Sentiment Analysis based on textual data

Author
FNU Mamta
Akshar Shah
Vishnu Bangalore Thirumalesha

1 Abstract

In a relatively short period of time, social media has gained significant importance as a mass communication and public engagement tool for political and governance purposes. Rapid dissemination of information through social media platforms such as youtube, Twitter, provides politicians and campaigners with the ability to broadcast their message to a wide audience instantly and directly while bypassing the traditional media channels. Sentiment analysis or opinion mining is the field of study related to analyze opinions, sentiments, evaluations, attitudes, and emotions of users which they express on social media and other online resources. The revolution of social media sites has also attracted the users towards video sharing sites, such as YouTube. The online users express their opinions or sentiments on the videos that was on neutral stages like Presidential debates from multiple news channels which was uploaded before Nov 1. We have analyzed opinions posted by users about several Presidential Election 2020 videos. However, there are numerous challenges towards realizing this goal effectively and efficiently, due to the unstructured and noisy nature of social media data. We will investigate the nature and characteristics of the political discourse that took place on twitter and youtube during the American Presidential elections of November 2020. The goal is to perform exploratory sentiment-based analysis of data collected from tweets and youtube comments that was gathered before Election Day.

2 Introduction

Sentiment analysis is a machine learning technique that includes the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, aims to determine the attitude of a speaker or a writer with respect to some topic.

With the rapid growth of online information, a large amount of data is at our fingertips for this kind of analysis. However, the sheer volume of information was a daunting challenge itself. To separate relevant information from the irrelevant, and to gain knowledge from this unprecedented deluge of data, automatic algorithm is essential. In this project, we explored the use of Support Vector machine algorithm which is a supervised machine learning algorithms in learning sentiment classifier.

Today social networks have a very important role in online social discourse and especially during pre-elections period. Online Socoal Networks can either be used productively to perform dissemination, communication or administration in elections. The content posted by the users can represent their political belief or express negative opinion towards a political party or ideology. Twitter and YouTube while constitute two of the most popular online social networks attracting million of users daily, capture a very important proportion of this online discourse. Through analysis of this online discourse we can discover the main tendencies and preference of the electorate, generate patters that can distinguish users' favoritism towards an ideology or a specific political party, study the sentiment prevailed towards the political parties or even predict the outcome of the elections. Analyzing the sentiment is a necessary step towards any of these directions. Sentiment in these analyses is represented by a variable which gives polarity for 'Positive as 1' and 'Negative as 0'. Sentiment analysis usually requires 'text normalization', an initial preprocessing of the corpus in order to extract the lexical features that can significantly affect the performance. The steps of the preprocessing include tokenization, expansion of abbreviations

and removal of stop words. Each member has equally contributed in all phases and drives of the project as no part of work was individually distributed.

3 Literature Review:

There are 2 main learning techniques for sentiment analysis, one being unsupervised learning which relies on, for example, clustering and the other being supervised learning which makes use of labeled data for training the classifier. Most frequently used supervised algorithms for training a supervised model are Naive Bayes, maximum entropy, and support vector machines.(Pang et.al) Sentiment analysis can be divided into three components which are subjectivity classification, sentiment classification and complimentary tasks. From the data that is collected subjectivity classifier identifies each data point as a representation of an opinion or not.(Pang and Lee) Next being the polarity analysis, allows us to gain insight by quantifying the sentiment of the text. Complimentary tasks generally involve using the metadata such as location and user information for evaluating how tweets or comments based on state location correlate with the real world sentiment of people towards those candidates in this problem. Li et.al. have proposed various semi-supervised techniques to solve the issue of shortage of labeled data for sentiment classification . They have used under sampling technique to deal with the problem of sentiment classification i.e., imbalance problem. In their study, an attempt has been made to classify sentiment analysis for movie reviews using machine learning techniques. Two different algorithms namely Naive Bayes (NB) and Support Vector Machine (SVM) are considered. These two algorithms have also been implemented earlier by different researchers and results of all versions of implementation have been compared. It is observed that SVM classifier outperforms every other classifier in predicting the sentiment of a review.

