US Politics by Social Media

Project Checkpoint Report

Rachel Ng Min Yee
University of Colorado, Boulder
Rachel.Ng@colorado.edu

Xinyi Lu
University of Colorado, Boulder
xilu2783@colorado.edu

Anuragini Sinha
University of Colorado, Boulder
ansi3987@colorado.edu

1 Introduction

Social media platforms are widely used by citizens during the US election period to express their opinions. This project examines the sentiments of Twitter users towards the different political parties and how Tweets can affect the outcome of the US presidential elections. With the use of natural language processing techniques such as keyword extraction, text classification, topic modeling and opinion analysis, we aim to extract key information from social media data and study the political behavior of Twitter users. This will be helpful for election candidates to assess their odds of success based on previous case studies and to help them tailor their campaigns based on the sentiments on the ground.

2 Related Work

Multiple projects have explored the sentiments of tweets related to the US election. Ali et al. proved that removed tweets posted after the 2020 US Election Day sided with Joe Biden while those before Election Day were more favorable about Donald Trump. This was done by performing sentiment analysis on tweets that were subsequently removed from Twitter and comparing their results to accessible tweets and accounts across time [1]. In another research project, Caetano et al. discovered that political homophily level rises when there are close connections and similar speeches. This was accomplished by analyzing the tweets of users during the 2016 US elections and classifying sentiments into six groups: Trump supporter, Hillary supporter, whatever, positive, neutral and negative [2].

Finally, Ansari et al. found out the dominance of support for a single party on Twitter for the 2019 General Elections of India by using the Long Short Term Memory (LSTM) classification model to predict the sentiments and results of the elections. Ansari et al. also compares the results from the LSTM model with other machine learning models [3].

Our approach is different from past projects as we aim to perform opinion analysis using natural language processing techniques to answer key questions. Some questions we can possibly answer are: which groups of people have these opinions? How do opinions change over time? Subsequently, with insights from the data, we can find out the possible factors that led to Trump's large voter share despite the controversies surrounding him in the 2020 US Presidential Election.

3 Proposed Work

3.1 Dataset

The dataset we chose to work on is a collection of tweets containing #DonaldTrump, #Trump, #JoeBiden, #Biden hashtags from the 2020 US presidential election [4]. There are two subcollections in the dataset - one for the tweets related to Donald Trump and one for Joe Biden. With 970,919 and 776,886 rows respectively, there is a total of approximately 1.72 million rows of data in our dataset. Of all the tweets, 394,400 or roughly 20% of them were tweets from the US.

The data set has a total of 21 columns and a Data Dictionary describing the names and definitions of the columns of the dataset can be found in the appendix. The specific columns that we found interesting were columns created_at, tweet, likes, reteweet_count, user_followers_count, user_location, lat, long, city, country, continent, and state_code.

3.2 Subtasks

With the above dataset, we hope to perform text classification, opinion analysis as well as a case study on Donald Trump and his relatively successful voter share despite his abrasive personality and the serious controversies surrounding him. These analyses can be further split into the following subtasks:

Analysis	Subtasks
Text Classification	Name Entity Recognition • Which parties are involved in this tweet? • Which candidates are involved in this tweet? Keyword Extraction • Which keywords are the most important/relevant in the tweet? Sentiment Analysis • Is the tweet a positive, negative or neutral one?

	• What identity is the tweet supporting / criticizing? Topic Modeling • Grouping tweets that share common topics • Grouping tweets that share the same sentiment • Which topics were most discussed? • What were the topics that supporters of each party cared the most about? • Which party's supporters were more vocal about their opinions? • Which party's supporters generally had the bigger following on Twitter?
Opinion Analysis	Temporal Analysis • What are the predominant opinions over time? Categorization of Opinions according to: • State • Country • In the US vs outside of the US • Democratic vs Republican
Case Study on Donald Trump	• How did opinions change over time? • Did the timeline coincide with certain events? • How did Twitter specifically help him/prevent him from swinging favor? • What factors helped him garner his large voter share?

