Hirad Nourbakhsh

Directed Studies - United States Midterm Elections

Introduction

Among the fanfare and posturing characteristic of the American political landscape, the buildup to, and outcome of, the United States Midterm Elections stand out for their uncanny ability to serve as an unbiased assessment of the current president's performance. Historically, the party in power typically performed worse than the opposition during the midterms, and particularly successful candidates were tapped by party insiders as potential presidential candidates. While in the past voter turnout was considerably lower for midterms compared to presidential elections, as the political landscape in the U.S. becomes increasingly polarized, engagement during midterms has increased significantly, with the past election cycle (2022) being particularly contentious in light of marred diplomatic relationships internationally and claims of voter fraud and election stealing locally. Such controversy has inevitably led to a rich hotbed of social media activity, which we will mine and analyze to better understand the key issues and networks defining this landscape.

This report is divided into three main sections. Part I describes the process of text mining Twitter feeds pertaining to historical and current midterm elections, and performing topic analysis to identify the main issues on voters' minds during the midterm election cycle. Part II involves a comparison between the social networks surrounding the midterm elections in the two months before, and the month after them. In Part III, we will mine headlines pertaining to the midterm elections from BBC News, perform a second round of topic modeling, and identify changes in policy or the emergence of new firebrands in either political party following the elections.

Part I - Text Analysis and Topic Modeling

Overall methodology

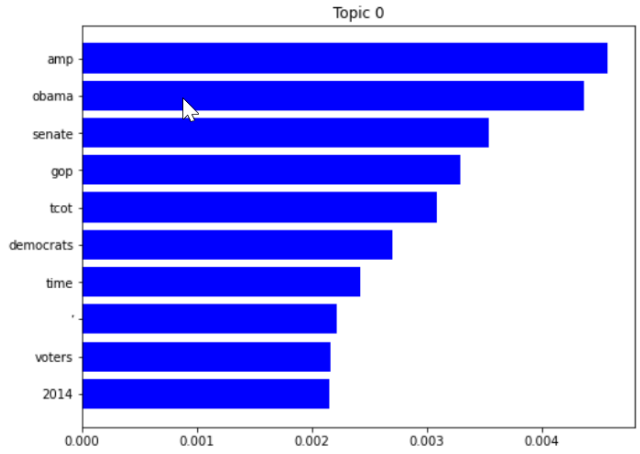
There were three separate rounds of topic modeling performed. The first and second rounds of topic modeling used twitter data from the 2014 and 2018 midterm election cycles, respectively. The third round of topic modeling was performed using tweets captured from the 2022 midterm election cycle. In all three cases, tweets were taken from two months before election day until election day itself.

In each round, twitter feeds were mined using SNSscraper with the hashtag #midterms. 4000 tweets were mined in each round. Information pertaining to the tweet's user, the tweet's date, the

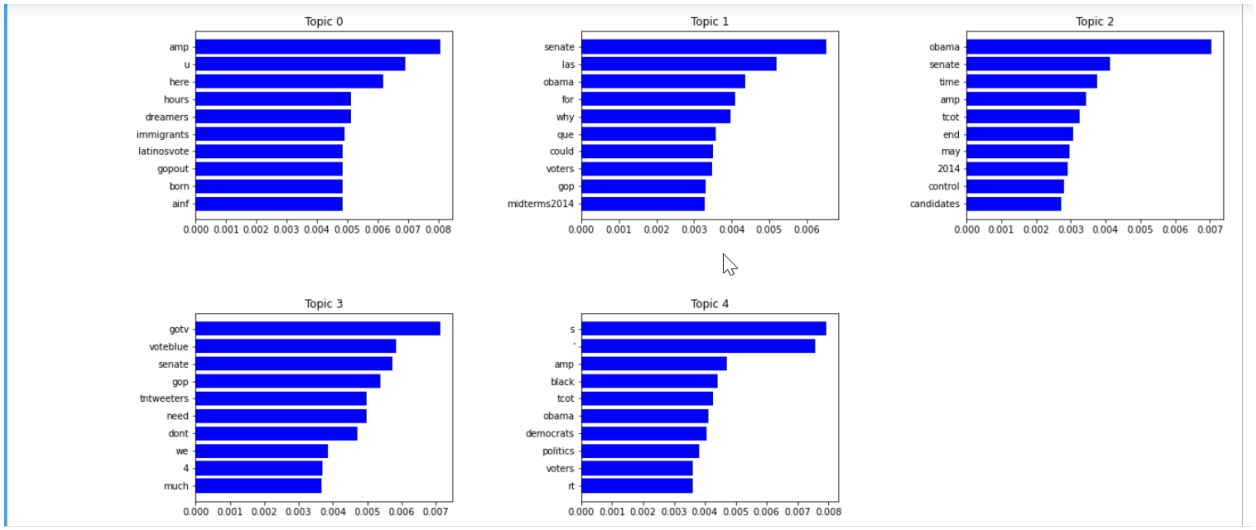
tweet's number of likes, its source, and the tweet itself was captured. Tweets were cleaned to remove hyperlinks. Next, stopwords were removed using the nltk package. Tweets were lowercased, punctuation was removed, and the tweets were tokenized. Next, the bag-of-words was generated, and words that were deemed irrelevant or noise were removed. The tokenized tweets were converted into a dictionary, which was used to generate the tweet corpus, which in turn served as the input to the LDA topic model. The overall goal in this section was to use the results of topic modeling from the 2014 and 2018 elections and the hindsight knowledge of the election results to predict the overall performance of the two main political parties in the 2022 election.

