Analysis of Midterm Election On Twitter

By Kyle Arbide and Zhaoyang Chen

Abstract

The 2022 United States elections were held on Tuesday, November 8, 2022. During this midterm election year, all 435 seats in the House of Representatives and 35 of the 100 seats in the Senate will be contested. Thirty-nine state and territorial gubernatorial and numerous other state and local elections will also be contested. Twitter is a popular social media platform that allows people to keep track of what's being said about certain topics, for example, the midterm elections. It aggregates comments on the topic so that people can not only read about what is happening right now, but also be able to identify the positive and negative comments.

Introduction

Using the Twitter API, we set up a continuous query which allowed us to gather a significant dataset of tweets and threads related to the topic of the upcoming midterm elections.

The election tweets will then be run through a topic modeling algorithm, such as BERTopic, to subdivide the tweets into classes correlated to the political or election related issue they are discussing. Combined with sentiment analysis, these topic clusters will provide us an overview on public sentiment regarding key issues. If the data proves to be granular enough, we can map these sentiments across election zones in the United States.

Tools and Algorithms

The first step will be setting up the call to the Twitter API to collect data related to the upcoming election. This call should run for a significant portion of the semester as to finish with a significant collection of tweets, as well as a representative sample for the entire run up period of the election. Election related tweets were identified by specific keywords and hashtags that are included in the tweets content itself. This query as follows: *midterm elections", "midterm election", "American voters", "american voters", "Senate race", "senate race", senate elections, #Midterms2022, #MidtermElections, #Election2022*. Our dataset should consist entirely of tweets containing one or many of these tokens. Use of Python libraries, such as pandas and numpy,are required to do the data preprocessing and data manipulation. Exploratory analysis will be key to understanding the nature of the twitter data. TF-IDF weighting will allow us to extract the important and related terms within the tweets. Doing this before and after topic modeling will provide a better understanding of how the topic clusters were created. Once topics are created and properly interpreted, we can add additional filters to our data which provide us additional context into the topics being discussed around specific users, areas, and time periods. Each of these filters will be readily customizable through an interface to allow for easy understanding and customization for the end user.

Related Works

The processing and analysis of twitter data is very common, especially for beginner and intermediate level NLP projects. Topics for these kinds of projects vary from sports to significant societal events. However, from our research we were unable to find very many of these analyses that use the most updated and preferred methods from preprocessing and topic modeling. A study done in 2012 titled "Election result prediction using Twitter sentiment analysis" attempted to predict the results of upcoming elections using sentiment analysis on twitter,

but technologies and availability of open source transformer models has skyrocketed since this study. A more recent article was published on medium attempting to replicate this type of prediction on an election in Nigeria, but again the methods consisted of basic sentiment analysis and regression models. For our project we apply some of the latest innovation in Natural Language Transformer Models to create accurate and readable topics.

Data Description and Data Preprocessing

The data were collected continuously from September 20th to November 10th by using Social Feed Manager from GWU. It contains the information of over 800,000 unique tweets, with each data point consisting of 37 features. Among the 37 features, there are 10 features containing an amount of missing values larger than 50% percent of the total number. Also, some features contain information that is not relevant or important to our analysis. For columns that have a very small portion of missing data, a simple average number or majority number will be applied for the imputation. After careful consideration, a total of 21 features were dropped, and the final dataset still contains 16 features. “text”, “hashtags”, “user_location” are three of the most important features which will be used for analysis and models in this project. The “text” column contains the content of a specific tweet. The “hashtags” column contains just tags for each tweet, while “location” means the location where a twitter user has self-identified to be their home.

Natural Language Processing requires rigorous preprocessing and cleaning of text data before any kind of analysis can be done. Our preprocessing pipeline can be described with the following steps. First, we use the “fast_lang” model from the Python spaCy package to identify the language for all of our tweets. Any tweets found not to be in English were dropped from the dataset. Tweets contain some unique types of tokens we needed to account for in our preprocessing, specifically emojis and mentions. For the emojis, we were able to convert these tokens to text so they could be properly consumed by the language model. For example, 🎢becomes `:roller-coaster:`. Mentions are identified using regular expressions and are added to a new column, but remain in the text as well to provide context in the future. Standard practices also meant we removed numbers, punctuation, and all non-unicode characters. Finally, tweets containing less than 10 words were dropped, as they would be difficult to model to topics in the future.

