Predicting Tweet Reach of Select 2020 Presidential Candidates from Sentiment Analysis

Salem Abuammer, Ekrem Guzelyel and Muhammad Shareef

Department of Computer Science, Illinois Institute of Technology, Chicago

Abstract

Social media usage is part of our daily lives and is playing significant roles in the campaign strategies of politicians. We focused on one platform, Twitter, to determine whether the negativity or positivity of a tweet has a correlation with the reach of the tweet. We collected 1,000 tweets from each of four select Democratic Presidential Candidates, ran sentiment analysis using IBM's Natural Language Understanding API, then used the output of this as features into a linear regression model that predicts whether a tweet would be high reach or low reach. After carrying out these steps, we found that tweets exhibiting disgust, fear, and sadness were more likely to be high reach across the board, and tweets with positive sentiment scores were more likely to be classified as low reach.

Keywords: Machine Learning, Sentiment Analysis, Presidential Election

---

1. 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. A common phrase in journalism is "If it bleeds, it leads" meaning that stories involving death and invoking fear have a higher likelihood of being the ratings leader. Is similar behavior exhibited for negative, fearful, or depressing tweets? Do Americans resonate more with positive messaging, or negative messaging?

With the 2020 election coming up, there is plenty of data being generated by politicians that is available to analyze. Specifically we will utilize Twitter, a social media site which allows users to post messages known as tweets to their user profile, known as their feed. Originally tweets were arbitrarily limited to 140 characters, resulting in brevity, but this was updated in 2017 to 280 characters. These tweets can either be public and viewable to the world, or private and viewable only to individuals who are approved by the tweet poster. There are four ways in which a tweet can be interacted with: viewed, replied to, liked, and retweeted. A retweet is a repost of another Twitter user's tweet on one's own profile. In order to gauge the reach of a tweet, one must analyze the impressions and engagement of that tweet. Twitter impressions are a tally of all the times a tweet has been seen in people's feeds, whereas engagements are a count of all the times an individual interacted with a tweet. To answer our question about which type of political messaging spreads further on twitter, we need a multi-step analysis.

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 tweet's sentiment as input to a model that will 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 future presidential campaign strategies.

2. Background and 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 user's 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 22nd to November 8th 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 in order to predict the results of the election. Instead 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 spreads further, and by extension whether a negative or positive campaign resonates more with the American people.

3. Data Collection

The data we collected 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 at the time of the project proposal. 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.

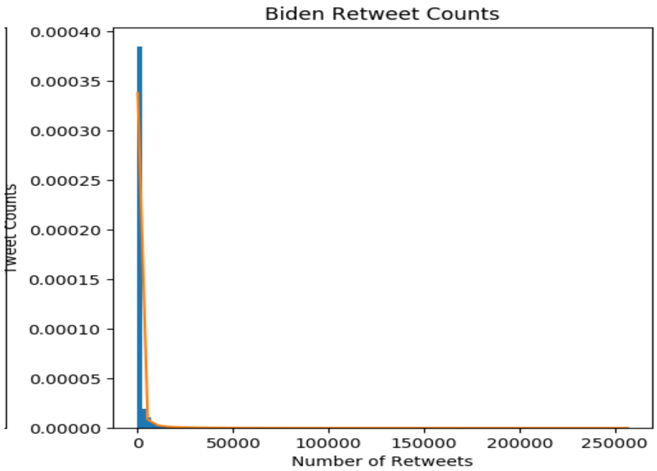
We collected 2,989 tweets for Elizabeth Warren, 2,186 tweets for Andrew Yang, 2,863 tweets for Joe Biden, and 2,231 tweets for Bernie Sanders. This was done using a custom python function which made calls to the Twitter API, got the maximum number of recent tweets, and filtered out retweets.

