DATA 340 Natural Language Processing
Final Paper - Professor Tucker
Josh Rogers
May 16, 2023

Using Natural Language Processing for Analysis of Politicians' Tweets

Abstract

This project centers around tweets from U.S. politicians. The data set used for this project was created by Kyle Pastor in 2018 and consists of approximately 84,000 tweets. This study sets out to answer several questions: whether one political party is more rhetorically positive or negative on social media than the other; whether, based on previous tweets, we can create a model that predicts a politician's political party affiliation; and finally, whether a model of this nature performs better when only using nouns and adjectives as opposed to multiple forms of speech that are preprocessed using tokenization, lemmatization, and stopword and punctuation removal. As someone with a data science and government background, I was surprised to see how little research has been done in this field. Other than research conducted during the 2016 Presidential election, there has not been much done since. However, communication is critical in a congressional office, and analyzing tweets from politicians provides a closer look into their psyche and strategy. For the sentiment analysis of politicians, I used nltk's VADER tool to assign positive, neutral, or negative sentiments to each tweet. From there, I would create a graph showcasing the results, as well as a breakdown of percentages. Other than the preprocessing steps (tokenization, lemmatization, stopword removal, and punctuation removal), the methodology was the same for the model building. First, I used TfidfVectorizer to convert the words to vectors, then used logistic regression to create a model. From there, I produced a classification report and confusion matrix of the model.

The sentiment analysis showed that generally, Republican rhetoric skewed positive, whereas Democrat rhetoric skewed negative. Since Republicans had a unified government in

2018, it is possible that Republicans attempted to promote a more positive outlook on social media. In contrast, Democrats were more likely to use negative language, most likely to motivate their base to vote in the 2018 midterm elections. I wouldn't be comfortable asserting this without further research, but a similar result with the parties' positivity and negativity reversed could be drawn from a dataset from 2021. Both of the models had reasonable accuracy rates, with the noun and adjective model only returning a 68% accuracy rate and the traditionally preprocessed tweets returning a 75% accuracy rate. Unfortunately, I could only use a subset of the data due to memory issues, which caused entire sessions to crash. However, given the opportunity to work with more data, the classification results could be even higher.

Introduction

Over the last decade, politicians have increasingly utilized Twitter as a primary source of communication with constituents and to express their views. As someone studying Government and Data Science in college, I wanted to combine these two and research statistics using politicians' tweets. I had two primary questions: is one party more rhetorically negative than the other; and can someone build a model that predicts whether a tweet is from a Democrat or a Republican? The first question is important because analyzing tweets' sentiments could indicate whether one party is more responsible than the other for the negative perception most people have of the current political landscape in the U.S., and the second question is important because there are several ways that this type of analysis can be utilized. With more development, this model could identify if certain politicians are lean a certain way on a political spectrum. The model could help track the overton window shift over time and could also help analyze critical topics that politicians debate. This can be useful for campaigns, journalists, and political scientists. Holistically, this research aims to analyze a growing form of communication among politicians over the last decade.

Literature Review

The first piece of literature I read was by Andrzej Szymanski, a data scientist from Heidelberg, Germany. He ran a project centered around the British media's political sentiment. Sentiment analysis is vital in both of our projects. His research is also why this research utilized VADER (Valence Aware Dictionary and sEntiment Reasoner). As Szymanski stated, VADER differentiates from other sentiment programs: "It evaluates sentiment based not only on words themselves but also takes into account case, punctuation, and emoticons."[1] While our project goals and what we are analyzing differ slightly, how he utilized VADER provided a solid foundation for my work.

Another article I analyzed was "Using logistic regression method to classify tweets into the selected topics."[2] While our topics varied heavily (they had hundreds of different classifications while mine is only binary), it provided a framework for proceeding: preprocessing; text feature extraction; and machine learning. In terms of preprocessing, they utilized tools that individuals with knowledge of Natural Language Processing are already familiar with, such as tokenization, stopword removal, and punctuation removal. Additionally, they utilized a bag of words, another familiar tool. Their model produced a 95% accuracy rate, which is highly successful. Overall, both readings provided an excellent framework to carry out what was needed.