For the last step of our data preparation, we conducted another feature engineering technique by creating two new features called “clean_text” and “mentions.” “clean_text” are the cleaned data which were used to do the exploratory data analysis, visualization and modeling. “mentions” is the data extracted from earlier concerning mentioned users. Finally the size of the dataset is 523794 rows and 19 columns

523792	526563	2022-09-22T00:13:20+00:00	aFathersLog	They would love to coup Lula and keep Bolsonar...
523793	526564	2022-09-22T00:13:48+00:00	IFollowTheFacts	House passes bill to prevent stolen elections,...

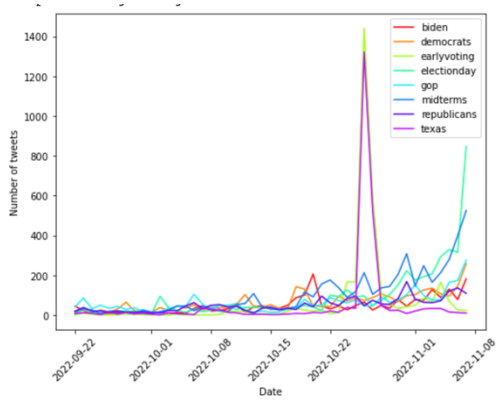
523794 rows × 19 columns

Exploratory Data Analysis:

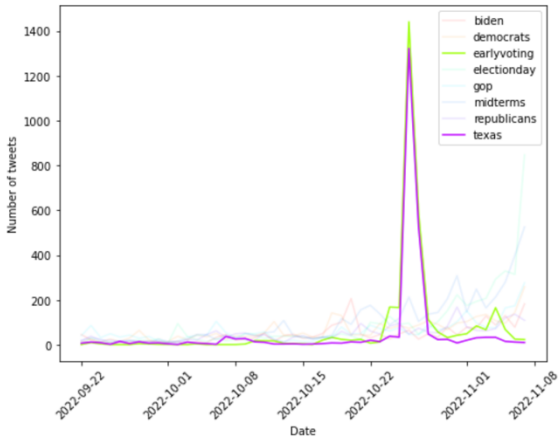
In our dataset, the hashtags in a tweet will play a big role in the exploratory data analysis step. We can get a basic idea of some popular topics by looking at the highest number of tweets by hashtag. For preparation of the hashtag analysis, we

created a new dataset called hashtag_tweets which only includes data points that have hashtags. Those entries with multiple hashtags are also split into multiple rows to ensure each hashtag is counted as one occurrence.

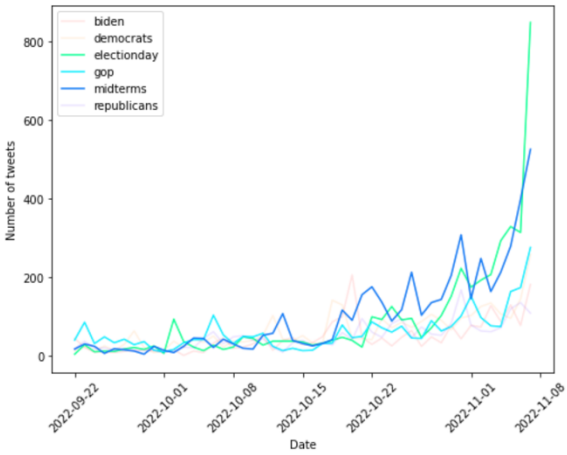
Grouping the data by hashtag and date, we found the number of tweets per tag per day. If we group the data by the hashtag only, we get the overall most used tags. The plot below shows the most popular hashtags in a time period from 2022/9/22 to 2022/11/8.



