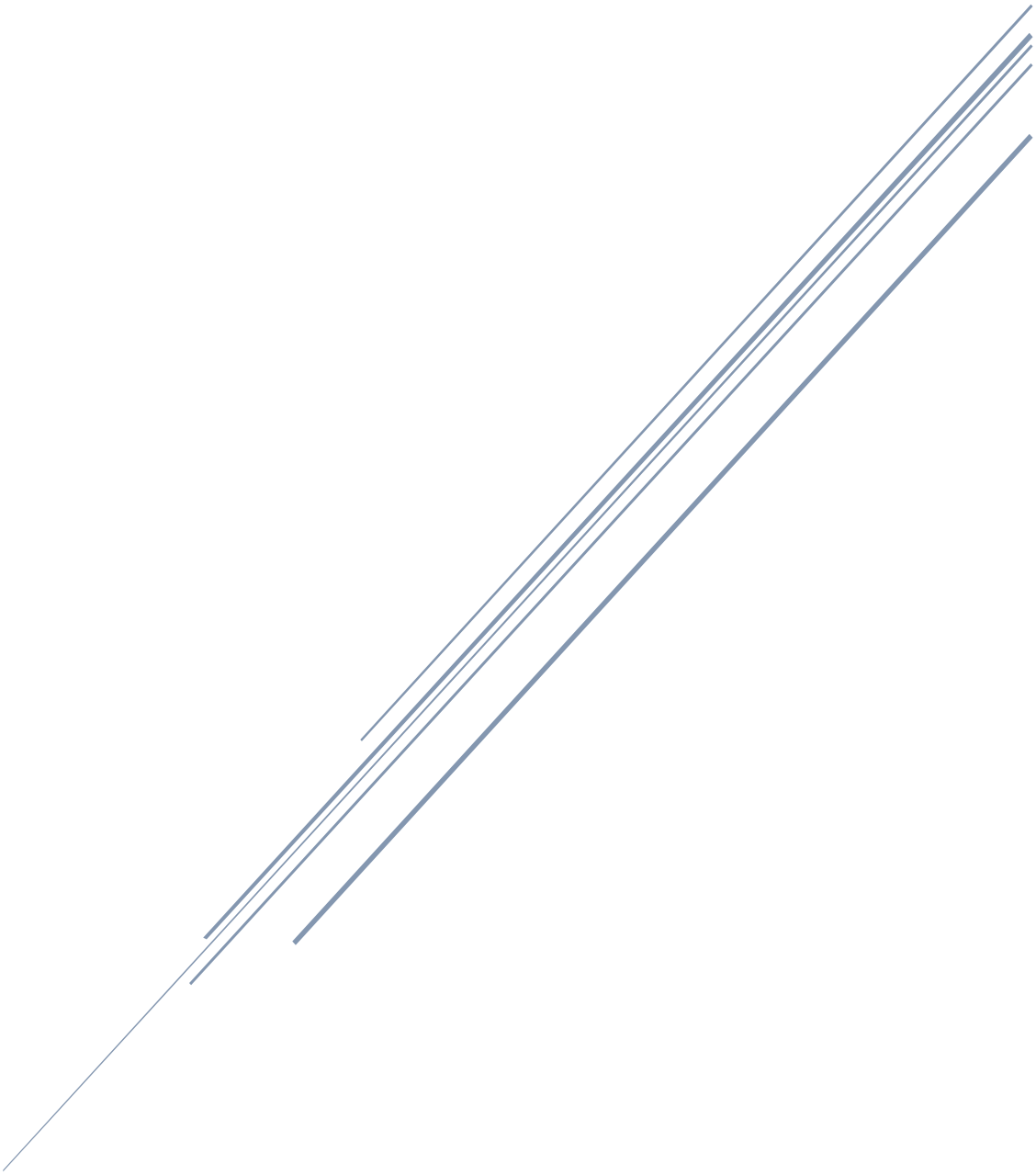
# US PRESIDENTIAL ELECTION 2020

## WITH ECONOMICS AND TWITTER



The Chinese University of Hong Kong
ECON 4140 Yam Tsz Chung 1155122291

**Abstract – 2020 United State presidential election held under the covid situation. In this paper, we succeed in predicting election winner in each county with Logistic Regression using economic variables but fail to identify swing counties. The second part of this paper, we built a multinomial naive bayes classifier for classifying election candidates related tweets with nearly 90% accuracy and applied sentiment analysis with TextBlob to discover swing states. The results comply with the election result in selected states but with some limitations.**

# Introduction

In my last paper, although we failed to identify which county will flip their vote in 2016 election with the models and economic variables we built, but we still can predict most counties voting result. The main idea of the paper is to evaluate our previous models and try to find another way to predict the swing votes.