4 Method

4.1 Dataset Generation

Twitter Data set was collected based on tweets that had replies to @realDonaldTrump and @JoeBiden. Around 20,000 tweets were collected directly referred to the candidates. Youtube data set was collected using Youtube's Data API. We chose the videos that was on neutral stages like Presidential debates from multiple news channels which was uploaded before Nov 1. Around 80,000 comments were collected. Comments data set was cleaned to have regular expression and filtered for having either Donald Trump or Joe Biden referenced in the comments. Combing both the data set we finally have 28,618 comments on the candidates.

4.2 Data Preprocessing

The most prominent step while building any machine learning model is data pre-processing as it will directly affect the result of our model. The more we pre-process the data, the more accurate the model performs. We removed emoticons during pre-processing to make sure the model learns from the text and not just memorizing emoticons. We selected only English comments . Initially, for both Twitter and YouTube, we follow a prerequisite set of steps for pre-processing of the corpus (tweets - comments); by removing punctuation symbols, URLs, by modifying mentions and hashtags and by removing starting characters of (@ and #).

This procedure removes the text noise and finally it allows the identification of the entity that was discussed by the users. The next step includes the transformation of the text to lower-case and the tokenization of the collected tweets. As a result of the previous steps, this sentiment analysis will be performed on lower cased and normalized sentences.

4.3 Implementation

After cleaning the dataset features can be extracted from it. The features are tokenized word of a comment. These words need to be converted to numerical vectors so that each review can be represented in the form of numerical data. The training is done on movie review using Stanford dataset for IMDB. Once the model is built our dataset can be used to get labels. Each comments is exacted and then vectorized. The vectorization of features are done using the following method.

4.3.1 TF-IDF

TF-IDF stands for Term Frequency and Inverse Document Frequency. This is a very common algorithm to transform text into meaningful representation of numbers which is used to fit machine algorithm for prediction. It is a tool for the purpose of extracting keywords from a document by not just considering a single document but all documents from the corpus. In terms of tf-idf a word is important for a specific document if it shows up relatively often within that document and rarely in other documents of the corpus.

4.3.2 Term Frequency

The tf-formula is a ratio of a term's occurrences in a document and the number of occurrences of the most frequent word within the same document. We would end up with stop words yielding high scores – and even if those would have been discarded before, a lot of words naturally show up often in a long text but aren't relevant to the specific document.

$$TF(t) = \frac{No.\ of\ times\ t\ appears\ in\ a\ document}{total\ no.\ of\ terms\ in\ the\ document}$$

4.3.3 Inverse Document Frequency

It represents the inverse of the share of the documents in which the regarded term can be found. The lower the number of containing documents relative to the size of the corpus, the higher the factor.

$$IDF(t) = log(\frac{Total\ no.\ of\ Documents}{No.\ of\ Documents\ with\ t\ in\ it})$$

4.3.4 TF and IDF

The TF*IDF algorithm is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term. The higher the TF*IDF score (weight), the rarer the term and vice versa. For a term t in a document d, the weight Wt,d of term t in document d is given by:

$$Wt, d = TFt, d \log(N/DFt)$$

Where:

- TFt,d is the number of occurrences of t in document d.
- DFt is the number of documents containing the term t.
- N is the total number of documents in the corpus.

4.3.5 Classifier: SVM

Support Vector Machine is a supervised machine learning algorithm which can be used for both classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. The basic idea behind the training procedure is to find a hyperplane, represented by vector that not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. (Singh, Piryani, Uddin and Waila, 2012) stated that Support Vector machine is a kind of vector space model based classifier which requires that the text documents should be transformed to feature vectors before they are used for classification. Training examples from the two different classes are separated by the hyperplane

$$g(x) = W^T x + b$$

Mathematically, this hyperplane can be found by minimizing the following cost function:

$$J(W) = \frac{1}{2} W^T W = \frac{1}{2} ||W||^2$$

subject to the separability constraints

$$y_i(W^T x_i + b) >= 1; i = 1, 2, 3...l$$

5 Results

Below are the results for SVM on the Imdb Review data set.

Accuracy	Precision	Recall
0.87	0.87	0.88

Based on polarity assigned the overall positive and negative comments count is taken and plotted on simple bar graphs.

