# Using Sentiment Analysis of Select 2020 Democratic Presidential Candidates to Predict Tweet Reach

Salem Abuammer, Ekrem Guzelyel, Muhammad Shareef

**Problem Overview:**

Social media usage has increasingly become a daily part of our lives; Facebook reports more than 2.1 billion users use Facebook, Instagram, Whatsapp, or Messenger everyday on average [2]. In October 2019, Twitter reported 145 million daily active users [3], and Snapchat reported 210 Daily active users [4]. This vast audience is a prime target for advertisers, but increasingly, for politicians as well.

Social Media played a significant role in the campaign strategies of both 2016 presidential candidates. In a hearing held by the Senate Intelligence Committee in 2017, Facebook lawyer Colin Stretch testified that both campaigns spent a combined $81 million on Facebook advertisements[5].

Much has been said about the negativity of these campaigns, and of politics in general. CNN analysis of data showed that of 69,500 ads spotted by Kantar Media between November 1-5, 92% had either negative messages or focused on differences between the candidates. Only 3% focused on positive messages about Clinton, and 5% focused on positive messages about Trump[1]. Trump was criticized profusely for his usage of negative rhetoric during and after his campaign, yet he was still elected in 2016. This brings up the question as to whether he was elected in spite of his negativity, or, if he was elected because of his negativity. Do Americans resonate more with positive messaging, or negative messaging?

With the 2020 election coming up, there is plenty of data available to analyze. Our approach is two-fold. Firstly, we will use sentiment analysis to determine whether or not a tweet by a Democratic presidential candidate is positive or negative. Then, we will use the sentiment as part of a model to predict the “reach” of a tweet. The question we wish to answer is, “Do negative tweets by Presidential candidates spread further than positive tweets?” Answering these questions can give us insight into the American populace, and help future presidential campaign strategies.

**Data**

The data we will be collecting is Tweets from four Democratic Presidential Candidates: Bernie Sanders, Joe Biden, Elizabeth Warren, and Andrew Yang. We chose the first three candidates because they were the front-runners for the nomination. We choose Yang partly because he appeared in fourth place in some polls, but also because he is a political outsider like Trump and is relatively young, so his online strategy may be significantly different from the others.

This data will be collected using the Twitter API, and it will be the 1000 most recent tweets from a candidate, ignoring retweets.

**Method**

To conduct the sentiment analysis, the Watson Natural Language Understanding collection of APIs from the IBM Cloud will be used. This set of APIs can analyze text to help understand keywords, entities, sentiment, and even allows for the creation of custom APIs. It also allows for insight into the intent of the text. We initially were using the NLTK, but this collection of APIs is much more robust and returns additional information related to sentence structure, entities.

We will be constructing four models, one for each candidate, to predict the reach of a tweet given the text of the tweet. The reason we are creating a model per candidate is that some candidates have a much lower follower count than others, so their reach will inherently be lower. For example, Bernie has nearly 10 times the number of followers as Yang (9.91 Million to 986 thousand), so even if he tweets the exact same thing as Yang, it will have a further reach than Yang's tweet. Creating the model per candidate allows us to specifically see the impact a tweet's sentiment has on its reach.

As of now, we have not concretely decided on which specific model to use. We will test out Naive Bayes, Logistic Regression, Random Forests, and Neural Networks and see which gives us the highest reach prediction accuracy.

**Intermediate Results**

So far, 1,000 tweets per candidate, not including any retweets or quote tweets, were collected from each of the candidates. Preliminary sentiment analysis was conducted using NLTK, but we found the IBM collection of APIs to be more accurate, and as such have switched to using that for Sentiment analysis. We created four matrices per candidate from the sentiment analysis results: an emotion matrix, a keyword matrix, a sentiment matrix, and a character matrix. The emotion matrix contains the probability values that the tweet is expressing anger, disgust, fear, joy, and sadness. The keyword matrix is a one-hot like encoding of relevance values from 0 to 1 for each keyword in the tweet.
The sentiment matrix contains a sentiment score per tweet, from -1 to 1. The character matrix contains the number of characters per tweet. Each of these matrices will be combined per candidate, ultimately creating 4 feature matrices, 1 per candidate, to be plugged into the model we decide on using.

**Related Work**

In [6], Caetano et al. collected 4.9 million tweets from 18,450 users and their contact networks between August and November 2016. They identified political and non-political tweets on each users timeline, then carried out sentiment analysis on each tweet. They used the SentiStrength tool to perform the sentiment analysis, and they used the sentiment analysis results to define six user classes regarding sentiment toward Trump and Hilary Clinton: Whatever, Trump Supporter,

Hilary Supporter, positive, negative, and neutral. They then analyzed the twitter homophily, that is, the tendency of individuals to follow or interact with those similar to themselves.

In [7], Hamlin and Agrawal wrote a program to collect tweets that mentioned one of the two candidates(not both), sorted these tweets by state based on the user's location description tag, then created their own Sentiment analysis algorithm that used SentiWordNet to classify these tweets as positive or negative to the candidate. They then looked at the percentage of positive/negative tweets per state to determine the favorability of the candidate per state, compared the favorability of each candidate, and predicted that the candidate with the higher favorability would win that state. In total, they collected 1,873,150 tweets, and correctly predicted the results for 34 states. Some drawbacks of this method is that it failed at detecting sarcasm and could not detect sentiment from complex word clauses or phrases consisting of multiple words, since it worked by summing up values of individual words. For example, the phrase "Hitlery Clinton" was not detected, even though it was extremely negative towards Hillary.