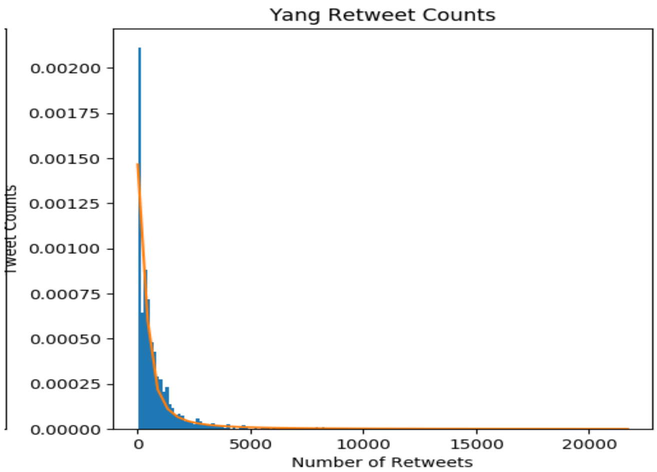
From these collections of tweets, we ignored the 1,000 most recent tweets taken for analysis. The reason for this was to make sure that the number of retweets for those tweets had stabilized, as they were older and less likely to be appearing in users' feeds. If the 1,000 most recent tweets were taken instead, it would be possible that the number of retweets significantly increases after the time we collected the data, resulting in inaccurate assessments of the tweet reach. We ended up analyzing 1,989 tweets for Elizabeth Warren, 1,186 tweets for Andrew Yang, 1,863 tweets for Joe Biden, and 1,231 tweets for Bernie Sanders.

4. Approach

After collecting the data, we performed basic data analysis to see the number of retweets in order for us to draw a cutoff for high reach and low reach for each candidate.

We plotted a histogram of the number of retweets for each of the candidates. We expected these histograms to be long-tailed, since there are usually a smaller number of really high reach tweets and a much larger number of low reach tweets. The histograms are shown below. Our initial estimated cutoff points for "high reach" was over 5000 retweets for Warren, over 7500 for Bernie, over 15000 for Biden, and over 2500 for Yang. These values were approximations obtained by looking at the elbow points of each respective graph, to see where the long-tail distribution begins.





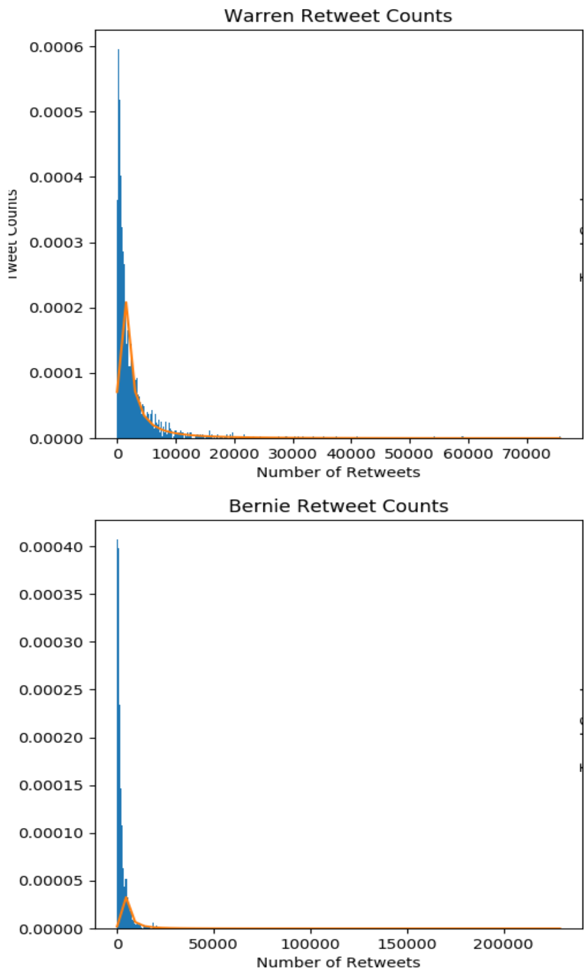


Figure 1: Histogram and Gamma distribution for number of retweets

These datasets ignored the most recent 1000 tweets of each of the candidates, since the more recent tweets were still in the immediate attention of the public, which means the retweets for those tweets are bound to fluctuate.

After the data was collected and the high reach cutoff point was estimated, we fit an inverse gamma distribution to the histogram of each candidate. This was to get a more accurate and less arbitrary cutoff for what defines a high reach tweet. After getting the results from the inverse gamma distributions, we made the cutoffs for Warren and Biden so that high reach tweets were in the upper 15% of the tweet dataset and low reach tweets were the bottom 80% of the tweet dataset. The remaining 5% of tweets at the boundary region were removed, to help create a more defined distinction between high and low reach tweets. Since Biden and Yang had more tweets with lower reach we made the cutoff at bottom 45% upper 55% for Biden, and bottom 57.5% upper 67.5% for Yang to balance the ratio.