Methodology & Dataset

Collecting data was the first step in this process, which proved to create several challenges. The initial dataset, which I created by collecting tweets using snscrape, contained approximately 18000 tweets from several Republican senators. However, more data needed to be collected to successfully complete this project. Unfortunately, Elon Musk altered Twitter's API so that one would have to pay in order to scrape tweets. I had no success scraping tweets even

---

[1] Andrzej Szymański. "*Political sentiment of British media in 2019*".
[2] S.T. Indra et. al. "*Using logistic regression method to classify tweets into the selected topics*".

after experimenting with alternative options such as GraphQL and Osome. This created a significant roadblock in my research. Fortunately, though, I found a dataset on Kaggle with over 80,000 tweets from both Democratic and Republican Congressmen from 2018.[3] The dataset consisted of three columns, the Politician's handle (e.g., RepDarrenSoto), the politicians' party (Democrat or Republican), and the tweet they sent out.

For each question, there was a different methodology. For the sentiment-party relationship, the pipeline is simple. After loading in the VADER Sentiment analyzer, I created a definition called "analyze_sentiment." The function applies the sentiment analyzer to each tweet and then extracts the sentiment polarity score. Then I used an if-else loop that classified the tweet as either positive, negative or neutral. After this, I produced a graph that showcased the results.

I wanted to experiment with two methods for building the model. For the first one, I would preprocess with traditional NLP tools, such as tokenization, lemmatization, stop word removal, and punctuation removal. After this, I would use TfidfVectorizer as a text feature extraction method. I then used the TfidfVectorizers fit_transform function to transform the preprocessed tweets, then convert that into an array, and set that as the X variable. Then, I would set the tweet user's party as the y variable. Then, using train_test_split, I split the data into a training and validation set. Finally, I fit the model to retrieve the accuracy score.

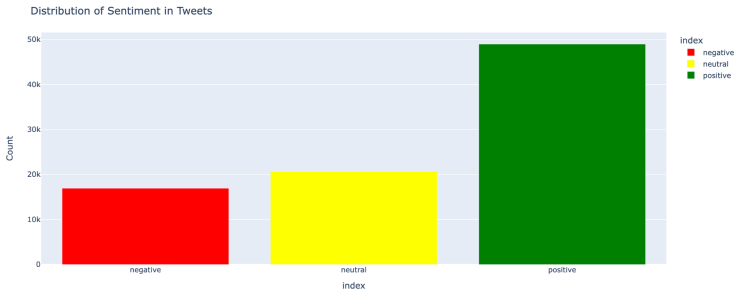
After this, I wanted to experiment with analyzing only the nouns and adjectives in a tweet. I hypothesized that specific nouns and adjectives that politicians would use (e.g Republicans might use “border,” whereas Democrats may use “gun control”) would differ more than just having all words in a tweet and, in turn, produce a better model. I repeated the steps with the generic logistic regression model, except for Spacy, which I only used to pull the nouns and adjectives in each tweet.

---

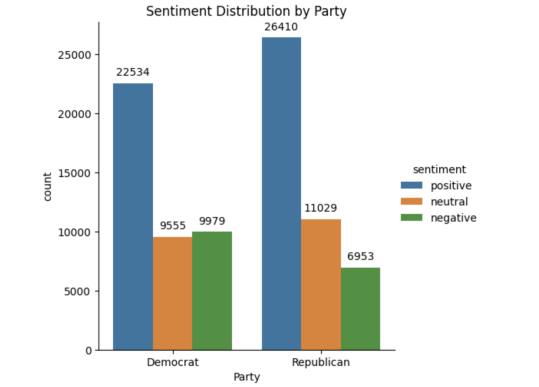
[3] Kyle Pastor. “*Democrat Vs. Republican Tweets*”.

Results

After running the VADER function on the dataset, it determined that there were 48,944 positive tweets, 20,584 neutral tweets, and 16,932 negative tweets. This makes sense because both parties wanted to energize their voter bases in the election year.



The results also showed that Republicans were more likely to post positive tweets, whereas Democrats were more likely to post negative tweets. This makes sense since, in 2018, Donald Trump was the sitting President, and Republicans had a unified government, with control over the White House, Senate, House of Representatives, and Supreme Court.