The 2020 United State presidential election result was out, as polls suggested, Joe Biden won the election with 306 electro votes[1]. Some articles suggested that since the election was very close to the outbreak of covid, the policies and measures by Donald Trump affected the voting a lot[2]. By the fact that this year accept mailing their votes, more than 159 million people voted in 2020[3], which is the first time more than 140 million people voted in U.S. history. In the first part of this paper, we will update the economic variables and apply them into the models with transformation. Same as last paper, we will predict the election result by county and try to identify swing counties. With the much more complex voting situation, can our models perform well as last time while predicting the results?

As a well-developed country as US, social media is part of everyone's life. People can freely share their ideas on the platform. Different from traditional channel like television shows, social media allows feedback and comments, politicians like Trump can interact with his haters and followers. Even the cost of using social media as a promoting mean is much lower than setting up campaigns all over the country. Under the pandemic, social media plays an

[1] https://www.bbc.com/news/election/us2020/results
[2] https://www.bbc.com/news/election-us-2020-54390559
[3] https://www.usatoday.com/story/news/factcheck/2020/12/30/fact-check-fals-president-than-were-registered-u-s/4010087001/

important role in the election[4] as candidates can interact with all people and not bounded by the restriction in covid. In the second part, we will try to understand the relation between social media and election by analyzing Twitter. We will build a classifier to separate positive and negative tweets and discover the connect between them with the election.

## Data Collection and Cleaning

In the first section of this paper, I will update the economic variables with 2019 record provided by the US government Census Bureau[5], these data include employment, education, population, and poverty for each country. To indicate which candidate won in 2016 and 2020 elections, I assigned "1" as Republican (Donald Trump) won and "0" as Democrat (Joe Biden) won in every county. The data frame has a total of 51 columns (features) and 3111 rows (counties).

To understand social media, I choose twitter as my starting point to understand social media in election. Twitter is an extremely popular platform which allow users to use hashtag to clearly identify what topics they are discussing. In this paper I used Twitterscraper to collect tweets from 1st October to 8th November, exactly 1 month before election until the end of the election. As we would focus on the two dominant candidates Donald Trump and Joe Biden in this paper, we will collect two sets of tweets. The first set contains keywords with #trump and #donaltrump and the second set contains #joebiden and #biden. Which directly shows that the tweets are these two candidates related. There are total of 1.8 million tweets collected.

After collected the data, we need to clean and process it to help us in the analysis. For the economic variables, I normalized the data and performed dimensionality reduction on the dataset with Linear Discriminant Analysis. Since LDA uses class information, it will perform better than principal component analysis. We will use the transformed data as the inputs in the models. Since we can expect election has a high autocorrelation, we will use the election result in 2016 as a dummy variable in the model.

For the text-data, we will prepare our twitter dataset for further analysis. We first converted

[4] https://www.marketwatch.com/story/the-election-is-being-fought-on-social-media-amid-the-pandemic-11602529760
[5] https://www.census.gov/data

the tweets to lower case, removed special characters (@#$& etc.) and hyperlink that difficult for the machine to analyze. As this paper focus on US election, I used Pycld2 (Google Chromium's embedded compact language detection library) to detect tweets in English language to keep the model simple. I also removed stop words and stem with NLTK and get polarity between 1 (positive tweets) and -1 (negative tweets) with TextBlob.

## Model Result in Economic Variables

Since we only have a total of 3111 counties to train and test our model, the model size is not large enough. We will apply Cross-Validation with GridSearchCV with 25% as testing size to train out models which is 778 counties for testing. The best model is logistic regression with c = 1. The result of the model returns us a 0.98 F1-score in testing set and 0.97 when using all counties as input and predict election result in every county. The model was able to predict 3033 counties correctly and it performed quite well like the last paper (see Figure1. Predicted Counties Election Result). It can even predict some swing counties correctly.

