Course Project: Final Report

Project: Evolution of Donald Trump Twitter Sentiment: 2016 to 2020 Presidential Election
Author: Lee Waters

Abstract

This document contains the *Final Report* for the course project in the Fall 2020 offering of Data Foundations at The University of Texas at San Antonio. This study is the result of a series of preceding steps, including: topic generation; peer and instructor feedback; *Final Proposal;* and a *Midterm Report* that was submitted previously. This report details the author's completed progress including: 1) a description of the data set; 2) final experimental results; 3) validation of experimental results; 4) a discussion of the study's findings; 5) and a conclusion confirming the completion of final progress checks outlined in the *Final Proposal*.

1 Description of data set

There are two data sets used in this study – one for each respective presidential election. The 2016 United States Presidential Election was downloaded from the George Washing University Libraries Dataverse at Harvard Dataverse (Littman et al., 2016). The Harvard Dataverse is a free data repository available to all researchers, including those not affiliated with the university. Within the Harvard site, there exist individual *dataverses* constituting a virtual repository for organizing, managing and showcasing data sets.

The George Washington Libraries Dataverse contains the *2016 United States Presidential Tweet IDs*, which is used in this study. The dataset contains the tweet ids of approximately 280 million tweets related to the 2016 United States Presidential Election. They were collected between July 13 and November 10, 2016 by the GWU research team. This is the four-month period prior to the 2016 United States Presidential Election, which took place on November 8, 2016.

The tweet ids from this dataset are broken up into 12 collections, including: candidates and key election hashtags; Democratic candidates; Democratic Convention; Democratic Party; election day; first presidential debate; GOP Convention; Republican candidates; Republican Party; second presidential debate; third presidential debate; and vice presidential debate. As per Twitter's Developer Policy, tweet ids may be shared publicly but tweets may not. Therefore, the actual tweets are "hydrated" from the corresponding tweet ids. The resulting dataset consists of a CSV file containing 647,648 entries. The attributes of this dataset are date of post, username, location tag, and tweet text.

A second source is used for the tweets related to the 2020 Presidential Election. The data comes from the IEEE DataPort. The platform offers free uploads of any dataset up to 2TB. This allows users to retain and manage their valuable research data, while also permitting free access to other datasets on the platform. The IEEE DataPort currently has over 425,000 global users and over 1,500 datasets.

The actual dataset is the *US November 2020 Election Tweets Dataset* (Ibrahim, 2020). The dataset includes tweets from the 2020 Presidential Election that contain #USAelection or at least one of the following keywords relating to the four candidate parties.

<u>Keywords about the Democratic Party</u>

- @DNC
- @TheDemocrats

- @Biden
- @JoeBiden
- "Our best days still lie ahead"
- "No Malarkey!"

Keywords about the Green Party

- @GreenPartyUS
- @TheGreenParty
- "Howie Hawkins"
- @HowieHawkins
- "Angela Walker"
- @AngelaNWalker

Keywords about the Libertarian Party

- @LPNational
- "Jo Jorgersen"
- @Jorgensen4POTUS
- "Spike Cohen"
- @RealSpikeCohen

Keywords about Republican Party

- #MAGA2020
- #NovemberElection
- @GOP
- @Trump
- @POTUS
- @realDonaldTrump
- @Pence
- @Mike_Pence
- @VP
- "Keep America Great"

This dataset contains three million tweets from July 1 to August 12, 2020. The 2020 US Presidential Election occurred on November 3. This dataset does not capture tweets occurring after the first presidential debate which occurred on September 29, 2020. This differs from the dataset used for the 2016 Presidential Election.

The 2020 dataset consists of six attributes, including: date of creation; from user id; to user id; language; retweet count; and tweet id. The dataset is restructured and includes 3,053,897 tweets containing the date, user id, location tag, and tweet content.

2 Final experimental results

Both datasets are successfully loaded into a JupyterLab Notebook as part of the Anaconda distribution. The objectives of the study are to 1) compare the public perception of Donald Trump before the 2016 and 2020 US Presidential Elections using 2) lexicon and machine learning sentiment analyses. The methods employed use techniques learned throughout Data Foundations class.