Sentiment Distribution Percentages by Party

	Positive	**Neutral**	**Negative**
Democrats	53.57%	22.71%	23.72%
Republicans	59.49%	24.85%	15.66%

Following the sentiment analysis, I decided to run a logistic regression on the extracted nouns and adjectives of the tweets. First, I ran a TF-IDF on each party to see if there were any key differences. While they were not as strong as I anticipated, they were there. One term that appeared on the Republicans' term frequency list much higher than on the Democrat's list was "tax." This was notable since taxation was a significant point of Republican policy in 2018. Conversely, some words that appeared on the Democrats' list that were not on the Republicans' were "students," "women," "families," and "health."

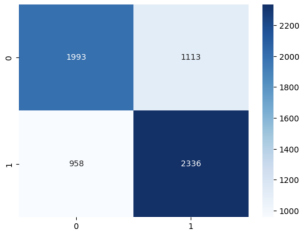
Republicans (left) and Democrats (right) Term Frequency Lists

	term	term_frequency
3	#	4006
30391	today	2151
7483	Today	1591
27168	people	1360
30362	time	1273
29839	students	1233
9028	bill	1182
26217	more	1164
31535	women	1058
12136	families	1019
12870	great	970
30123	tax	910
13087	health	901
10438	country	883
31395	week	865

	term	term_frequency
3	#	4006
30391	today	2151
7483	Today	1591
27168	people	1360
30362	time	1273
29839	students	1233
9028	bill	1182
26217	more	1164
31535	women	1058
12136	families	1019
12870	great	970
30123	tax	910
13087	health	901
10438	country	883
31395	week	865

A common problem with both logistic regression models was the memory for both the train_test_splits. Even after using a generator, the system continued to crash from RAM issues. Therefore, I had to use a smaller sample size. In the future, this is something I would like to work on more. Despite this, I successfully created a model with a 68% accuracy rate.

	precision	recall	f1-score	support
Democrat	0.68	0.64	0.66	3106
Republican	0.68	0.71	0.69	3294
accuracy			0.68	6400
macro avg	0.68	0.68	0.68	6400
weighted avg	0.68	0.68	0.68	6400



I repeated a similar process with the entire dataset. After running a TF-IDF on both parties again, the most notable difference is that Democrats tweeted about Trump significantly more than Republicans. In journalism, it is a common practice to direct attention to the opposition (e.g Fox News writes more articles about Joe Biden in 2022, and CNN writes more articles about Trump in 2018).

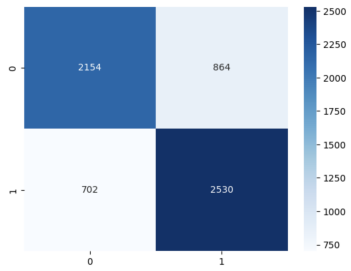
Republicans (left) and Democrats (right) Term Frequency Lists

	term	term_frequency
63955	rt	9076
12687	amp	4701
67182	today	3848
66892	thank	2687
67592	trump	2489
69536	work	2341
56170	join	1930
69739	year	1829
57551	m	1788
21650	great	1784

	term	term_frequency
64933	rt	9993
68017	today	4870
12919	amp	4569
67730	thank	3709
21201	great	3012
67454	tax	2575
14219	bill	2377
22250	house	2366
70106	work	2309
15		2064

After running this model, it returned a 75% accuracy rate. This made the hypothesis null. However, I was happy with the accuracy rate and believe that if I could have solved the memory problem, the accuracy rate would have been even higher.

	precision	recall	f1-score	support
Democrat	0.75	0.71	0.73	3018
Republican	0.75	0.78	0.76	3232
accuracy			0.75	6250
macro avg	0.75	0.75	0.75	6250
weighted avg	0.75	0.75	0.75	6250



Discussion

The results of the analysis of politicians' sentiments made sense. It is important to note that this dataset was from 2018, when Republicans held the Presidency and both chambers of the United States Congress. Those with power are often happy, and those in the opposition are often not. If one were to rerun this procedure using a dataset from 2022, I would hypothesize that Democrats' tweets would be more positive and Republicans' tweets would be more negative. In addition, it would be interesting to utilize a dataset during a midterm year where one party controls the U.S. The House of Representatives and the other control the Senate. If we were to analyze data from 2024 a few years down the line, Democrats would still come out as the more positive party, as they currently control the Senate and, more importantly, the Presidency.