For The model finding swing counties, I added a dummy variable to indicant which counties is a swing county in 2016 election and split the data with 30% testing size where 934 counties for testing. The best model is decision tree with entropy, maximum depth = 5, minimum samples leaf = 2 and minimum samples split = 2. The model failed to identify swing counties in training and testing state. While it can only identify 4 out of 79 swing counties in all input (see Figure2. Predicted Swing Counties Result). The model performed badly in both stages. Same as last paper, economic variables cannot accurately identify swing county.

We have tested to use economic variables to predict the election result in 2012, 2016 and 2020, although the models performed quite well in getting the result, they were not able to identify swing state. We can conclude that whether a state will or will not swing its vote may not depends on its change in economic status. Can we identify the swing states by social media?

## Model Result in Twitter

To understand social media, we will build ourselves a multinomial naive bayes classifier to classify positive and negative tweets towards the two candidates.

The reason of choosing multinomial naive bayes is that we assume features are independent and it is important for our study while making the model simple. First, we vectorize the tweets with TF-IDT (frequency-inverse document frequency) as this vectorizer will return with the weight to evaluate how important a word is to our model. I assumed that the important words would repeat many times in different tweets, TF-IDT can help us to identify and weight them. After that we dropped the neutral statement with polarity equals 0 as we are doing a 2-class classification and the datasets left with both candidates positive and negative tweets. Total Tweets left for Donald Trump = 395501 and Joe Biden = 30274.

To train our models, we split both datasets with 33% testing size (Trump = 130516 and Biden = 99905) and assign 1 for polarity greater than 0 and -1 for tweets with polarity smaller than 0. I also apply a five-fold cross validation with GridSearchCV to train and test the models.

After running the codes and algorithms, the model returns us with a 0.82 F1-score in testing set and an overall performance at 0.88 in the Trump set (see Figure 3. Donald Trump Tweets MultinomialNB). Similarly, 0.83 in testing set and a 0.87 overall performance in Biden set (see Figure 4. Joe Biden Tweets MultinomialNB).

The models perform well in separating positive and negative tweets in both sets. After evaluating the model, I assumed a negative tweet in Trump set is a pro Biden tweets and vice versa, the model returns us that Pro-Biden tweets = 365,566 more than Pro Trump tweets = 332,676 (see Figure 5. Percentage of Pro-Candidates Tweets). The result saying that twitter users prefer Biden over Trump.

## Social Media and Swing States

### I. Preparation

In this section, I grouped the tweets based on their user's profile location by states and assume they are voters in that specific state. After that I chose the states with enough tweets in both swing state and stable state. For both state tweets set, I also drop the neutral statement. Furthermore, "1" in Trump set meaning 1 tweet supporting Trump and "-1" meaning 1 tweet supporting Biden, vice versa in

Biden set. Then that we can compare sentiment percentage with the voting result percentage.

The swing state chosen are Georgia, Pennsylvania, Arizona, Wisconsin, and Michigan. The stable states used are New York, California, Florida, and Columbia.

### II. Compare results

The swing states sentiment results grouped by states are showed in the appendix (see Figure 6. Swing States Sentiment Results). After that we compute the margin between two candidates and compare it with the election result margin in each state. Where (D) meaning Democrat (Joe Biden won) and (R) meaning Republican (Donald Trump won).

In the swing states we chose, the sentiment results comply with the voting result, that Democrat won all sates while the margin between sentiment and voting were very close. The sentiment mostly over-estimate the voting margin with around 1% in each swing states (see Figure 7. Swing States Margin).

The result in stable states is more or less the same, the stable states sentiment results grouped by states are showed in the appendix (Figure 8. Stable States Sentiment Results). The sentiment results also comply with the voting result with Florida won by Trump and other won by Biden (Figure 9. Stable States Margin). But the margins are not accurate enough as sentiments under-estimate the voting margin by a lot compare with previous observation in swing states.

## Conclusion

To conclude this paper, economic variables were able to predict the voting result on county level with over 90% of accuracy and we had tested it with 2012, 2016 and 2020 elections. These variables are not able to discover swing counties so that we can say that a county changing their vote is not deal to the change in their economic status but for other reasons.