In [8], Joyce and Deng conducted sentiment analysis of tweets for the 2016 US Presidential Election. For their sentiment analysis, they used the OpinionFinder Lexicon, which contains approximately 1,600 positive and 1,200 negative words, combined with another Lexicon to account for misspellings. They used the National Language Toolkit (NLTK) to implement the Naïve Bayes algorithm, and collected 5,000 positive ad negative tweets for each candidate (20,000 total). They then correlated this sentiment data with national opinion polls, and found a correlation coefficient of approximately 40%-60%

In [9], Heredia et al. analyzed 3 million tweets collected from September 22$^{nd}$ to November 8$^{th}$ that were related to either Trump or Hilary. They collected these tweets using the twitter API and partitioned these tweets by state based on the location in the user's profile. They used another dataset of 1.6 million tweets, 800,000 positive and negative, to train a convolutional neural network (ConvNet) for determining sentiment in the 3 million tweet election dataset. The ConvNet was implemented using TensorFlow 1.1.0 with the Python API. They predicted the winner of each state based on the percentage of positive tweets related to each candidate.

In [10], Pak and Paroubek used the twitter API to collect a corpus of tweets to create a dataset of 3 classes: positive, negative, and neutral. They queried for text emoticons, such as ":-)" or ":-(" to train a classifier to recognize positive and negative sentiments. They assumed an emoticon within a message represented the emotion for the whole message, and so all the words in that message were associated with the emotion. Tweets were cleansed by removing URLSs, usernames, twitter specific words such as RT. Stopwords were removed, and n-grams were created out of consecutive words. They then built a sentiment classifier using the Naive Bayes classifier. A large issue with this approach is that emoticons have fallen out of general use in favor of emojis, and these can be used sarcastically, so the accuracy could be lower.

Our project is similar in that it is using sentiment data from Twitter to give insights into a US Presidential election. It differs from all of these approaches because we are not analyzing the sentiment of the general population towards the presidential candidates and trying to predict the results of the election; we are analyzing the sentiment of the candidates' tweets themselves. Then, we will create a model to predict the "reach" of a tweet based on the sentiment. This will

help us to determine whether a negative or positive tweet, and by extension a negative or positive campaign, resonates more with the American people.

**Who does what?**

Muhammad: Refining results, webapp/command line interface, report creation, presentation creation
Ekrem: Model creation, training, webapp/command line interface, report creation, presentation creation
Salem: Data collection, sentiment analysis, webapp/command line interface, report creation, presentation creation

**Timeline**

The remaining steps are

- Conducting the sentiment analysis for all tweets (11/3)
- Creating the model to predict reach using sentiment scores as an input (11/9)
- Testing (11/10)
- Refining the model(11/11)
- Creating the web application and command line interface(11/16)
- Analyzing results (11/19)
- Connecting the model to the API (11/19)
- Presentation (11/20)
- Final report (11/25)

**References**

[1] Wallace, Gregory(2016, November 8). Negative ads dominate in campaign's final days. *CNN*. Retrieved from https://www.cnn.com/2016/11/08/politics/negative-ads-hillary-clinton-donald-trump/index.html

[2] E. Culliford, P. Dave, and A. Rana (2019, October 30 ). Facebook sales grow as users tick up; Zuckerberg defends political ads. *Reuters*. Retrieved from https://www.reuters.com/article/us-facebook-results/facebook-beats-on-profit-as-costs-grow-slower-shares-rise-idUSKBN1X92H6

[3] Associated Press(2019, October 24). Weak profit, revenue, overshadow Twitter user growth. *The Washington Post* Retrieved from https://www.washingtonpost.com/business/weak-profit-revenue-overshadow-twitter-user-growth/2019/10/24/228632f2-f669-11e9-b2d2-1f37c9d82dbb_story.html

[4] T. Maglio, and S. Bursch (2019, October 22). Snap Q3 Earnings: Snapchat's Daily Active Users Rise to 210 Million. *Yahoo Entertainment*. Retrieved from https://www.yahoo.com/entertainment/snap-q3-earnings-snapchat-daily-202011514.html.

[5] Open Hearing: Social Media Influence in the 2016 U.S. Election: Hearing Before the Select Committee on Intelligence of the United States Senate. 108th Congress, First Session(2017)

[6] J.A. Caetano, H.S. Lima, M.F. Santos, H.T Marques-Neto, “Using sentiment analysis to define twitter political users’ classes and their homophily during the 2016 American presidential election”. J Internet Serv Appl. 2018;9:18.

[7] T. Hamling and A. Agrawal, “Sentiment analysis of tweets to gain insights into the 2016 US election”. Columbia Undergraduate Sci J. 2017;11:34–42.

[8] B. Joyce and J. Deng, "Sentiment analysis of tweets for the 2016 US presidential election," 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, 2017, pp. 1-4.

[9] B. Heredia, J.D. Prusa, and T.M. Khoshgoftaar, “Location-Based Twitter Sentiment Analysis for Predicting the US 2016 Presidential Election.” The Thirty-First International Flairs Conference, 2018.

[10] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.” LREc, vol. 10, pp. 1320-1326. 2010.