2014 Elections - Analysis

Interestingly, the first round of topic modeling with the #midterms hashtag produced several words pertaining to midterm exams ("study", "coffee", "TA", "biochem"). These words were removed prior to developing the LDA model. It is worth noting that this issue did not occur using the 2018 and 2022 data, presumably because Twitter's usage as a platform for political discussion has been far more prevalent in the years following Trump's presidency. The LDA model for the 2014 midterms displayed a marked decrease in coherence score as the number of topics increased, with the highest coherence score of -6 associated with a single topic. This is most likely due to the lower use of Twitter during the Obama years, as mentioned previously. The most notable result in this topic was that of #TCOT, which stands for Top Conservatives on Twitter. The hashtag was particularly popular among Tea Party followers (and candidates) for locating like-minded Republicans online. It is noteworthy that no similar hashtag or topic associated with the political left (i.e. Democrats) appeared in this topic. Given that the Republicans retained control of the House of Representatives and won control of the Senate in 2014, it is unsurprising that the one partisan hashtag that appears in this topic is associated with conservative politicians.



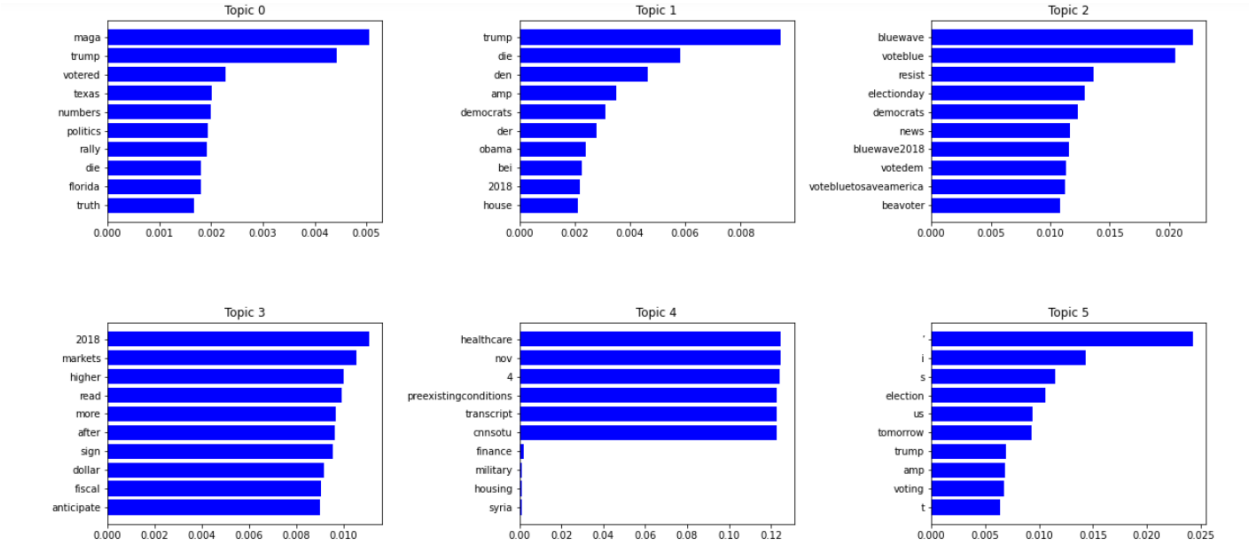
While the highest coherence score was associated with a single topic, I also graphed five topics in order to better understand the underlying distribution of the data. Other noteworthy results in these other topics included “dreamers,” “immigrants”, “latinosvote”, “gopout”, and “voteblue.” #GOPout and #VoteBlue were hashtags associated with the Democrats’ campaigns, evidently designed to stir up opposition against Republicans. The DREAM Act was implemented by the Obama administration to provide a path to permanent residency for illegal immigrants who entered the US as minors. The majority of the DREAMers were of Latino descent. Thus, the topics of DREAMers, immigrants, and #latinosvote indicate the significance of immigration and illegal immigration to the 2014 midterms, especially when considered against the backdrop of a possible Republican takeover of both the House and the Senate (Republicans were overwhelmingly opposed to the DREAM Act, and President Trump rescinded it in 2017).