To identify swing states, we turn our focus from economic to social media. As we explained in the previous parts, social media may reflect election result. We first built a multinomial naïve bayes classifier that can classify positive and negative tweets with nearly 90% accuracy in the texts. The sentiment score in swing states' tweets can reflect the states voting result and the vote margin. Which it can also reflect the voting result in stable states, but the problem of both stages is the under and over-estimation of the voting margins.

## Limitation

The total number of tweets reduced a lot when they are grouped by states, the compare with the original dataset. This may deal to most of the tweets have no location record it is difficult to get a hand full of tweets for the comparison. Even the location recorded does not necessarily mean the account owner is a voter in that state. We chose to use Twitter as the starting point, but Twitter is not the only social media used by US people. Other platform like Facebook and Pinterest are also popular within United State.

I also tried to account for the covid influence in both economic and twitter analysis. Although State level data are easy to get, the county level data is more difficult and messier. In scrapping twitter, if I add #covid as a keyword, the dataset will contain a lot of election-unrelated tweets that are not our focus. In this paper, I assumed the keywords used have already accounted for the covid effect as covid appears in the data sets many times.

## Suggestions

For suggestion, one can collect more keywords to do the sentiment analysis. In this paper we only use #DonaldTrump, #Trump, #JoeBiden and #Biden, which there are many keywords that can also be used like Anyone but Trump, For the People and Families First. With a better machine, we can collect more keywords. With the increased keywords, we must deal with unrelated tweets that also share the keywords. In later work, we can also account for a longer period. We only use the closest month for this paper as I think this month can reflect immediate effect on the election. With a longer timeseries, we can assume earlier month have smaller effect and closer month have more effect on the election.

For later work, we can collect more social media post to reduce the bias as people choose which social media they usually use. An article suggested that around 69% of users in Twitter

are Democrats[6]. If we can collect more post and social media users data, we can reduce the bias created by this.

My data in this paper cannot fully utilize Neural Network, I tried using ANN supported by TensorFlow, but the result did not improve by a lot, this is because my datasets are rather simple and less complex. In the future if we account for a longer period and include more keywords, the model will have a lot of inputs thus using Neural Network may have a better result. Instead of using TextBlob to get the sentiment score for text-data, we may use Neural Network to build a model specifically for election related text and this may work better than using TextBlob.

We can also extract knowledge in the text data including the relationship between texts and knowing what issues are the most concerned by twitter users, even planning the election strategies.

[6] https://techcrunch.com/2020/10/15/pew-most-prolific-twitter-users-tend-to-be-democrats-but-majority-of-users-still-rarely-tweet/

