Forecasting the winners and losers of 2020 elections from User predictions on YouTube comments

ENGL.681.01 - Final Project (ENGL68101.2221)

1st Shrinivas Sampath Muthupalaniyappan
Data Science Department (M.S), Rochester, New York, United States
sm3533@g.rit.edu

***Abstract*—Social media users often make explicit predictions about upcoming events. Such statements can be used to forecast the winners of presidential election. Can popular beliefs on social media predict who will win? To answer this question, we build a corpus from three YouTube channels namely MSNBC, CNN and FOX News. The first experiment is to perform topic modeling and calculate sentiment polarity on the two clusters to label the users intenion of voting. And then the second experiment was to manually create a list four different comment types and calculate the cosine similarity score for each other against the four different types comment types. Our model requires no manual labeling for sample comments to forecast the winner of 2020 presidential elections. We also trained the labeled corpus with BERT base cased and tested it against the MSNBC corpus.**

***Index Terms*—Word Embedding, Cosine Similarity, Topic Modeling, Forecast**

I. Introduction

In the digital era, there are millions of people broadcast their thoughts and opinions online [1]. These include opinions about people, movie reviews, presidential elections, sports, etc. Such statements can be useful to predict who is going to win [1]. For example, (a) in Table 1 criticizes about Trump, which tell the user doesn't like trump and the author must be voting for Biden, which is a strong assumption here. Whereas (b) clearly states that Trump is going to win. In contrast, (c) does not tell anything about the likelihood of Trump or Biden Wining. Prior work has made predictions such as presidential election, Oscar, Football, Presidential election (Swamy et al., EMNLP 2017) [2]and elections using tweet volumes (Tumasjan et al., 2010) [3] and also linking text sentiment to public opinion time series [5]. In this paper we explore whether the "wisdom of crowds" (Surowiecki, 2005) [3], as measured by users' explicit predictions, can predict outcomes of future events. We show how it is possible to forecast winners of 2020 Presidential Election from YouTube comments, by aggregating many individual predictions that assert an outcome. Previous approaches uses mechanical turk annotation to label the dataset (Swamy 2017) [2]. But in this paper we propose a new method to label the sample dataset and test the label quality by taking random sample of 100 labeled data. This approach does not require any manual label annotation but at last a human needs to check the sample of labeled corpus to measure the accuracy. This paper uses a set of two experiments to label the 2020 presidential winner from YouTube comments. i)Topic Modeling and Sentiment Analysis using TextBlob. ii) A novel method using a few human collected examples related to four different comment types converted to word embedding and compared with each comment from the user to forecast the winners. Moreover, this paper is extended to supervised learning after labeling the dataset.

a) " I am voting and trump needs to go he is horrible!!"
b) " vote red folks. cnn lies trump !!!"
c) " I'm voting neither trump nor Biden"

Table 1

II. Data

A. Data Collections and Preparation

The following YouTube channels' content is scraped to create the dataset: FOX News, CNN, and MSNBC. Around 2 billion data is stored in a json file format. All of them publish user comments from beginning of the channel till Aug 2021. The data was downloaded by using Python3 script and with the help of beautiful library. The data and script code scraped from website belongs to different author. Then we extract relevant data from all the three network and store it in separate CSV files.

B. Data Sets

As mentioned earlier, the dataset is collected from CNN, FOX news and MSNBC. While building the dataset, Fifteen days before the presidential election (held on November 3rd 2020) and five days after the presidential election data are again extracted from the originally scraped data. Further atleast one from each set should be presented in the comment to filter our relevant data set1("won", "vote", "win) and set2("trump,"donald,"joe","biden"). The distribution of each network is shown in Table 2.

III. Approach

Two experiments were performed to label the users intention. For forecasting the users intention only CNN corpus and 15 days of data before presidential election is used in both the experiments. In our first experiment in

News Network	No. of comments
CNN	45169
FOX News	61129
MSNBC	36729

Table 2

determining the winner of 2020 presidential election Topic Modeling and Sentiment Analysis is chosen. On top of preprocessed corpus topic modeling is applied to capture the overall cluster of trump and Biden. Topic modeling uses a library called BERT topic modeling to capture the overall structure of the data. The BERT Topic Modeling resulted in 2 clusters capturing the trump and Biden cluster containing 11932 and 6877 comments but around 9064 comments were unassigned to any of the clusters. Now our hypothesis is, trump cluster users might comment positive or negative about him. Same applies to Biden cluster. The Hypothesis is shown in Table 3 for sentiment Analysis task.