2018 Elections - Analysis

The 2018 twitter scrape provided far more relevant and richer text data compared to the 2014 elections. Despite the fact that the coherence score was slightly lower for 6 topics than 1 topic based on the LDA model, 6 topics were chosen for topic modeling in order to provide greater interpretability. An overview of the six topics’ results provides an excellent high-level overview of the issues on voters’ minds during the 2018 midterms. In topic 0, the results of “Trump”, “MAGA”, “Texas”, “Florida”, and “rally” collectively indicate the Trump campaign’s tactics. President Trump was then focused on holding mass rallies in two of the most populous and influential Southern US states in order to galvanize his base and attract more voters to ballot boxes. “Topic 2” shows a hotbed of Democrat political activity, as indicated by results such as “bluewave”, “voteblue”, “resist”, and “votebluetosaveamerica”. From 2016-2018, the Republicans had enjoyed a majority in the House, Senate, and the Presidency, allowing for considerably unhindered leeway when it came to passing legislation. Topic 2’s results are indicative of fervent political activity by the Democrats to get their base to vote for them and

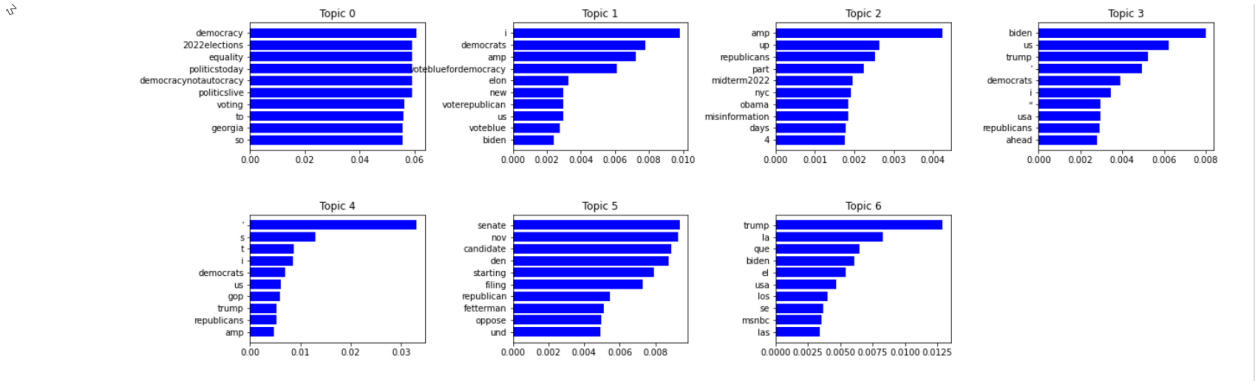
allow their representatives to stop Trump's agenda. It is worth noting that this number of left-leaning topics were not observed in 2014, where the Democrats performed much poorer than in 2018. Topic 3 indicates the fluctuation of global capital markets with respect to a then-indeterminate midterm election outcome with results such as "markets", "dollar", "fiscal", and, tellingly, "anticipate". Last, Topic 4 reveals the issues occupying voters' minds and politicians' platforms: "Healthcare", "finance", "military", "housing", and "Syria." With regards to the election outcome, the most noteworthy correlation with the topic modeling was the abundance of pro-Democrat and anti-Republican results. Given that the Democrats regained control of the House of Representatives in the 2018 midterm elections, a correlation between higher voter engagement online and better election performance by those voters' preferred party can be observed across both the 2014 and 2018 midterms. We will use these patterns to attempt to suss out the winning party in the 2022 midterms.