**Figure 1. Predicted Counties Election Result**

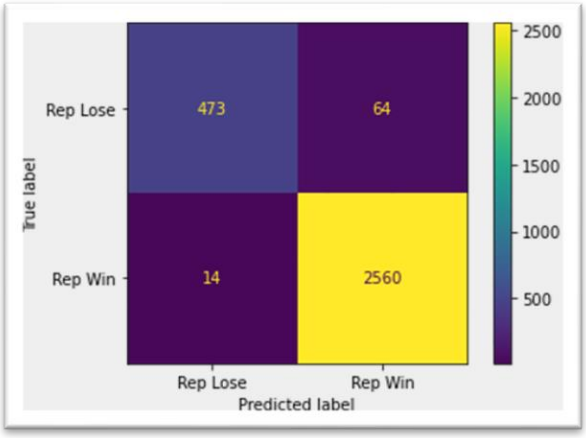



**Figure 2. Predicted Swing Counties Result**

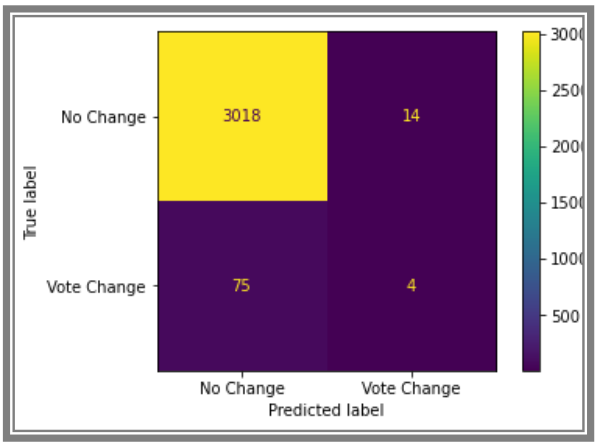



**Figure 3. Donald Trump Tweets MultinomialNB**

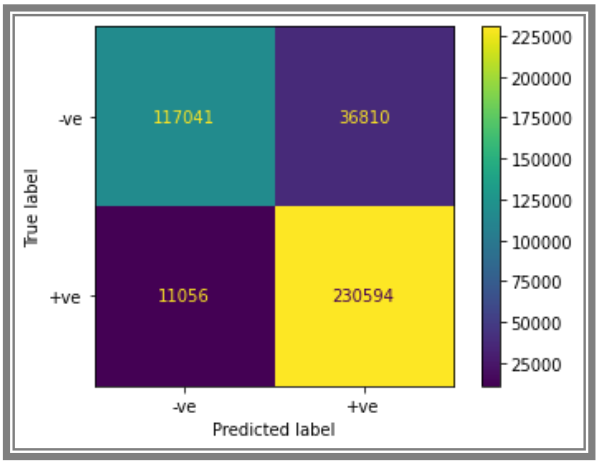


**Figure 4. Joe Biden Tweets MultinomialNB**

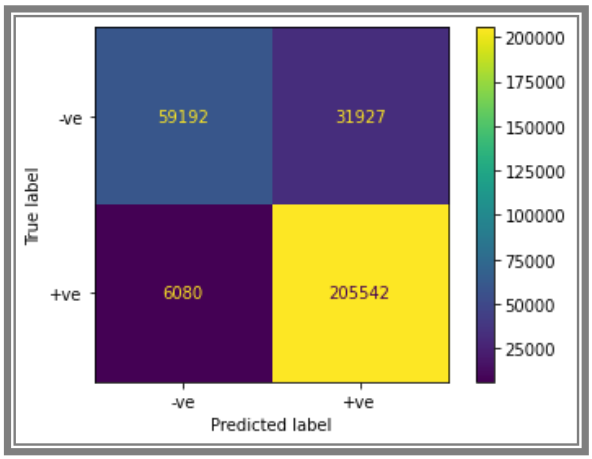



**Figure 5. Percentage of Pro- Candidates Tweets**

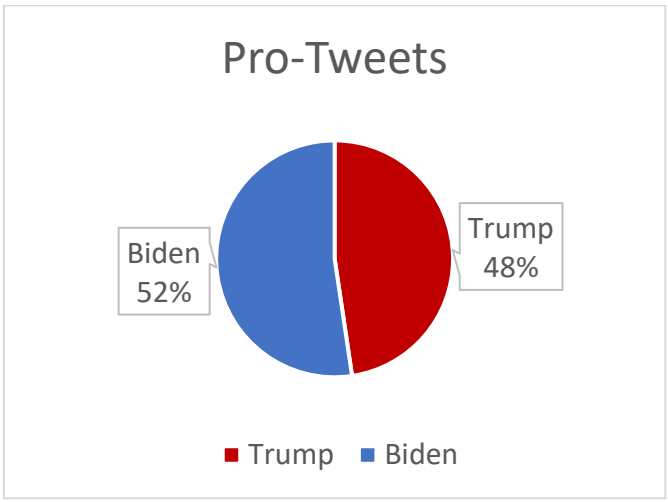



**Figure 6. Swing States Sentiment Results**

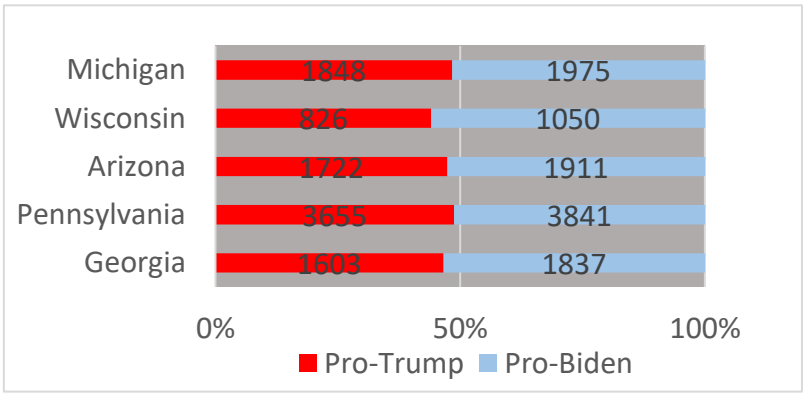


**Figure 7. Swing States Margin**

| State Name | Sentiment Margin | Voting Margin |
|---|---|---|
| Michigan | 1.6% (D) | 2.78% (D) |
| Wisconsin | 5.0% (D) | 0.63% (D) |
| Arizona | 2.6% (D) | 0.31% (D) |
| Pennsylvania | 1.2% (D) | 1.16% (D) |
| Georgia | 3.0% (D) | 0.24% (D) |

**Figure 8. Stable States Sentiment Results**

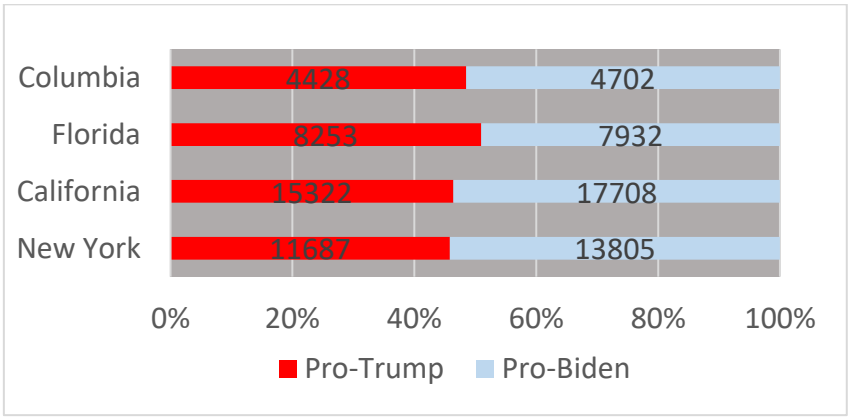



**Figure 9. Stable States Margin**

| State Name | Sentiment Margin | Voting Margin |
|---|---|---|
| Columbia | 2% (D) | 87.6% (D) |
| Florida | 1% (R) | 3.3% (R) |
| California | 3% (D) | 13.5% (D) |
| New York | 4% (D) | 23.1% (D) |

# REFERENCE


1. US Election 2020. Retrieved from BBC. https://www.bbc.com/news/election/us2020/results
2. Anthony, Z. (2020 October). *Trump Covid: How will this affect US election?* Retrieved from BBC News: https://www.bbc.com/news/election-us-2020-54390559