2.1 Lexicon sentiment analysis

A *LexiconClassifier* class is created to assist with sentiment annotation. The class has several constituent functions. These functions predict whether a given tweet is positive, negative or neutral by counting the positive and negative words in a tweet. The language is "scored" from the files *positive_words.txt* and *negative_words.txt,* which contain manually curated positive (e.g. good, great, awesome) and negative words (e.g., bad, hate, terrible). These files were first

introduced in Data Foundations *Problem Set 2* and are also used in this analysis.

The defined class is used to perform sentiment analysis on the 2016 and 2020 election tweets. These results differ somewhat from the *Midterm Report*, because the first 100 tweets are removed and used for validation. A number of metrics are calculated for each data set. The results are reproduced below:

2016 Analysis

- Total Number of Tweets: 647,548
- Total Number of Positive Tweets: 271,978
- Total Number of Negative Tweets: 109,701
- Total Number of Neutral Tweets: 265,869
- Most Positive Tweet:

  @Ted_Strickland TRUMP NOW Trump Forever Trump Now Trump Forever Trump Now Trump Forever Trump Now Trump Forever

- Most Positive Tweet Score: 8
- Most Negative Tweet:

  @cnni @cnnbrk @CNN LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS LIARS!! https://t.co/jrnPHFET6T

- Most Negative Tweet Score: -15
- Total Number of Users: 257,204
- Average Number of Tweets per User: 2.52
- User with the Most Tweets: PolitickDick (1,001)

2020 Analysis

- Total Number of Tweets: 3,053,797
- Total Number of Positive Tweets: 1,176,515
- Total Number of Negative Tweets: 623,914
- Total Number of Neutral Tweets: 1,253,368
- Most Positive Tweet:

  @GOP GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION GOOD GREAT RIGHT DECISION

- Most Positive Tweet Score: 30
- Most Negative Tweet:

  @realDonaldTrump @OANN Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie Lie

- Most Negative Tweet Score: -70
- Total Number of Users: 1,005,918
- Average Number of Tweets per User: 3.04
- User with the Most Tweets: tsartbot (444)

2.2 Machine learning sentiment analysis

A sentiment analysis is then performed for the same group of tweets for years 2016 and 2020 using machine learning techniques. The first step is to split the data into training and test sets. The first 100 tweets are used for this purpose; of which, 80 are used for the training dataset and 20 are used for the testing dataset. These 100 tweets are manually annotated for sentiment to train and test the model.

Example Tweet & Annotation

- Tweet 1: RT @HuffPostBiz: Trump's new economic plan is short on specifics, other than tax help for wealthy https://t.co/vGefdXaiDP https://t.co/Obh9â□¦
- Manual Annotation: Negative

Feature engineering is then used to convert the text to a set of representative numerical values. Scikit-Learn's CountVectorizer is used for this process and the vectorization is performed with unigrams (Scikit-learn, 2011).

A linear SVC machine learning model is then trained on the data. Scikit-Learn's GridSearchCV is used to search for the best model with regularization parameters of 0.0001, 0.001, 0.01, 0.1, 1, 10, and 100; Micro F1 scoring, and five cross-validation folds

This model is then used to predict sentiment annotations for the remaining tweets. These results are reproduced below:

2016 Analysis

- Total Number of Tweets: 647,548
- Total Number of Positive Tweets: 198,030
- Total Number of Negative Tweets: 231,600
- Total Number of Neutral Tweets: 217,918

2020 Analysis

- Total Number of Tweets: 3,053,797
- Total Number of Positive Tweets: 282,857
- Total Number of Negative Tweets: 2,729,178
- Total Number of Neutral Tweets: 41,762

2.3 Summary of results

The results from the lexicon and sentiment analyses are reproduced in *Table 1* and *Table 2*. Net Tweets is the difference of Positive Tweets and Negative Tweets. Sentiment Score is the proportion of Net Tweets and Total Tweets.

There are several findings that can be inferred from these results. First, the lexicon analysis shows a positive sentiment for Trump in 2016 (25%) and 2020 (18%), whereas the machine learning analysis shows a negative sentiment for Trump in 2016 (-5.2%) and 2020 (-80%). The second result is that sentiment of Trump decreased from 2016 to 2020 for both lexicon (-7%) and machine learning analyses (-75%). Finally, machine learning saw a larger decrease in sentiment between 2016 and 2020.

Validation of experimental results and the respective accuracy of each approach will be discussed in the next section.