After determining the cutoff, we conducted a preliminary sentiment analysis on tweets. To conduct the sentiment analysis, the Watson Natural Language Understanding collection of APIs from the IBM Cloud was 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 were initially using the NLTK library, but the IBM collection of APIs was much more robust and returned additional information related to sentence structure, entities. We took samples from our data set and ran them through the sentiment analysis for both NLTK and NLU, and from our manual review, we found that the IBM NLU was more accurate at determining sentiment than the NLTK. We then decided to use the NLU for the full sentiment analysis.

We created a function in Python that takes in a list of tweets as an argument, calls the NLU API, and returns a Python dictionary mapping the tweet text to all of the sentiment scores. These results were used to train our models.

We decided on 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 one model per candidate is that there may be unaccounted for variables and factors for each candidate that affect their reach, so it would not be accurate to compare reach across candidates. 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 may have a further reach than Yang's tweet. Other confounding variables can be the demographic that follows the candidate, the campaign strategies, how popular they are, how long they've been with the party, etc. Creating the model per candidate allows us to specifically see the impact a tweet's sentiment has on its reach and control for confounding variables.

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 were combined per candidate, ultimately creating 4 feature matrices to be plugged into the model.

We initially were unsure about which specific model to use, so we evaluated multiple models using the same features and performed classification using these models. The models we used were logistic regression, Support Vector Machine(SVM), and Keras Deep Learning. We initially were using 3 labels and classifying tweets as high, medium and low reach, however the model accuracy did not go above 44% when doing so. As a result we decided to use only 2 labels, high and low reach, and conduct binary classification instead.

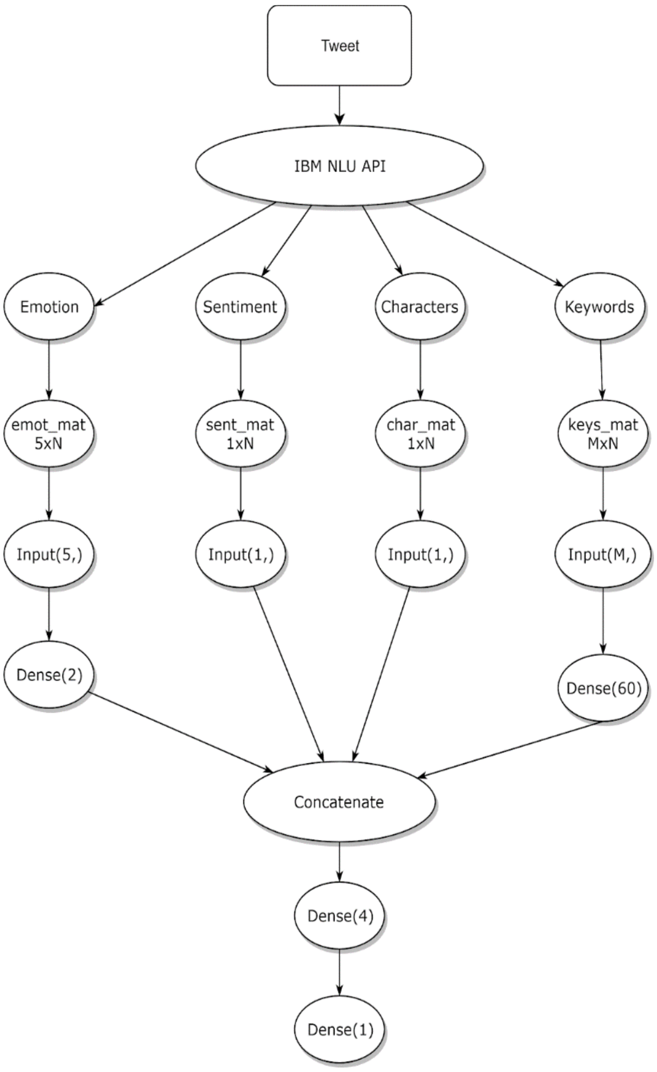


Figure 2: Keras Model Diagram

For each model, we just used the sentiment and emotion without corpus as a baseline, meaning that keywords and the subject matter of the tweets were ignored to simply find the direct effect sentiment and emotion had on reach. We then used sentiment and emotion with corpus and we found that the accuracy of the model significantly increased. This is because certain keywords significantly impacted a tweet's reach, for example, mentioning a politician like "Donald Trump" in a tweet by Warren had a bigger coefficient in the logistic regression model than any of the sentiment or emotion analysis alone. We thought that scaling the inputs of the model would give a more balanced result, but after applying standard scaling to the inputs we found no difference than without scaling. We ultimately decided on using the sentiments and emotions with corpus and without scaling as inputs for our models. The results are summarized in the tables below.