Cluster	Polarity	Users Intention
Trump Cluster	Positive	Vote Trump
Trump Cluster	Negative	Vote Biden
Biden Cluster	Positive	Vote Biden
Biden Cluster	Negative	Vote Trump

Table 3

In our second experiment we are treating the task of forecasting the winners of 2020 presidential election has the unsupervised labeling task. The unsupervised labeling task is accomplished by comparing the semantic structure of two sentences. For example the following two sentences ”china wants biden to win badly.” and ”if you’re voting for biden make sure you start touching up on your mandarin ... duolingo and babbel are free to download on any smart device”, both the sentences criticize about Biden. Since our dataset contains only two candidate names ”Biden” and ”Trump” in this scenario the users intention must be not to vote biden instead the user votes trump. In our approach, we believe these assumption helps us to forecast who are going to win the 2020 presidential election. At first we manually created a text file containing 4 different types of sentiment and for each sentiment four assumptions were made to know the users intention towards voting. As shown in Table 4.

Comment Types	Number of Example Comments	Users Intention
Criticizing Biden	30	Vote Trump
Appreciate Biden	30	Vote Biden
Criticizing Trump	30	Vote Biden
Appreciate Trump	30	Vote Trump

Table 4

IV. Implementation

A. Preprocessing and Data Exploratory

As mentioned earlier we are considering the CNN corpus only for our labeling task. By looking into the users comments the following preprocessing steps were undertaken removing Unicode(eg: //u2fed), lower-casing the characters, negation words were converted to their original form (eg: ’don’t’ to ’do not’), Regex rules were applied on top of this to remove new lines symbol, URL, strings contained in parenthesis. The architecture is shown in Figure 2.

Then applied few Data Exploratory Analysis to understand the distribution of the keywords(eg: ’win’, ’biden’) as shown in Figure 1. And also calculated the top 20 bigram_trigram distribution of the data. Finally tried to find the highest, lowest and average word count to trim the data. From the analysis the average word count was 222 words but for our analysis we don’t need that much amount of words to determine who people are voting. According to the average words present in the corpus, all the user comments are trimmed to ± 20 words to save memory and run time.

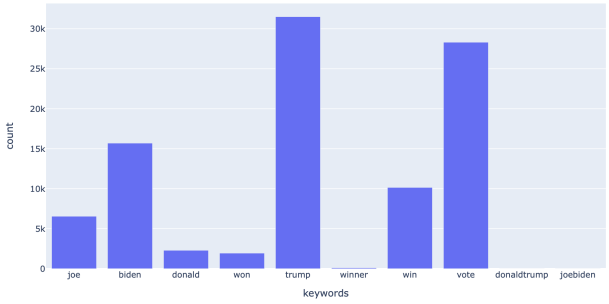

Figure 1

B. Experiment 1

In our experiment 1, the preprocessed text is fed to the Topic modeling layer to capture the overall cluster as explained in the Approach section. Then according to the polarity of the sentiment the user comments are labeled. Inorder to know the efficiency of the sentiment classifier. We randomly sampled 100 comments along with the labels and human annotated the voting preference from the users comments. The result was really bad, the sentiment classifier was able to correctly label only 50% of the label correctly. The output is shown in results section.

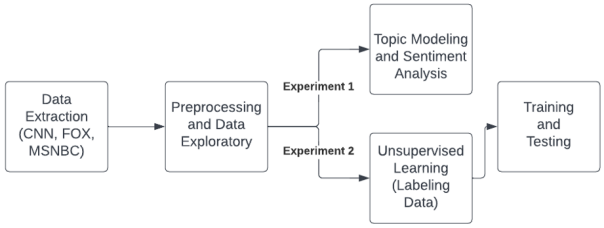

Figure 2

C. Experiment 2

In the experiment 2, As mentioned in the Approach section, we used the manually written human comments for four different categories each containing a total of 30 comments. Additional preprocessing step was performed to remove alphanumeric values for this experiment. On top of this we wrote a few regex pattern to capture the explicit user comments who wanted to vote for Biden and Trump. Around 4190 comments were labeled by explicit user comments out of 27873 comments. And then for unlabeled users, we calculated the cosine similarity for each user comments against manually created four different types of comments list. The argmax of all the comment list was chosen as the desired intention of the user according to Table 4. The result is explained better in results section.