We believe the work to be sufficient and feasible for our group size of 3 at the undergraduate level as we are approaching the topic of the 2020 US presidential election from multiple different angles. If time permits, we could possibly look into adding other datasets and comparing the 2016 election to the 2020 election that we are currently analyzing.

4 Evaluation

Once we have completed our analysis for the project, we plan on evaluating the results by using the following methods:

- For the correlations we find, we plan to further validate them against Tweets made during the 2016 US presidential elections, if possible.
- We will calculate the correlation coefficient for the correlations to measure their strength.
- If a strong correlation is found, then we will create a correlation model with the attributes that are found to be useful from our research.
- We also plan to analyze baseline performances.

5 Milestones

Task	**Start Date**	**End Date**
Data Preprocessing Text mining, natural language processing (text classification, named entity recognition, topic modeling)	10/6	11/2
Data Analysis Opinion analysis (categorization, temporal)	11/3	11/23
Data Visualization Charts (bar plot, line charts, scatter mapbox, choropleth map)	11/24	11/30
Evaluation Evaluation of metrics, documentation (report writing, presentation slides)	12/1	12/8

6 Tools

We are using Python for data pre-processing and model building. Some packages that we are using include: pandas, numpy, langdetect, re, matplotlib, seaborn, plotly, ast, wordcloud, nltk, textblob, flair and vader.

7 Completed Work

7.1 Data pre-processing

Data cleaning is the process of identifying and correcting errors, duplicated and incomplete data from a dataset. To start off, date columns (ie. ‘created_at’, ‘user_join_date’, ‘collected_at’) are converted to datetime types and numeric columns (ie. ‘tweet_id’, ‘likes’, ‘retweet_count’, ‘user_id’, ‘user_followers_count’) are converted to integer types. Next, rows that have NA in the ‘country’ columns are dropped as we intend to analyze tweets by country. We also standardized country names by converting ‘United States of America’ to ‘United States’.

Following this, we conducted text cleaning on the ‘tweet’ column of the dataset. Using the package langdetect, we detected the language of the tweets and filtered for English tweets to facilitate sentiment analysis and topic modeling of tweets. Next, using the package re, punctuation and numbers are removed, and tweets are converted to tokens. As stop words are not useful for our analysis, we removed stop words as well. Finally, we converted tokenized words to their base form using stemming and lemmatization functions from the nltk library.

7.2 Sampling

Since our dataset is too huge (1.72 million rows), we decided to test our approach on a smaller dataset. To achieve this, we sampled for the first 20000 rows from the raw data. Subsequently, after performing data cleaning, we have 8175 rows of cleaned data and we will proceed to perform exploratory data analysis.

7.3 Exploratory Data Analysis

Since our dataset is too huge (1.72 million rows), we decided to test our approach on a smaller dataset. To achieve this, we sampled for the first 20000 rows from the raw data. Subsequently, after performing data cleaning, we have 8175 rows of cleaned data and we will proceed to perform exploratory data analysis.

7.3.1 Bubble Map

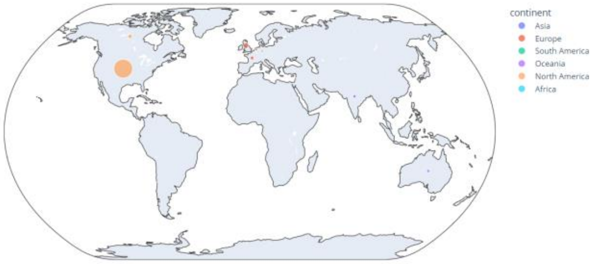


Fig 1: Distribution of tweets across continents

Fig 1 shows a bubble map of the distribution of sampled tweets across the world. As expected, the United States has the greatest number of tweets, followed by some in Europe and very little in the other continents. Since we want to analyze the sentiments of American voters who are mostly based in the United States, there will be less noise in our data and hopefully, this will give us better results.