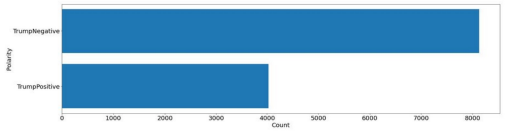

Figure 1: Trump's Polarity

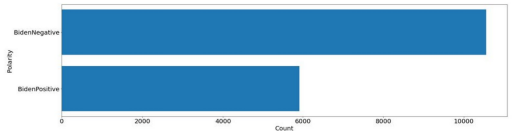

Figure 2: Biden's Polarity

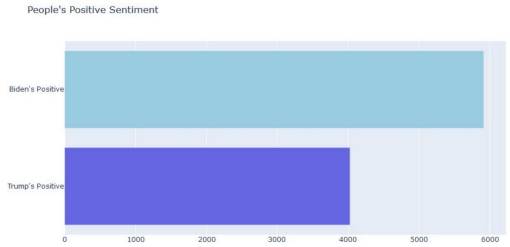

Figure 3: Positive Sentiments of the Candidates

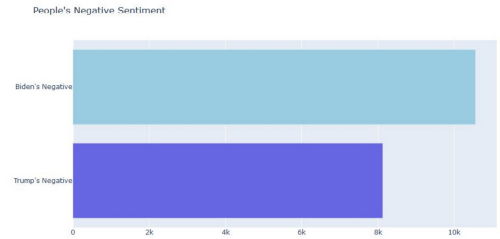

Figure 4: Negative Sentiments of the Candidates

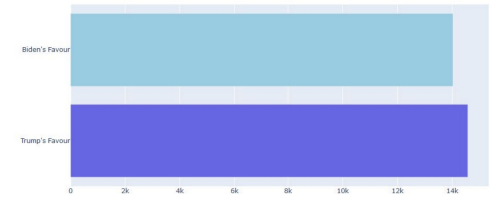

Figure 5: Overall Sentiments of the Candidates

Table with the count as below:

With the numbers Joe Biden seems to be more in news by getting comments from users, but though he leads in the positive comments, he is also leading in receiving the negative comments, so overall picture shows slightly to favor Trump (combining the candidates positive and opponents negative).

6 Discussion

For the above results and implemented method the major problem is in assigning polarity to the neutral comments , right now there might be neutral comments which will still be categorized as positive and negative with polarity 1 and 0 respectively. The training dataset is on Stanford's supervised data set for movie reviews and the model predicts polarity on the model built on IMDB movie review. If we try different dataset, it will give different results. So we can also conclude that context matters in sentiment analysis. Or to bypass the error we need to manually label the dataset and train model on that so that the context remains the same and predicted results are highly accurate.

7 Conclusion

One additional part we wanted to have was aspect based sentiment analysis, which could have given a better insight for the data we collected. Aspect Based Sentiment Analysis systems receives a set of texts as input such discussing a particular entity. It attempts to detect the main or the most frequently discussed aspects of the entity such as 'WikiLeaks', 'Democrats','Grand Old Party' and to estimate the sentiment polarity of the texts as per aspect that is whether the opinion of each aspect is positive, negative.

References

[1] Analysis of Twitter and YouTube during USelections 2020 Analysis of Twitter and YouTube during USelections 2020 by Alexander, Maria Oikonomidou, Despoina ,Polyvios Pratikakis,Sotiris Ioannidis arXiv:2010.08183v4 [cs.SI] 10 Nov 2020

[2] Compass: Spatio Temporal Sentiment Analysis of US Election by Debjyoti Paul, Feifei Li, Murali Krishna Teja, Xin Yu, Richie Frost, KDD 2017 Applied Data Science PaperKDD'17, August 13–17, 2017, Halifax, NS, Canada

[3] NLP-based Feature Extraction for the Detection of COVID-19Misinformation Videos on YouTube by Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, Simon Hegelich, 2020.nlpcovid19-acl.17

[4] N. Zainuddin and A. Selamat, "Sentiment analysis using Support Vector Machine," 2014 International Conference on Computer, Communications, and Control Technology (I4CT), Langkawi, 2014, pp. 333-337, doi: 10.1109/I4CT.2014.6914200.

[5] A Support Vector Machine Approachfor Detection of MicrocalcificationsIssam El-Naqa, Student Member, IEEE, Yongyi Yang*, Member, IEEE, Miles N. Wernick, Senior Member, IEEE,Nikolas P. Galatsanos, Senior Member, IEEE, and Robert M. Nishikawa, IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 21, NO. 12, DECEMBER 2002

[6] Movie Review dataset: Potts, Christopher. 2011. On the negativity of negation. In Nan Li and David Lutz, eds., Proceedings of Semantics and Linguistic Theory 20,636-659