Beyond the project, we extended the labeled data and created a BERT uncased model to train in CNN corpus and tested against the MSNBC corpus. The result is explained in the results section.

V. Results

In our analysis, we divide the preprocessed data into two experiments. In our first experiment, all the comments were clustered and a sentiment classifier labeled the data according to the polarity from Table 3. Then by random sampling we labeled the users vote, which resulted in 50% accuracy. Keeping this as our baseline, second experiment was performed on the same preprocessed data but calculted the cosine similarity of users comments towards four different lists of comment types created by us. Which resulted in 76% accuracy. The result of the forecasting was shown in Figure 3 and 4. There were 24% error rate in the second experiment. But compared to the first experiment, the second experiment looked promising. With the help of labeled dataset we tried to create a supervised model and then tested against the MSNBC corpus comments which resulted in 66% accurcay, 58% precision and 52% recall. In order to calculate the metrics, we labeled the test data manually by randomly sampling the MSNBC test data.

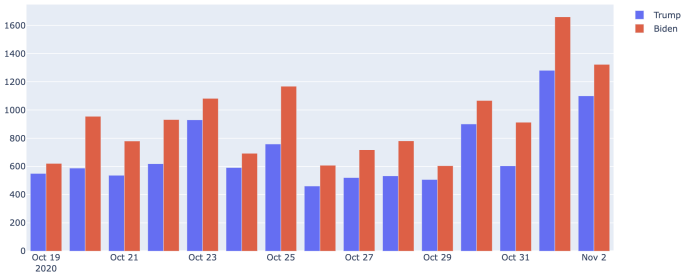

Figure 3

Conclusion

In this project, we successfully showed how can we label the YouTube comments using unsupervised learning and forecast the winners of 2020 presidential election. It was clear from the results Biden have approximately 14000 votes compared to Trump 10000 votes. Beyond the scope of this project, we extended the work to supervised learning to train a model with the labeled CNN corpus and test it against MSNBC corpus. In future, this work can be extended by applying active sampling [4] instead of manually picking the sentences from the data and then comparing it with the argmax of users comments. For this work, we will be using very few human labeled data and then by active sampling we can effectively find the nearest neighbours in the embedding space.

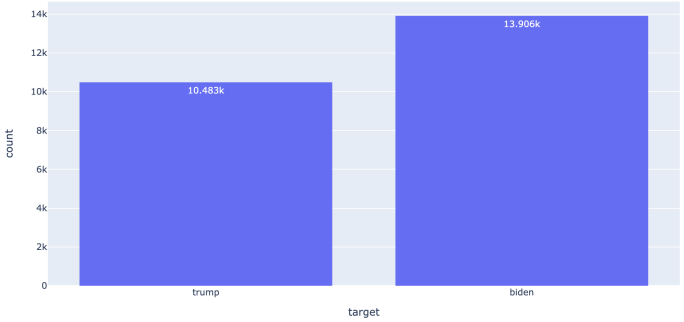

Figure 4

References

[1] Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. “i have a feeling trump will win.....”: Forecasting Winners and Losers from User Predictions on Twitter. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1583–1592, Copenhagen, Denmark. Association for Computational Linguistics, 2017.

[2] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 2010.

[3] James Surowiecki. The wisdom of crowds. Anchor Books, New York, NY,2005.

[4] Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas; Palakodety, KhudaBukhsh, Carbonell; AAAI 2020.

[5] From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series; O’Connor, Balasubramanyan, Routledge, Smith; ICWSM 2010.

[6] Does the @realDonaldTrump Really Matter to Financial Markets?; Benton, Philips; American Journal of Political Science, 2020..