7.3.2 Word Cloud

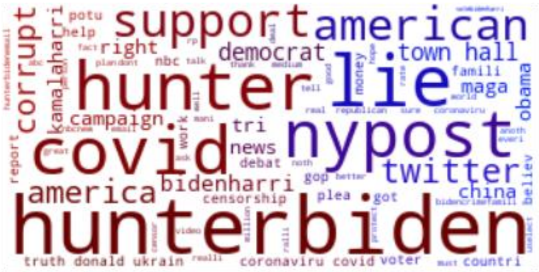


Fig 2: Overall Word Cloud of all Tweets

Fig 2 shows a world cloud for all the tweets in the sample dataset. This is helpful because it gives us an idea on the words that were most frequently used in discourse on twitter in the period leading up to the election.

From the overall word cloud, the most frequent words are “hunterbiden” and “hunter”. A cursory search shows this to be related to Joe Biden’s son, Hunter Biden’s tax crimes and false statement related to a gun purchase that came to rise during the period leading up to the 2020 US Presidential Elections. It is interesting that Joe Biden managed to win the election despite this being the trending topic regarding the elections on twitter. Perhaps, it is because people speak about it with sympathy towards Joe Biden. It could be useful to analyze how the media managed to separate Joe Biden from his son and diminish Biden’s culpability in this issue in the public eye.

Another big word is “lie”. Lies were a big thing in the 2020 US Presidential Elections with both candidates having their fair share of scandals. It would be interesting to analyze how their respective scandals affected voter perception.

“Covid” was another big word, not surprising due to the fact that the elections occurred in the middle of the pandemic. It would be insightful to see if covid did in fact play a part in Trump’s loss because of the way that he handled the pandemic in his term and the election period.

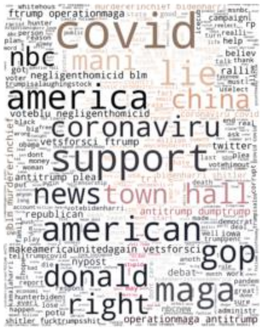


Fig 3: Word Cloud for Trump

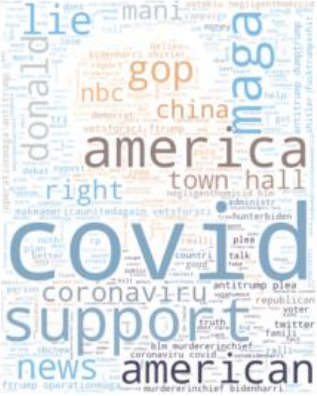


Fig 4: Word Cloud for Biden

Fig 3 and 4 show word clouds for each of the candidates. This is useful in helping us compare between the topics of discourse for both candidates. Using "covid" as an example, we can see that those who talked about Biden cared a lot more about the coronavirus than those who talked about Trump, something to think about for each party as they move forward and tackle more health crises. "Maga" is also very prominent on Joe Biden's word cloud, highlighting the success of Trump and his branding. It might be helpful for the Democratic party to look towards their own branding to help them in future elections.

Overall, these word clouds help give a big picture of the topics of discourse on Twitter in relation to the 2020 US Presidential Elections that we can focus on in our future analyses.

7.4 Sentiment Analysis

Sentiment analysis determines whether a text is positive, neutral, or negative based on its overall valence. The sentiment analysis is based on assessing text valence differently depending on the library or algorithm. These algorithms include the bag-of-words approach used by TextBlob and Vader algorithm, where the text is considered the sum of its constituent words, and valence or sentiment is calculated for words and combined to get a representative valence for the sentence, which informs if the text is positive, negative or neutral. Another approach is the word embedding-based model, where words are represented as vectors of numbers in an n-dimensional space This mapping from individual words to a continuous vector space can be generated through various methods: neural networks, dimensionality reduction, co-occurrence matrix.