We can see that “early voting” was the most popular topic on the date of 10/24/2022 which has about 1500 tweets. Also, Texas was a popular topic from 10/23-10/25. Clearly there is a large correlation between the “early voting” and “texas” hashtags, so we highlight the two in the next plot. We theorize that this is caused by the occurrence of a specific event outside of twitter concerning these two topics, but other factors could be influential as well, such as bot spamming of the two tags, or some kind of glitch with the API call. We will explore this more later on.



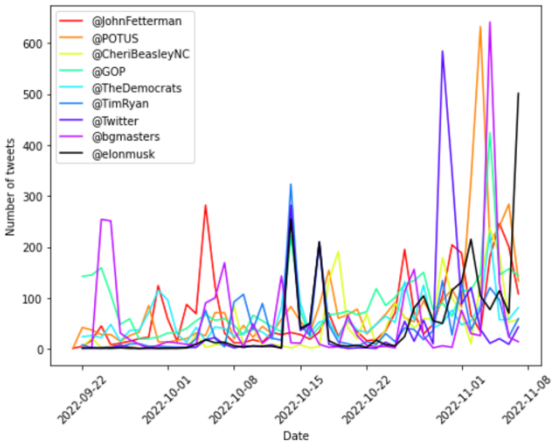
Other hashtags such as “electionday”, “gop”, and “midterms” show a different kind of pattern. These tags show slow, continuous growth throughout the collection period, with some small variations in popularity. We also see the largest uptick come right around election day on November 8th.



The next part for our exploratory data analysis is looking into the “mentions”. Unlike “hashtags”, “mentions” typically only refer to a certain user or certain brand. By taking all the mentions, we can see that “@GOP” is mentioned 4238 times, and “@POTUS” is mentioned 4016 times.

	mentions	size
0	@GOP	4238
1	@POTUS	4016
2	@GETTRofficial	3575
3	@JohnFetterman	3161
4	@TheDemocrats	2783
5	@bgmasters	2673
6	@elonmusk	2358
7	@TimRyan	2125
8	@CheriBeasleyNC	2108
9	@Twitter	2105

It is also interesting which mentioned users are most frequent for different time periods within our dataset.



It is hard to distinguish any relationships from this plot, so we can utilize pearson correlation to understand if there is any relationship between the mentioning of different users each day. The output is the correlation matrix shown below. The highest correlation can be seen between "@CheriBeaslyNC" and "@GOP", with a score of 0.5.

mentions	@CheriBeasleyNC	@GETTRofficial	@GOP	@JohnFetterman	@POTUS	@TheDemocrats	@TimRyan	@Twitter	@bgmasters	@elonmusk
@CheriBeasleyNC	1.000000	-0.058840	0.500995	0.391340	0.449025	0.262234	0.210384	0.250206	0.338988	0.129644
@GETTRofficial	-0.058840	1.000000	-0.080490	-0.210582	-0.246517	-0.143705	-0.037567	-0.143202	0.167680	-0.197046
@GOP	0.500995	-0.080490	1.000000	0.311756	0.403340	0.642527	0.347950	0.073183	0.677250	0.370203
@JohnFetterman	0.391340	-0.210582	0.311756	1.000000	0.253626	0.221902	0.163832	0.212464	0.282855	0.226106
@POTUS	0.449025	-0.246517	0.403340	0.253626	1.000000	0.253470	0.187959	0.094453	0.102630	0.366139
@TheDemocrats	0.262234	-0.143705	0.642527	0.221902	0.253470	1.000000	0.516009	0.265954	0.394619	0.379597
@TimRyan	0.210384	-0.037567	0.347950	0.163832	0.187959	0.516009	1.000000	0.530097	0.073816	0.500255
@Twitter	0.250206	-0.143202	0.073183	0.212464	0.094453	0.265954	0.530097	1.000000	-0.137032	0.338062
@bgmasters	0.338988	0.167680	0.677250	0.282855	0.102630	0.394619	0.073816	-0.137032	1.000000	-0.003329
@elonmusk	0.129644	-0.197046	0.370203	0.226106	0.366139	0.379597	0.500255	0.338062	-0.003329	1.000000

Topic Modeling

BERTopic is a topic modeling technique that leverages transformers and c-TF-IDF to create dense clusters allowing for easily interpretable topics whilst keeping import words in the topic descriptions. It is built on top of BERT, one of the most powerful transformer models developed by Google. BERTopic produces large models that take long periods of time to train, so in an attempt to increase efficiency we filtered the dataset to only contain original tweets and replies to tweets. This left is with around 100000 tweets for model training. The rest of the tweets were transformed after model training.

