Article

Using Candidates' Tweets to Predict an Election Outcome

Political Research Quarterly
2025, Vol. 78(1) 323–340

DOI: 10.1177/10659129241286827
journals.sagepub.com/home/prq


Francisco Afonso[1], **Paulo Rita**[1], **and Nuno António**[1]

Abstract

Understanding social media's role in political communication is crucial in the evolving media landscape. Motivated by the transformative impact of social media on political engagement and discourse, this research fills an under-explored academic gap, studying the effects of geographic focus—local versus national—in candidates' tweets on U.S. Senate election outcomes. It reveals a modest but significant correlation between the nature of political discourse and election competitiveness. Interestingly, strict adherence to party-centric topics did not significantly influence electoral success. The study assessed the performance of regression and classification models in forecasting election outcomes, with classification models demonstrating superior results. Both models provide a new benchmark for future studies in political communication on social media. These findings bear considerable implications for political practitioners, indicating that election success is not merely guaranteed by echoing party-centric issues or predominantly adopting a national communication scope.

Keywords

machine learning, predictive modeling, topic analysis, political communication, social media political marketing, Twitter

Introduction

The advent of social media has dramatically transformed political engagement and discourse, creating an environment where political actors and constituents interact continuously and publicly (Effing, Van Hillegersberg, and Huibers 2011; Loader and Mercea 2011). While this constant engagement gives politics an air of ephemerality, it also enables the convergence of individuals with similar viewpoints, amplifying their collective political messages. Motivated by a desire to comprehend how such synchronous communication impacts election outcomes, this study asks: Does the content and scope of social media political discourse indeed influence voter behavior and, consequently, election results?

Considerable research has been dedicated to disentangling the intricate dynamics between politics and citizens in the online sphere (Brito, Filho, and Adeodato 2021; Gayo-Avello 2013). It is recognized that platforms such as Twitter offer an essential channel for candidates and political parties to communicate with voters, mobilize supporters, and shape public opinion (Das et al. 2022).

Existing studies in this field generally fall into two categories. The first type investigates citizens' behavior, exploring their sentiments toward candidates (Grover et al. 2019; Liu et al. 2021) and how they connect and discuss political topics with one another (Barberá et al. 2015; Hanna et al. 2011; Jungherr 2016). The second category scrutinizes the conduct of political candidates, precisely their focus topics (Ausubel 2019; Feezell 2018; Granberg-Rademacker and Parsneau 2021) and the geographic scope of their discourse (Das et al. 2022; Schürmann 2023).

Despite these extensive efforts, there is a significant gap in understanding the influence of geographic scope (local vs. national) and the topical focus of candidates' tweets on election outcomes. This study aims to fill this gap by taking a cue from Schürmann's work highlighting this dynamic during the 2017 and 2021 Federal German Elections. Schürmann (2023) found that candidates from highly competitive districts referenced their districts

[1]NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Lisboa, Portugal

Corresponding Author:
Paulo Rita, NOVA Information Management School (NOVA IMS), Universidade NOVA de Lisboa, Campus de Campolide, Lisboa 1070-312, Portugal.
Email: prita@novaims.unl.pt

almost five times more often than those in non-competitive communities, indicating the potential impact of local-national dynamics in political communication.

Our research targets the 2022 Midterm Elections in the United States, chosen due to the recentness of the data and the relatively untapped nature of this event for such a study. As Americans reportedly spend over 7 hours online daily, with nearly 2 hours on social media platforms (Data Reportal 2022), Twitter, a preferred source for current events (Odabaş 2022a, 2022b), is selected as the study platform for its potential to mirror and influence real-world political events.

Given the ongoing trend of election nationalization (Alemán and Kellam 2008; Carson and Sievert 2018; Hopkins 2018) and the concerted communication strategies of political parties (Barberá et al. 2015; Iyengar and Westwood 2015), we aim to discern whether national and party-wide discussion topics can explain Senate election results. Additionally, we examine whether each race, especially tightly contested ones, has unique topical dynamics. Given this context, the research's question is, "Can we predict who will win the Senate race based on their Twitter content?"

The study begins with a review of the literature on the interplay between Twitter and politics, followed by an explanation of the research methodology, a presentation of results, and a discussion of the findings. The study aims to enrich the understanding of political communication on social media and its influence on election outcomes.

Literature Review

Twitter and Politics

Social Media platforms have risen since the early 2010s, becoming an integrated part of everyone's life, with 59.3 percent of the world population using these platforms daily, which means 94 percent of internet users use social media. In the United States, 75 percent of its population actively uses Social Media, spending around 2 hours daily on these platforms (Data Reportal 2022).