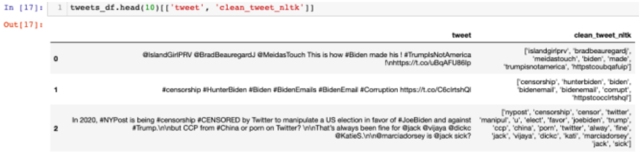
We will use TextBlob and Vader models to analyze the sentiment of tweets for Biden and Trump. We will also use a deep learning model such as Flair or an alternative approach that uses aspect-based sentiment analysis. Flair requires a lot of computing power, so we may have to limit the sample size we can analyze.

We add a column for data sources to identify tweets from Biden and Trump. This will allow us to see the sentiments of tweets from their supporters.

All three models, TextBlob, Vader, and Flair output a continuous number between -1 and 1. For our study, we will use classification and will convert these numerical values into categories. For example, we can classify them as negative if the score is less than 0, positive if the score is greater than 0, and neutral if the score is 0. These buckets for categorization can be changed as needed for analysis.

- TextBlob

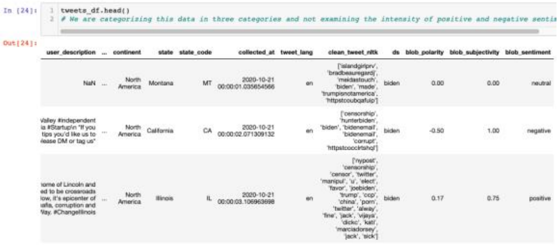
TextBlob is a simple model that calculates the polarity and subjectivity of a tweet. Subjectivity estimates how factual versus opinionated a text is and polarity estimates how positive, negative or neutral is the text. We are using a sample of the total dataset for the computation. Clean tweet is a tokenization of the actual tweet. We started with 10,000 tweets for each candidate and cleaned them to arrive at this sample of data for analysis. A larger cleaned sample can be tried to see if the results vary a lot.



```
In [17]:  tweets_df.head(10)[['tweet', 'clean_tweet_nltk']]
Out[17]:
```

	tweet	clean_tweet_nltk
0	@IslandGirlPRV @BradBeauregardJ @MeidasTouch This is how #Biden made his ! #TrumpIsNotAmerica f\nhttps://t.co/uBqAFU86lp	['islandgirlprv', 'bradbeauregardj', 'meidastouch', 'biden', 'made', 'trumpisnotamerica', 'httpstcoubqafuip']
1	#censorship #HunterBiden #Biden #BidenEmails #BidenEmail #Corruption https://t.co/C6clrtshQl	['censorship', 'hunterbiden', 'biden', 'bidenemail', 'bidenemail', 'corrupt', 'httpstcocclrtshql']
2	In 2020, #NYPost is being #censorship #CENSORED by Twitter to manipulate a US election in favor of #JoeBiden and against #Trump.\n\nbut CCP from #China or porn on Twitter? \n\nThat's always been fine for @jack @vijaya @dickc @KatieS.\n\n@marciadorsey is @jack sick?	['nypost', 'censorship', 'censor', 'twitter', 'manipul', 'u', 'elect', 'favor', 'joebiden', 'trump', 'ccp', 'china', 'porn', 'twitter', 'alway', 'fine', 'jack', 'vijaya', 'dickc', 'kati', 'marciadorsey', 'jack', 'sick']

Fig 5: Screenshot of cleaned tweets after using nltk



```
In [24]:  tweets_df.head()
          # We are categorizing this data in three categories and not examining the intensity of positive and negative sentim
Out[24]:
```

	user_description	...	continent	state	state_code	collected_at	tweet_lang	clean_tweet_nltk	ds	blob_polarity	blob_subjectivity	blob_sentiment
	NaN	...	North America	Montana	MT	2020-10-21 00:00:01.035654566	en	['islandgirlprv', 'bradbeauregardj', 'meidastouch', 'biden', 'made', 'trumpisnotamerica', 'httpstcoubqafuip']	biden	0.00	0.00	neutral
	Valley #independent ia #Startup\n "If you tips you'd like us to lease DM or tag us"	...	North America	California	CA	2020-10-21 00:00:02.071309132	en	['censorship', 'hunterbiden', 'biden', 'bidenemail', 'bidenemail', 'corrupt', 'httpstcocclrtshql']	biden	-0.50	1.00	negative
	iome of Lincoln and ed to be crossroads low, it's epicenter of afia, corruption and Nay. #ChangeIllinois	...	North America	Illinois	IL	2020-10-21 00:00:03.106963698	en	['nypost', 'censorship', 'censor', 'twitter', 'manipul', 'u', 'elect', 'favor', 'joebiden', 'trump', 'ccp', 'china', 'porn', 'twitter', 'alway', 'fine', 'jack', 'vijaya', 'dickc', 'kati', 'marciadorsey', 'jack', 'sick']	biden	0.17	0.75	positive