The original run of the model produced over 1000 topics ranging in size from 10-35,000.

Topic	Count	Name
-1	34246	-1_people_ballot_get_win
0	1034	0_twitter_elonmusk_elon_musk
1	932	1_abortion_abortionrights_forcedbirth_reproductive
2	737	2_florida_desantis_charliecrist_vbm
3	649	3_inflation_economy_reduction_recession
...	...	...
998	10	998_everyones_finances_worlds_hopeless
999	10	999_archer_bonuses_crossbones_ps
1000	10	1000_mailbox_stitt_flag_tulsa
1001	10	1001_fraud_fraudulent_legit_nuisance
1002	10	1002_tightened_warning_sign_dr

ows × 3 columns

The first row, topic = -1, will serve as a baseline for our topic model. This topic contains tweets that have no significant

meaning beyond general election verbiage. The terms in the “name” column represent the most frequent terms for that topic.

Topics with only 10 tweets feels too small for what we are trying to accomplish, so we are able to decrease the number of topics automatically with BERTopic. After some experimentation and trial-and-error, we landed on reducing the total number of topics to 75. The results of this are shown here:

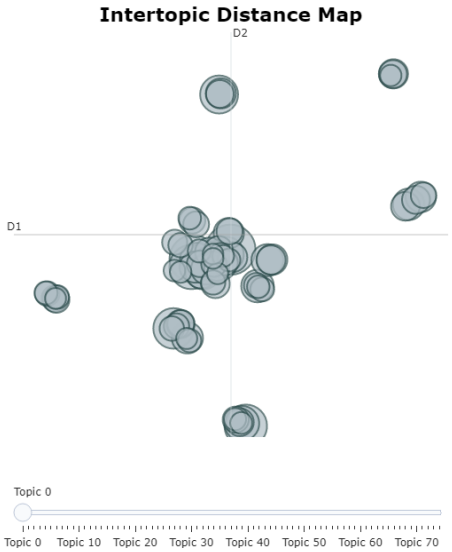
Topic	Count	Name
-1	46838	-1_ballot_gop_trump_republicans
0	1418	0_abortion_abortionrights_roe_women
1	1268	1_twitter_elonmusk_elon_musk
2	1101	2_ohio_ryan_vance_tim
3	1017	3_early_voting_earlyvoting_find
4	943	4_florida_desantis_rubio_charliecrist

We see that the most popular topics have changed slightly and the overall number of tweets in each topic has increased significantly. Remember, this section was created using only a subset of the tweets, so we expect the number of tweets per topic to go up across the board.

Additionally, we experimented with an additional method which included training a second topic model to re-run those tweets categorized as topic = -1. The purpose of this was to see if there was any additional information we could gather from these general tweets. We found some success using this method, shown below:

Topic	Count	Name
-1	33679	-1_democrats_ballot_gop_voting
0	1812	0_women_abortion_rights_children
1	1342	1_fetterman_oz_pennsylvania_droz
2	767	2_georgia_walker_herschel_herschelwalker
3	756	3_russia_ukraine_russian_putin
4	681	4_money_dark_billion_spending

As you can see, we were still able to decipher some meaningful topics from the remaining tweets by re-running these tweets. However, we ultimately chose to stick with only the initial model, as running both models would take too much computation and runtime across the entire dataset.



Looking back at the first model, we can take the first two principal components for our features and plot them to understand the similarity between our topics. In the plot above we see some large overlaps between many of our topics. This suggests that the overlapped topics have some basic similarities. The amount of overlaps in this plot suggests that for future work, we could explore decreasing the number of topics we choose to include.