After conducting research, while there has been a large amount of political sentiment analysis, especially on the 2016 election, I could not find any articles that conducted a similar study to mine. This study is also different from the 2016 Presidential election, as that race was for an open presidential seat. If someone were to run a similar experiment to mine, they would come away with a similar conclusion, as the analysis results were not surprising. Since Twitter's API has changed, it is hard to analyze this data. However, I believe that this analysis shows that those in power (Republicans in 2018) tried to create a more positive outlook on the country's standing, whereas Democrats (the opposition) used more negativity to rile supporters up and motivate them for the upcoming midterms.

Utilizing nouns and adjectives was less successful than I anticipated. After analyzing each party's TF-IDF, I believed that since words such as "border," "students," and "women" were used more by one party, it would produce a better model. However, I found an article from Bo Pang and Lillian Lee stating that more types of language produced had a higher accuracy rate than models that only used a specific type of speech.[4] Following the results, it makes sense

[4] Bo Pang and Lillian Lee. "*Opinion mining and sentiment analysis*".

that offering a model with more text forms would produce a more accurate result. That being said, the model's 68% accuracy rate was satisfactory for purposes of this paper. However, similar to the other logistic regression model, the results would have been even more accurate if I had access to more memory and data for this project.

I was happy with the outcome regarding the primary logistic regression model. Despite the smaller sample size needed because of the memory issue, I was happy that this model had a 75% accuracy rate. Like the point above, the model could have been even more accurate with a more robust RAM system. While similar studies have been conducted using GRU models and other methods, this was the first one I could find that utilized Logistic Regression specifically. In the future, I would like to build more on this by experimenting with other models, such as neural networks or decision trees. Overall, though, I was content with this result.

Building a model like this can be helpful in several ways. First, it can be used to identify current political trends in parties. If one were to split the classifications up even more (progressive vs. moderate Democrat), it could help hone the classification of politicians' ideologies. Additionally, this can be used as a tool to track trends in both parties and analyze how they are responding to a given issue. Additionally, this tool can be used in a variety of fields. As someone who previously worked on campaigns, having a tool to make this classification would have been helpful for opposition research. While this research was more proof of concept, I have shown this model's potential with further development.

It should be noted that the future of this type of research is in jeopardy. Given Elon Musk's recent changes to Twitter's API, retrieving tweets for data has become almost inaccessible. While there may be continued development of tools that can scrape Twitter in the coming months, as of writing this paper, all previous tools capable of scraping Twitter must be functional, given the changes. While a private company has a right to restrict what is accessed, this puts a hold on necessary research. Social media has become a cornerstone of politics in multiple fields. Restricting academic research on sight halts important developments in this field

and could potentially pose a threat to the free exchange of information. While it is sad and a little unnerving that this work has been temporarily stopped, I am confident there will be a way to circumvent this roadblock in the coming months and years.

Conclusion

In terms of the sentiment of politicians' tweets, while further research is required, based on the data, an opposition party is more likely to tweet negatively. In contrast, a party in power is more likely to post positive messages. If Twitter's API becomes usable again, I would like to continue this research and see what a split congress's results show. Additionally, a politician's party can indeed be identified by utilizing their social media posts. Since the model was successful, I would like to move on to splitting the party classifications into more niche groups and make a model that can classify a politician more specifically. While Twitter's API now having limitations is a setback for academic research, we will hopefully develop more ways to retrieve this information.

References:

1. Andrzej Szymański. "*Political sentiment of British media in 2019*".
2. S.T. Indra et. al. "*Using logistic regression method to classify tweets into the selected topics*".
3. Kyle Pastor. "*Democrat Vs. Republican Tweets*".
4. Bo Pang and Lillian Lee. "*Opinion mining and sentiment analysis*".