Fig 6: Screenshot of data after applying sentiment analysis

From Fig 6, we also add a column named blob_sentiment, where we classify the numerical polarity values into categories named positive, negative, and neutral.

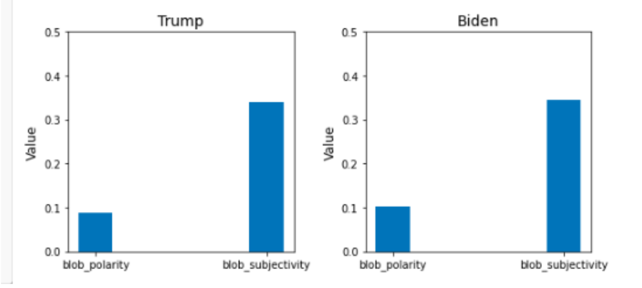


Fig 7: Tweet polarity for Trump and Biden

We can now plot the polarity and subjectivity of tweets in two ways: sentiment expressed per tweet and sentiment expressed per user. In Fig 7, we will compute an average polarity per candidate on a per-tweet basis. The limitation of this approach is that it does not handle spams. Imagine we have 1 user who tweeted 99 times, each having polarity -1 (opposer) and 1 user who tweeted once with polarity 1 (supporter). If we average across all tweets, we obtain -0.98. inferring support / opposing for the candidate would be limited in such a case as we are computing average sentiment per tweet.

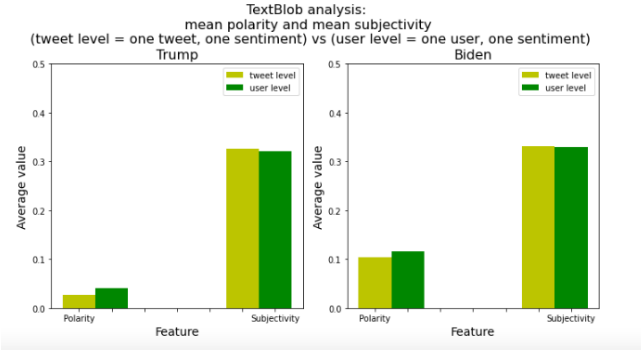


Fig 8: Polarity and subjectivity for Trump and Biden

Fig 8 shows the mean polarity and subjectivity for Biden and Trump. These are average values, so for Trump, the average polarity is a small positive number, whereas for Biden, the polarity is a slightly greater number for both per tweet and user basis. Biden supporters are tweeting more positive sentiments than Trump supporters.

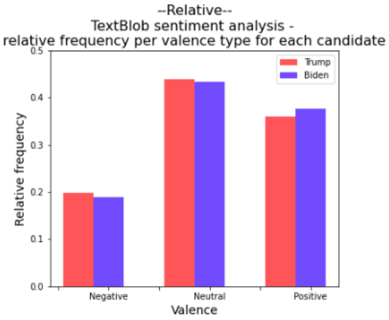


Fig 9: Relative frequency of sentiments for Trump and Biden

Fig 9 shows the relative frequency of sentiments and it is similar to the one in Fig 8.