2022 Elections

Several noteworthy trends appear here. First, all the main media outlets mentioned in the results (MSNBC, PoliticsLive) are left-leaning outlets. Next, names of politicians other than former or current presidents appear in the results, in this case, that of John Fetterman, who successfully ran against Dr. Mehmet Oz in the state of Pennsylvania. Third, there is a relatively greater proportion of pro-Democrat hashtags than pro-Republican: "#Democracynotautocracy", "#votebluefordemocracy", and "#voteblue", versus the "#voterepublican". As in previous rounds, names of former and current presidents (Trump, Biden, and Obama) and the two main political parties appear in relatively equal proportion across the topics, thus disallowing any meaningful conclusions. The mention of Georgia among the results is unsurprising, given the importance of that state to both parties' overall campaigns and increasing reports of voter suppression within it. What was noteworthy about the 2022 data was that despite constant

mentions of a “red wave” across American media outlets that would sweep the US and give the Republicans control over the House and the Senate, the distribution of the results in the topic modeling were remarkably similar to those of the 2018 midterms, that saw a Democratic win of the House and lackluster Republican performance, in so far as they suggested greater pro-Democrat voter engagement. Based on this data, I anticipated that the Democrats would at least maintain control of either the House and Senate, and that is indeed what happened. The Republicans won a razor-thin margin in the House of Representatives, but as of the time of writing, the Senate remains in Democrat control. In the next section, we will use network analysis to understand the shifts in social networks before and after the 2022 Midterm elections.



Part II - Social media analysis

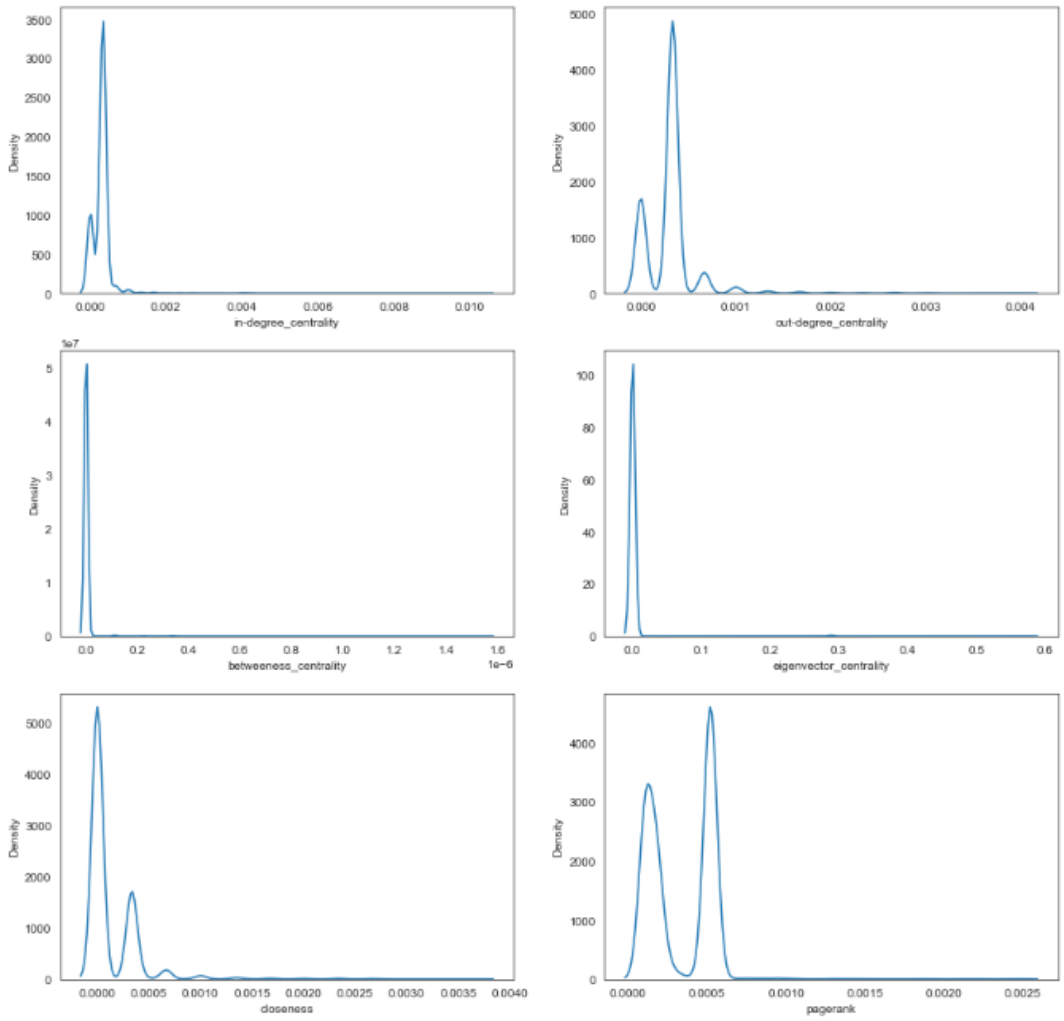
Overall methodology

For the network analysis, information pertaining to the username, the date of the tweet, and other users mentioned in the tweet was pulled using the SNScraper tool. The user mentions were used as an approximation of retweets, as SNSscraper does not provide extensive information on retweeting users and for the purpose of drawing a network graph and understanding relationships between accounts, user mentions was sufficient.