3. Adrienne, D. (2020 December). *Fact check: Over 159 million people voted in the US general election.* Retrieved from USA Today: https://www.usatoday.com/story/news/factcheck/2020/12/30/fact-check-fals-president-than-were-registered-u-s/4010087001/
4. Jon, S. (2020 October). *The election is being fought on social media amid the pandemic.* Retrieved from MarketWatch: https://www.marketwatch.com/story/the-election-is-being-fought-on-social-media-amid-the-pandemic-11602529760
5. Sarah, P. (2020 October). *Pew: Most prolific Twitter users tend to be Democrats, but majority of users still rarely tweet.* Retrieved from Techcrunch: https://techcrunch.com/2020/10/15/pew-most-prolific-twitter-users-tend-to-be-democrats-but-majority-of-users-still-rarely-tweet/
6. S. Ray, *6 Easy Steps to Learn Naive Bayes Algorithm (with codes in Python and R)*: https://www.analyticsvidhya.com/blog/2017/09/naive-bayesexplained/
7. S. Shepard. (2016 August). *The 11 states that will determine the 2016 election*, Retrieved from Politico: http://www.politico.com/story/2016/06/donald-trump-hillary-clintonbattleground- states-224025
8. M.P. Cameron, P. Barrett, and B. Stewardson, *"Can social media predict election results? Evidence from New Zealand"*, Journal of Political Marketing, 15(4), 2016.
9. "TextBlob: Simplified Text Processing", https://textblob.readthedocs.io/en/dev/
10. SentiWordNet, http://sentiwordnet.isti.cnr.it/
11.