- VADER

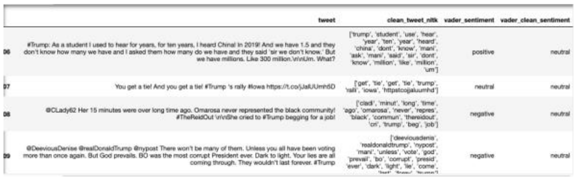
VADER (Valence Aware Dictionary and Sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media and works well on texts from other domains. The valence for the words in the dictionary was empirically validated by multiple human judges who are familiar with this space of tweet and microblogging. It uses some heuristics to recognize word negations (“cool” versus “not cool”) and word intensifiers (“a bit sad” versus “really sad”) to understand the sentiment. The limitation is that it cannot distinguish spelling mistakes and will consider them out of vocabulary words.

Vader is a pre-trained model. Vader outputs something like this:

{'neg': 0.0, 'neu': 0.436, 'pos': 0.564, 'compound': 0.3802}

Negative, neutral and positive are scores between 0 and 1.

The compound value reflects the overall sentiment of the text. It's computed based on the values of negative, neutral, and positive. It ranges from -1 (maximum negativity) to 1 (maximum positivity). There is no standard way to interpret compound. One can decide that whatever is larger than 0 is positive and lower is negative, while 0 means neutral. But we can also decide to look only at more extreme values, like above or below +/- 0.8, for example.



	tweet	clean_tweet_nltk	vader_sentiment	vader_clean_sentiment
06	#Trump: As a student I used to hear for years, for ten years, I heard China! In 2019! And we have 1.5 and they don't know how many we have and I asked them how many do we have and they said 'sir we don't know.' But we have millions. Like 300 million.\n\nUm. What?	['trump', 'student', 'use', 'hear', 'year', 'ten', 'year', 'heard', 'china', 'dont', 'know', 'mani', 'ask', 'mani', 'said', 'sir', 'dont', 'know', 'million', 'like', 'million', 'um']	positive	neutral
07	You get a tie! And you get a tie! #Trump 's rally #Iowa https://t.co/jJalUUmh5D	['get', 'tie', 'get', 'tie', 'trump', 'ralli', 'iowa', 'httpstcojjaluumhd']	neutral	neutral
08	@CLady62 Her 15 minutes were over long time ago. Omarosa never represented the black community! #TheReidOut \n\nShe cried to #Trump begging for a job!	['cladi', 'minut', 'long', 'time', 'ago', 'omarosa', 'never', 'repres', 'black', 'commun', 'thereidout', 'cri', 'trump', 'beg', 'job']	negative	neutral
09	@DeeviousDenise @realDonaldTrump @nypost There won't be many of them. Unless you all have been voting more than once again. But God prevails. BO was the most corrupt President ever. Dark to light. Your lies are all coming through. They wouldn't last forever. #Trump	['deeviousdenis', 'realdonaldtrump', 'nypost', 'mani', 'unless', 'vote', 'god', 'prevail', 'bo', 'corrupt', 'presid', 'ever', 'dark', 'light', 'lie', 'come', 'last', 'forev', 'trump']	negative	neutral

Fig 10: Screenshot of data after VADER

Vader recommends that data may be used uncleaned as cleaning strategies can introduce biases. For this analysis, we add a column in Fig 10 for clean and raw tweets to evaluate if they differ a lot.

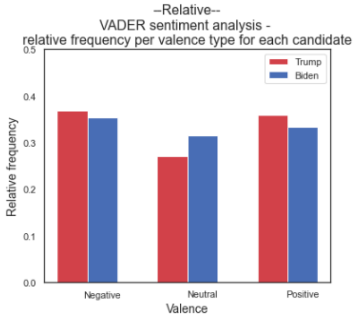


Fig 11: VADER sentiment analysis results

Fig 11 shows the VADER mode result, and it shows that a higher percentage of tweets for Trump is negative. On the other hand, Biden has a higher percentage of positive tweets. The broad patterns of results are similar to TextBlob. The results here show that both negative and positive trend for Trump is slightly greater than for Biden. This could be a result of using strongly positive and negative words in tweets as the Vader model is based on the analysis of the sentiment of each word which gets combined to get the sentiment of the sentence. Biden supporters are tweeting more neutral sentiment tweets based on this analysis.