Two rounds of network analysis were conducted. The first round used a total of 4000 tweets taken from two months before midterm election night until election night itself (i.e. September 6 - November 7 2022). The second round used 3000 tweets taken from election night until the 24th of November. Fewer tweets were used in the second round as the network analysis failed to converge with 4000 tweets. For each round, a network graph was plotted and the following measures were calculated: In-degree centrality, out-degree centrality, betweenness centrality, eigenvector centrality, closeness, and pagerank.

Network Analysis: Pre-Midterm Elections

A directed graph was created wherein edges were drawn from users to those they had mentioned. As mentioned previously, additional network metrics were calculated to gain further insight into relationships leading up to the 2022 Midterm elections.



We can see that for all of the metrics calculated, a handful of accounts score extremely highly, while others score low. This indicates that the overwhelming majority of political discourse revolves around a small number of accounts and individuals.

Next, the top scoring accounts for network metrics were listed. For betweenness centrality, the top scoring accounts were media publications (France's Le Monde, Wisconsin Now, Conversation) or left-leaning PACs and pundits: The Grande Ecole professor Correntin Selin, and the Democratic, pro-union political advisory committee MakeRoadActPA. Given that betweenness centrality measures the number of times a node appears on a path between other nodes, it is logical that the majority of accounts connecting others on Twitter during the build up

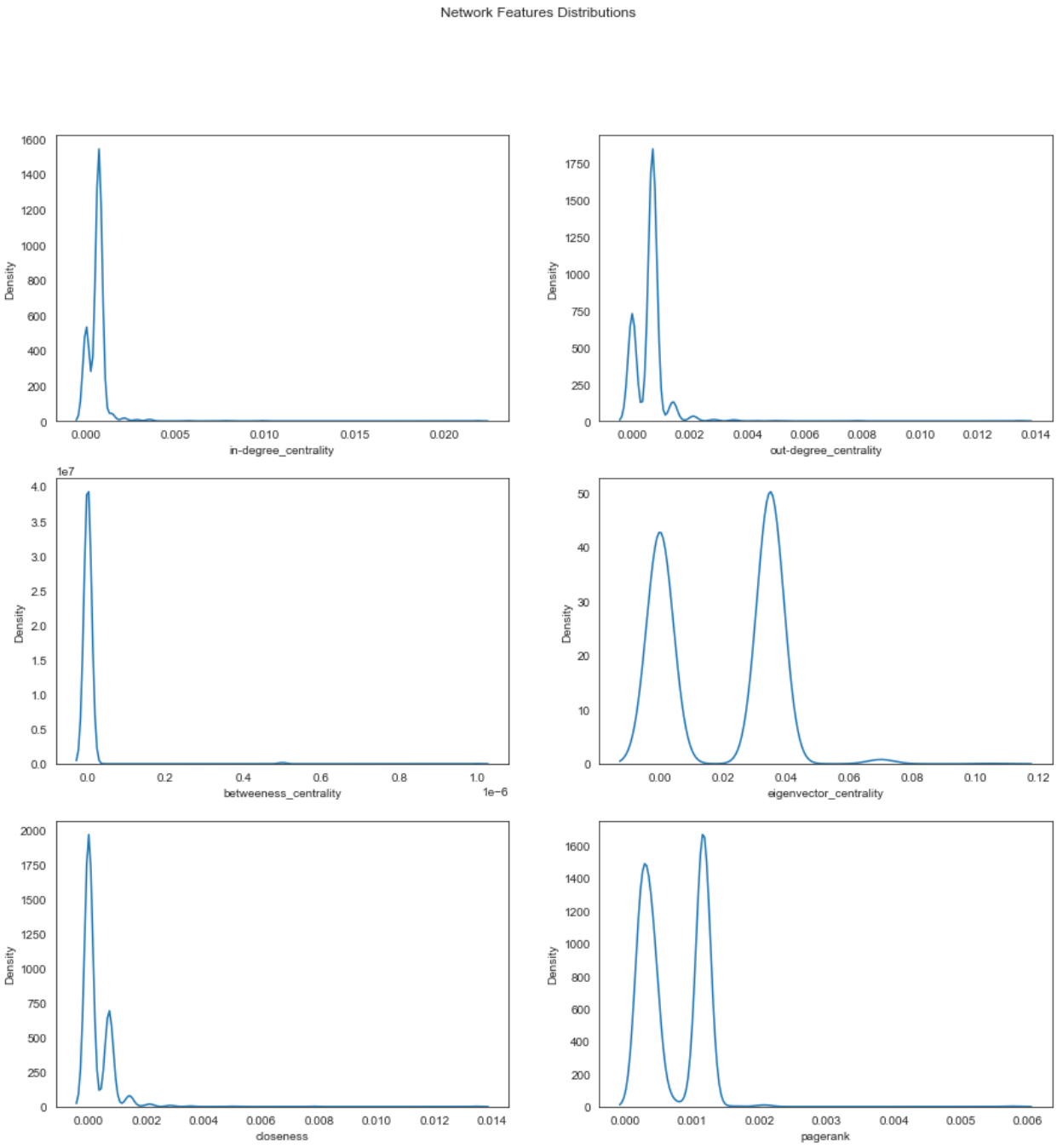
to an election are news outlets. Given Wisconsin's status as a swing state, it is unsurprising to see it achieves similar betweenness centrality scores as national publications such as Conversation and Le Monde. It is telling that a pro-Democrat PAC in Pennsylvania appears in the top 10 betweenness centrality scorers and its Republican counterparts do not. John Fetterman beat Dr. Oz in the state of Pennsylvania, and it is likely that the heightened activity of the PAC that supported him (and a lack of comparative donor activity online for Dr. Oz) directly contributed to this win.