5. Results

Initially, we tested each classifier using three different methods: without corpus, including corpus, and using standard scaler (Table 1). Including corpus increased the accuracy by 9% for logistic regression and 7% for support vector machines. However, scaling the inputs did not increase the model score. The best results for the Warren dataset were using the keywords as a feature without scaling any inputs.

	Binary Label (no corpus)	With Corpus	Scaled
Linear Regression	0.59	0.68	0.66
SVM	0.58	0.65	0.64
Keras		0.25	

Table 1: Deciding For Classifier (Test Accuracy)

After training the Keras deep learning model, we reached 76% train accuracy for Warren, however, the test accuracy did not increase beyond 25%, so we abandoned the model. The accuracy results are seen in Table 2. SVM with linear kernel gave a good f1 score for most candidates ranging between 0.64 to 0.78 on test data. However, the best results were taken from the logistic regression model which was better for every candidate. The logistic regression model gave us an f1 score of 98% on the train data and 66% on the test data for Warren; 94% on the train data 77% on the test data for Biden; 97% on the train data and 70% on the test data for Bernie; and 92% on train data, 80% on the test data for Yang.

Candidate / F1 Score	Logistic Regression		SVM		Custom Keras Model	
	Train	Test	Train	Test	Train	Test
Warren	0.98	0.66	0.99	0.65	0.76	0.25
Biden	0.94	0.77	0.98	0.69		
Bernie	0.97	0.70	0.99	0.64		
Yang	0.92	0.80	0.97	0.78		

Table 2: F1 Scores for Each Classifier

	Sad	Joy	Fear	Disgust	Anger	Sent.	Num. Chars	Most Effective Keywords
EW	0.007	0.463	0.791	**1.602**	**1.22**	-0.37	0.134	@realDonaldTrump Brett Kavanaugh
JB	0.311	0.883	0.437	0.291	0.32	-0.32	**1.982**	Hillary, America, Donald Trump, Families
BS	0.383	**-1.316**	0.062	0.565	0.962	-0.004	**-1.307**	Child, Dollars, Republicans
AY	0.607	-0.891	0.015	**2.222**	-0.480	**-1.044**	**4.434**	Americans, Thanks, Time

Table 3: Coefficients for each feature

When investigating the coefficients of the classifiers, we found the keywords and features differ in importance for each candidate(table 3). For Warren the most important factor is showing disgust. Mentioning political and public figures like @realDonaldTrump also increases the reach of her tweets, which makes intuitive sense as he is the incumbent President and the challengers need to show why they would make a better President than him.

Tweet length is another important factor as well. For Biden and Yang, the number of characters play a direct role on number of reach, while for Bernie it is the opposite.

Our analysis showed that sentiment scores are inversely proportional to the reach, meaning that the more negative the tweet, the higher number of retweet counts. However, this correlation is not as high as expected.

6. Conclusion

An interesting thing to note is the keyword analysis for Biden. Hillary is one of his most effective keywords, and this initially seems strange as she is not participating in the 2020 election. However, Biden is a low volume Tweeter with only about 3,200 tweets as of November 2019. In our data collection step, we ignored the 1,000 most recent tweets to ensure the retweet and like numbers did not fluctuate while we conducted the analysis. Due to Biden's low number of tweets and Hillary being an effective keyword, it is highly likely that his tweets from the 2016 election cycle were included in our model.

If we were to expand upon this research, we would do much more with time data from the tweets. We would make sure to collect tweets only after the candidate officially announced their candidacy for the 2020 election. In addition, we would refine the cut-off of recent tweets so it is based off of the time instead of the number of the tweet. In this project, we ignored the 1,000 most recent tweets but it would have been better to filter out tweets from the past 2 weeks.

In addition, adding features for images or videos would help make the model more accurate, as a video or image may be the reason why a tweet is high reach, but our analysis would falsely label the text as having high reach.

Overall, we found that positive sentiment had a negative coefficient for our logistic regression model. This means that tweets with positive sentiment values have a lower probability of becoming high reach than those in the reference group. In addition, tweets interpreted to exhibit disgust, fear, and sadness had a positive coefficient across the board, meaning that they were more likely to be high reach. However the coefficient values were not that high, so while there is an indication that negative messaging leads to higher reach, we cannot definitively conclude negative messaging causes higher reach.

7. 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.