha Nausheen et al., “Sentiment Analysis to Predict Election Results Using Python”, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018).

[2]. Jyoti Ramteke et al., “Election Result Prediction Using Twitter Sentiment Analysis”, 2016 International Conference on Inventive Computation Technologies (ICICT).

[3]. Alexander Pak, et al., “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, Proceedings of the Universite de Paris-Sud, Laboratoire LIMSI-CNRS.

[4]. Andranik Tumasjan et al., “Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[5]. Parul Sharma et al., "Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter", 2016 IEEE International Conference on Big Data (Big Data).

[6]. Pritee Salunkhe et al., "Twitter Based Election Prediction and Analysis", International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 10 — Oct -2017.

[7]. Amandeep Kaur et al., "Sentiment Analysis On Twitter Using Apache Spark", Carleton University Project Report.

[8]. Widodo Budiharto et al., "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis", J Big Data (2018) 5:51.

DOI: 10.48175/IJARSCT-V4-I3-018