Topic Plotting

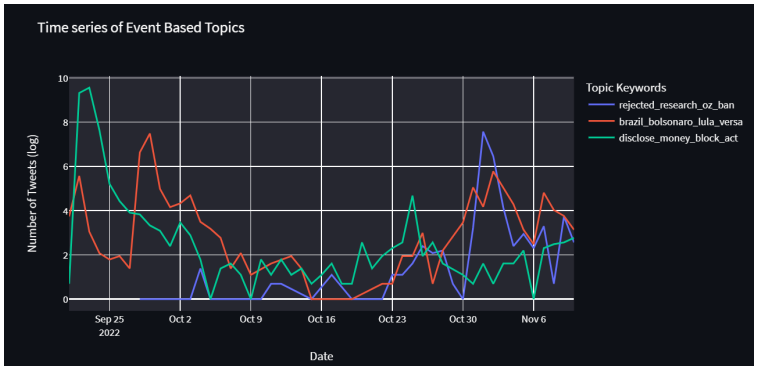
After we have done the topic modeling, we would like to make some visualizations and get more insights into the topics. This requires us to first go back and fit our trained model on all of the tweets in the dataset.

As mentioned before, we are hoping to show multiple time series visuals which will provide insight into those topics being discussed throughout the election. This means aggregating the data in a similar

method to that from the EDA section with our new topics. The output of this aggregation will look something like what is generated below.
This information will be fed into our visualizations, which is later uploaded to our interface. The result is the following visual generated on our web app, using a log-y scale to normalize the large increase in number of tweets closer to the election day.

Now this visual gives some good insight about the most popular topics in our dataset, but what about those event-based topics mentioned earlier in the EDA section? To find these topics we looked for high variance from date to date. We also normalized this variance by diving by the median value for the topic. The following topics ended up near the top of this measure, and we explored to see if these qualified as "Event-Based topics".

	topic	var	median	avg_var
72	71	6.393939e+06	6.0	1.065656e+06
0	-1	5.054518e+08	4491.0	1.125477e+05
55	54	7.732913e+05	10.0	7.732913e+04
52	51	1.233118e+05	3.0	4.110394e+04
64	63	4.454235e+05	14.5	3.071886e+04



The three topics shown above each have irregular patterns which we found interesting. The topic "disclose_money_block_act" in green shows a high spike frequency around September 20-25, much higher than what we see on election day. Similar spikes are seen for the blue and red topics in the visual. All three of these topics fail to peak in frequency on or near election day, which is vastly different from what we see in the previous visual. In future work, we would love to dive deeper into these topics to dissect what exactly drives these behaviors, and if we can conclude that these topics are less significant to the overall scope of the election given that their usage declines significantly from its peak before election day.

Summary and Future Work

Using a dataset of over 800,000 tweets surrounding the topic of the Midterm Election in 2022, we were able to generate accurate and readable topic models from our tweets. From these topics, we implemented multiple forms of analysis and visualization which provided additional insight into the important topics surrounding different periods of times and twitter users. This combined with our interface allows end users to generate their own analysis on locations and users relevant to their elections and interests.

This work should serve as a baseline for performing more advanced analysis on

twitter text and election reaction. Readable topics are vastly beneficial when combined with other algorithms such as election prediction. The results of a prediction algorithm generated from our topic modeling will be much more interpretable than some of the results from works done prior. Additionally, our text can be fitted to sentiment analysis models to allow for additional information into each of the topics themselves.

Finally, there is a vast amount of information remaining in our dataset which we elected to exclude from our analysis. Additional features such as “Number of Likes”, “Number of Retweets”, etc. could be leveraged in future analysis.

Source Code

All of our work can be found on Github here: https://github.com/kylearbide/miderm_election_tweets

References

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Idehen, Kelly. “PREDICTING NIGERIA 2019 ELECTION WITH AI — SENTIMENT ANALYSIS.” Medium, 2019, https://medium.com/@kelly-idehen/predicting-nigeria-2019-election-with-ai-sentiment-analysis-d5daecf291ad. Accessed 27 Nov. 2022.

J. Ramteke, S. Shah, D. Godhia and A. Shaikh, "Election result prediction using Twitter sentiment analysis," 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, pp. 1-5, doi: 10.1109/INVENTIVE.2016.7823280.