	Lexicon	*Machine Learning*
Net Tweets	162,277	-33,570
Total Tweets	647,548	647,548
Sentiment Score	25%	-5.2%

Table 1: Trump sentiment 2016

	Lexicon	*Machine Learning*
Net Tweets	552,601	-2,446,321
Total Tweets	3,053,797	3,053,797
Sentiment Score	18%	-80%

Table 2: Trump sentiment 2020

3 Validation of experimental results

The validation of experimental results is different for each sentiment-based approach. For the lexicon analysis, the first 100 tweets are annotated for both the 2016 and 2020 datasets. The accuracy of this method is evaluated as the proportion of equivalent lexicon sentiment predictions and manual annotations by the author. The accuracy is 33% in 2016 and 27% in 2020.

	2016	2020
Accuracy	33%	27%

Table 3: Lexicon validation results

The same manual annotations are used for the validation of the machine learning results. The data is trained on 80 training samples and manual annotations and then tested on the remaining 20. A Micro F1 validation score is calculated for the 2016 and 2020 models' performance on the training data set. The Micro F1 score is 52.5% in 2016 and 70% in 2020. Accuracy is then calculated for each model's performance on the test data set. The accuracy is 65% in 2016 and 50% in 2020.

It is unexpected that the accuracy of the test set (65%) exceeds the Micro F1 score (52.5%) in 2016. This is most likely due to too small of a training set for the machine learning model. This is one opportunity for future research to improve this study's design.

	2016	2020
Micro F1	52.50%	70%
Accuracy	65%	50%

Table 4: Machine learning validation results

When comparing the performance of the lexicon and machine learning approaches for sentiment analysis, the machine learning model outperformed its counterpart for both election years. One difficulty with the lexicon-based approach is that it only counts positive and negative words and is unable to detect sarcasm. Most news also has negative bias, so the retweeting of news events can be labeled as having negative sentiment even though the author's purpose is purely objective. These are just some examples of the issues with lexicon-based methods for sentiment analysis and indicate why the machine learning approach may have performed better for the purposes of this study.

4 Discussion

The objectives of the study are to 1) compare the public perception of Donald Trump before the 2016 and 2020 US Presidential Elections using 2) lexicon and machine learning sentiment analyses.

The results show that 1) public perception of Donald Trump decreased from 2016 to 2020 with both methods. However, the lexicon-based sentiment analysis rated public perception of Donald Trump as positive before both elections, which contradicted the results of the machine learning analysis.

During the validation of the experimental results, 2) the machine learning approach is determined to be the more accurate of the sentiment prediction methods.

The overall findings of this study are that machine learning is the more effective approach for sentiment analysis of political tweets, the public perception of Donald Trump was negative preceding the 2016 and 2020 elections, and public opinion of the president decreased during that time.

5 Conclusion and final progress checks

The *Final Proposal* also included final "exam" checks to be completed by project submission. This included additional lexicon-based analyses, polishing the code, and generating a final report and presentation. The additional lexicon-based analyses were 1) hate speech and offensive language and 2) flagging various health markers. The author planned to use lexicons that were introduced in the extra credit portion of *Problem Set 2*.

However, after instructor feedback from the *Final Proposal*, the author decided to use machine learning to predict sentiment instead of conducting other lexicon-based analyses. The author would then compare the results from the sentiment and machine learning approaches. This decision was based on the instructor's recommendation and a preference to dive deeper into one analysis (sentiment) with different techniques (lexicon-based and machine learning) rather than conduct multiple analyses (sentiment, hate speech, mental health indicators) with one technique (lexicon-based).

The machine learning approach was not considered previously, because this technique was not introduced until *Problem Set 3* after the submission of the *Final Proposal*. The author is very interested in machine learning, which is why he elected to explore this concept further.

All of these final progress checks are completed culminating in the conclusion of this project.

Acknowledgments

Thank you to Dr. Anthony Rios of The University of Texas at San Antonio for his assistance in gathering the data and his guidance throughout this study.

References

Ibrahim Sabuncu, August 14, 2020, "US November 2020 Election Tweets Dataset", IEEE Dataport, doi: https://dx.doi.org/10.21227/25te-j338.

Littman, Justin; Wrubel, Laura; Kerchner, Daniel, 2016, *"2016 United States Presidential Election Tweet Ids"*, https://doi.org/10.7910/DVN/PDI7IN, Harvard Dataverse, V3

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.