- Comparison of TextBlob and VADER results

Our dataset is not labeled implying that we do not know by labels if a tweet is positive, negative, or neutral. So, there is no way for us to compare predictions to some 'ground truth'. We can, instead, compare each algorithm's predictions to the ones from the other two. This work remains to be done.

8 Work in Progress

There are still areas that we can improve on in the data cleaning steps. First, we can try out different sampling methods. We can sample by time periods - for example, we can segment the data into two groups: tweets posted before the election day and tweets posted after the election day. By doing so, we can compare the sentiments of Twitter users, for example, whether there are more positive or negative tweets after the election day, which can be associated with the election results.

Building on the word clouds, we have also begun topic modeling. Instead of now looking at singular words and their frequencies, we use the Latent Dirichlet Allocation (LDA) model and the pyLDAvis library to analyze groups of words that are commonly used together.

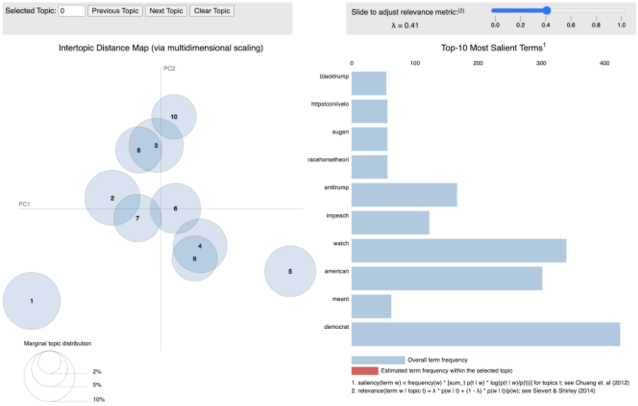

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 0.41
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
10
8
3
2
7
6
4
9
5
1
Marginal topic distribution
2%
5%
10%
Top-10 Most Salient Terms1
0
100
200
300
400
blacktrump
httpstconilvelo
eugen
racehorsetheori
antitrump
impeach
watch
american
meant
democrat
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig 12: Results from Topic Modeling

Beginning to play around with the configurations, the above figure shows a plot of 10 topics. Clicking on each topic bubble then displays the top 10 most salient terms with the estimated term frequency within the topic and the overall term frequency. We are currently identifying the specific topics from the salient terms and hoping to gain insights and explicit links from the topic modeling we have done so far.

REFERENCES

[1] Ali, R. H., Pinto, G., Lawrie, E., &amp; Linstead, E. J. (2022). A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-022-00633-z

[2] Caetano, J. A., Lima, H. S., Santos, M. F., &amp; Marques-Neto, H. T. (2018). Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election. Journal of Internet Services and Applications, 9(1). https://doi.org/10.1186/s13174-018-0089-0

[3] Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., &amp; Singh, K. P. (2020). Analysis of political sentiment orientations on Twitter. Procedia Computer Science, 167, 1821–1828. https://doi.org/10.1016/j.procs.2020.03.201

[4] Hui, M. (2020, November 9). *US election 2020 tweets*. Kaggle. Retrieved October 6, 2022, from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

APPENDIX

Data Dictionary

- created_at: Date and time of tweet creation
- tweet_id: Unique ID of the tweet
- tweet: Full tweet text
- likes: Number of likes
- retweet_count: Number of retweets
- source: Utility used to post tweet
- user_id: User ID of tweet creator
- user_name: Username of tweet creator
- user_screen_name: Screen name of tweet creator
- user_description: Description of self by tweet creator

- user_join_date: Join date of tweet creator
- user_followers_count: Followers count on tweet creator
- user_location: Location given on tweet creator's profile
- lat: Latitude parsed from user_location
- long: Longitude parsed from user_location
- city: City parsed from user_location
- country: Country parsed from user_location
- state: State parsed from user_location
- state_code: State code parsed from user_location
- collected_at: Date and time tweet data was mined from twitter