The top scorers in out-degree centrality and closeness displayed considerable overlap. In both cases, the majority of the top scorers were pro-Democrat and left-leaning. These included Sondra Thorsland, Sellin, ReneNow, and the progressive writer John Stauber. The one exception was Dr. Floridian, a staunch advocate of the GOP and in particular its push to rein in Big Tech. Based on these metrics, we can note that Democratic accounts displayed a greater ability to reach wider audiences and spread their message quicker than Republicans, yet another correlation with their better-than-expected performance.

Of note was that among the top scorers for in-degree centrality leading up to election night, almost all were individuals, organizations, or media outlets associated with the Democrats. These included POTUS (Joe Biden's account as the President), Joe Biden, MSNBC (a news outlet heavily favoring Democrats), TheDemocrats, John Fetterman, and the NY Times. The only exceptions to this rule were Elon Musk and the GOP. Even in the case of Elon Musk, the eccentric CEO did not expressly endorse the Republicans' platform, he merely told his followers to vote for them in order to keep a balance of power between the two mainstream parties. Thus, it appears that at the highest level, Democrats and their base were able to outperform the Republicans in terms of marshaling online activity toward them.

Network Analysis: Post-Midterm Elections

As in the case of pre-midterms, a directed graph was drawn, and network metrics were calculated.



We can again note a similar pattern to what was observed with the pre-election data, wherein a handful of accounts scored highly, while others scored extremely low. Top 10 scorers were calculated for each network metric. In terms of betweenness centrality, the majority of accounts with the exception of one ("Darius Bur") were pro-Democratic. Similar accounts appeared in out-degree centrality and closeness as well. As with pre-election data, the majority were pro-Democrat accounts, with the exception of the staunchly Republican RedWaveClips. Interestingly, a few politically neutral accounts appeared among the top out-degree centrality scorers, including OneLadyOneVote and RHCommonBridge. The narrow margins between

Democrats and Republicans and the election result that stayed indeterminate over several days likely contributed to the ability of non-partisan platforms to increase their reach.

The most interesting change in network metrics before and after the midterms was in eigenvector centrality, a measure of the influence of a node in a network. Several individuals appeared amongst the top scorers for eigenvector centrality following the midterms that were not among the top scorers before. These included Mitch McConnell, Jim Jordan, Kari Lake, Raphael Warnock, and Josh Hawley. Jordan, Lake and Hawley are all three ardent Trump supporters who backed his claim that the 2020 election was stolen in Biden's favor. Warnock is a Democrat who ran for US Senate in the state of Georgia, and as of the time of writing, he has not been conclusively declared the winner. McConnell is the Republican Senate Minority Leader whose relationship with Trump has soured in recent years. It is worth noting that despite the fact that the anticipated "red wave" failed to materialize and several Trump-backed candidates (including Lake herself) lost in their races, three of the former President's most vocal supporters apparently increased their influence (at least on social media) following the midterms. This comes at a time of widespread condemnation and ridicule of Trump, with several rising stars within the party such as Florida's DeSantis publicly blaming him for their party's lackluster performance and attempting to distance themselves from him. This trend of increased activity revolving around Republicans (and pro-Trump Republicans in particular) was further exemplified by the top scorers for in-degree centrality: Following the midterm elections, these now included McConnell, Lake, Jordan, and Lauren Boebert, the pro-Trump Colorado representative who infamously attempted to enter Congress while carrying a handgun. It appears that in light of the apparent dip in support for Trump, those politicians most closely associated with him have apparently reinforced activity from their base. Given the trend we saw previously with increased activity among Democrat supporters correlated with a stronger election performance, it is clearly too early to discredit Trump's influence among the Republican party. Indeed, based on these results, it is entirely possible that his base may be reenergized and vote him to power in 2024.

Part III - Web Scraping

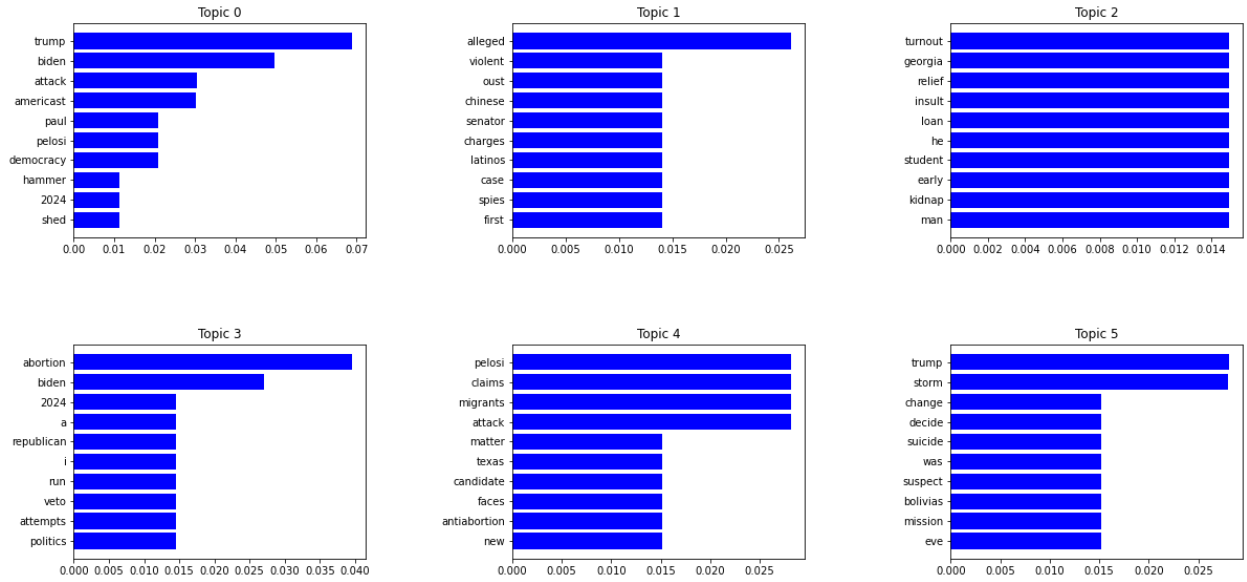
Overall methodology

BeautifulSoup was used to scrape headlines from BBC News. I used BBC News as opposed to American news outlets specifically because of its non-American status. I felt that because it existed outside of the United States and is funded in part by the British government, it would likely be a more impartial option for scraping news headlines than US outlets such as CNN or Fox News. I used the US News section of the BBC and searched "midterm elections." I scraped the first six pages of results, which began on November 24 and ended on September 15. I then split the scraped headlines into two datasets. The "Before" dataset was comprised of headlines printed from September 15 until November 8. The "After" dataset was composed of headlines

printed from November 8 until November 24. The ultimate goal was to compare the headlines before and after the midterm elections in order to understand whether the result led to a significant change in policy among either the Democrats or Republicans, or whether new candidates increased in popularity. For both datasets, a similar approach was followed as in section I. Headlines were lowercased, stopwords and punctuation were removed, and the headlines were tokenized. Bag-of-words models were generated, irrelevant words were removed, and the tokenized headlines were converted into a dictionary. This dictionary was used to generate the headline corpus which was fed into the LDA model.

Analysis of Results - Pre-Midterm Elections

For the pre-midterm headlines, 6 topics were chosen based on the LDA model, given that an increase in topics past 6 did not significantly improve coherence scores.

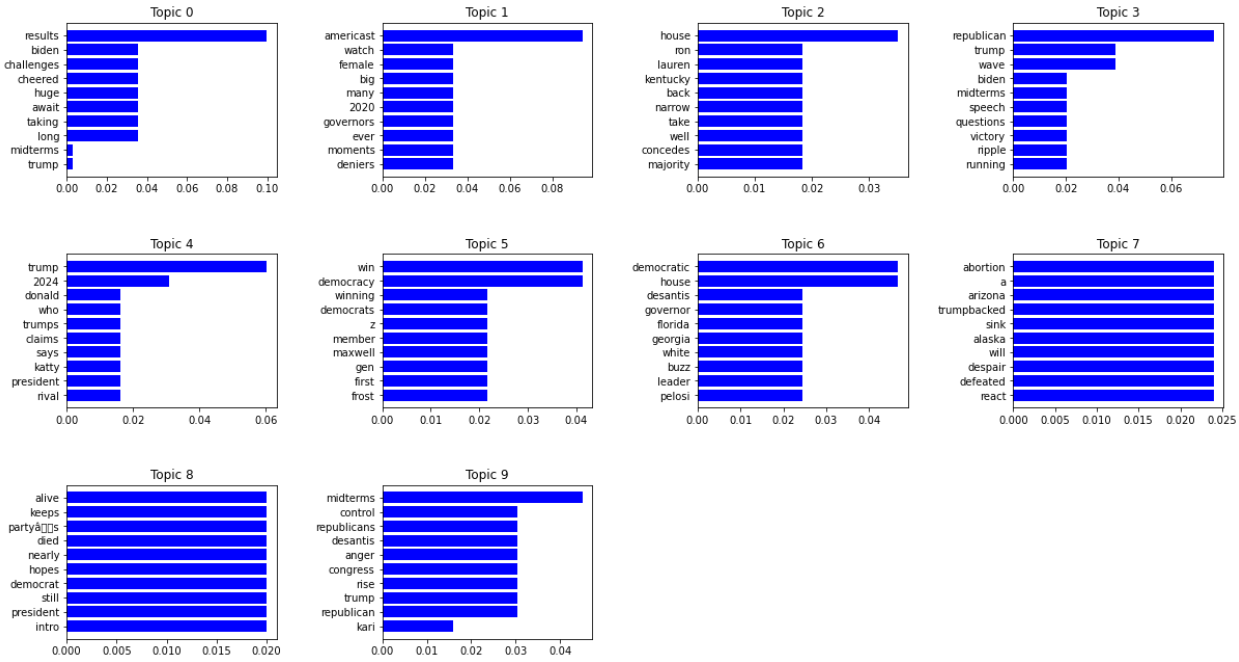


In terms of typical points of political contention in the United States, the most notable results include “loan”, “abortion”, and “migrants”. The names that appear among the results are those of Trump, Biden, and then-House Speaker Nancy Pelosi. Overall, both the issues and the people that appear in the headlines are expected and do not represent any significant surprises or changes from previous midterm elections in America. However, the post-midterm results show a markedly different landscape.

Analysis of Results - Post-Midterm Elections

10 topics were chosen based on the LDA model to optimize the trade-off between increasing model complexity and coherence scores. We can immediately note the presence of several high-profile names that were nowhere to be found in the pre-election topics: “Kari” (Kari Lake),

“DeSantis” and “Ron” (Florida governor Ron DeSantis), and “Lauren” (presumably referring to Lauren Boebert). Additionally, a number of states that did not appear in the pre-election topics, presumably because it was considered obvious at the time which way they would vote, show up in the post-election topics. These include Arizona, where the Democrat Mark Kelly defeated Trump-endorsed Lake in what was traditionally considered a Republican stronghold, Alaska, where moderate Republican and anti-Trump candidate Lisa Murkowski won by write-in (the first candidate to do so since the 1950s), and Florida, where another anti-Trump Republican delivered a resounding victory. The presence of these states in the topic modeling results indicates the impact that their elections had on the populace, given the fact that they are repeatedly being mentioned in a non-American news outlet on the other side of the Atlantic. Indeed, the wins of multiple publicly anti-Trump candidates across both political parties and their repeated mention in the media may serve to illustrate the American public’s fatigue with the former president, and a coming shift among the Republican establishment to procure a fresher face for their party.



Conclusion

Our results reinforce P.T. Barnum’s adage of “no such thing as bad publicity.” We can see that the more active a party’s base is on social media, the more likely they are to win. The results from Part I show a clear positive correlation between the number of mentions of a particular party, or that party’s stance on an issue, and the party’s overall performance in midterm elections. In each round, the party whose candidates or stance on a particular topic featured more prominently in the topic modeling was the party that performed better on election night. The results from the network analysis in Part II reinforce this pattern, as the party whose affiliates, be

it politicians, friendly news outlets, or PACs, who scored highest in betweenness and in-degree centrality, performed considerably better than expected during the elections. Given the increased eigenvector scores of pro-Trump candidates, it is possible that their popularity, although evidently overestimated leading up to the Midterms, may not have dipped as many pundits on the Left have hoped. Overall, we can conclude that those parties and individuals that appear most prominently in social media traffic, whether as network nodes or mentions in tweets, leading up to elections are likely to be the most successful during the elections themselves.