Twitter is the favorite social media platform for getting the news and keeping up with trending topics, and this is a new channel that politicians are trying to explore to convey their agenda. Accounting that 97 percent of the created content in the United States is produced by only 25 percent of its users (Odabaş 2022b), Twitter is a place where most users come to read and reply rather than create (Mitchell 2022; Odabaş 2022a), with 4 out of 10 American adults saying that Twitter increased their understanding of current events (Odabaş 2022a, 2022b). In the aftermath of major societal events, political activity on Twitter can get 14 times the volume of tweets and retweets than an average day (Shah and Bestvater 2022). Participation in political discussions on social media and using these platforms for political purposes is one of the strongest indications of political participation (Dimitrova et al. 2014; Rita, António, and Afonso 2023).

With a central role in the spread of information from citizens and candidates during campaigns, Twitter has become a source of information for contemporary political communication studies, primarily focusing on one of two aspects: (1) How politicians behave on these platforms to influence citizens and their participation, particularly during campaigns (Ausubel 2019; Feezell 2018; Granberg-Rademacker and Parsneau 2021); (2) How citizens engage in political conversation and share their views (Barberá et al. 2015; Grover et al. 2019; Hanna et al. 2011; Jungherr 2016; Liu et al. 2021). These research areas ultimately try to explain how these actors' dynamics can explain or predict election outcomes.

Midterm Elections Candidates and Twitter. First, the scope (local or national) of political discourse differs according to the type of election or political affiliation. According to Das et al. (2022), congresspeople and governors have a more nationalized focus than mayors, which is congruent with the increasing nationalization of American elections (Alemán and Kellam 2008; Carson and Sievert 2018; Hopkins 2018), especially given how polarized the American political environment is (Zingher and Richman 2019).

Despite this increasing nationalization, there is still some concern with local issues, with candidates discussing local issues such as infrastructure projects in their districts (Bode and Dalrymple 2016, 140). Schürmann (2023) found that in both the 2017 and 2021 Federal German Elections, candidates in highly competitive districts, with a marginality between winners and top-runners under 5 percent, referred "almost five times more often to their districts than candidates in non-competitive districts" (Schürmann 2023). Although German Federal Elections have similar dynamics to American Midterm Elections, several national parties in Germany might explain the contradiction with the nationalization of elections in the United States, in which Republican and Democratic parties mostly dominate.

H1a: Election winners tend to focus more on national issues when running in loose races.
H1b: Election winners tend to focus more on local issues when running in tight races.

Regarding spoken topics, there is also a difference between parties, with Republicans and Democrats speaking about different issues or giving different emphases to the same problem. According to Ausubel

(2019), in a study regarding U.S. House of Representative candidates' tweets during the 2018 midterm elections, Democrats focused on gun violence and healthcare, while Republican candidates focused more on Illegal Immigration.

When the topics referred to are the same, candidates from different parties use different diction or emphasis to refer to those topics (Ausubel 2019). For example, regarding Donald Trump's Tax Policy, while Democratic candidates tend to use the expression "Tax Cut," Republican candidates use "Tax Reform." The same happened with healthcare, more specifically, the Affordable Care Act. While Republican candidates refer to "Obamacare," Democratic candidates refer to the "Affordable Care Act." This tendency to prioritize various topics according to party identification (Goggin, Henderson, and Theodoridis 2016) is consistent with more significant trends in political communication and voting behavior (Abramowitz and Saunders 1998).

This relationship between partisanship and topic ownership can also explain the increasing nationalization of elections, with Americans caring "less about the specific person who represents them and more about the partisan balance of power in Congress" (Das et al. 2022).

H2: Election winners tend to Tweet in line with their party counterparts.

Citizens and Twitter. Research indicates that Twitter plays a pivotal role in political discussions and public opinion formation. It acts as a venue for political conversation and mobilization (Barberá et al. 2015; Dimitrova et al. 2014; Hanna et al. 2011). It can facilitate political participation but also magnify political divisiveness and misinformation.

Studies have confirmed an increased tendency towards homophily on Twitter, where users connect with those sharing similar habits, lifestyles, or moral views (Hong and Kim 2016; Vaisey and Lizardo 2010). This tendency fosters an "us vs. them" posture, leading to echo chambers of polarized ideas (Benton 2022; Gruzd and Roy 2014; Lee 2007; Yardi and boyd, 2010). This bias is more pronounced among Republicans (Iyengar and Westwood 2015; Jost et al. 2018), who often share negative content about left-leaning ideologies (Barberá et al. 2015; Shah and Bestvater 2022).

Despite the echo chamber effect, Twitter also fosters cross-ideological communication, primarily through hashtags and mentions (Conover et al. 2011; Jungherr 2016; Stieglitz and Dang-Xuan 2013). However, this exposure to differing views does not appear to diminish polarization (Bail et al. 2018). Left and right-leaning users form separate clusters but often discuss the same topics using different terms (Shah and Bestvater 2022).

The rise of social media has also facilitated the spread of fake news and the use of bot accounts to propagate misinformation, potentially impacting political involvement and election outcomes (Allcott and Gentzkow 2017; Del Vicario et al. 2016; Guess, Nyhan, and Reifler 2020; Howard, Woolley, and Ryan 2018). While this might impede bipartisan conversations and agreements (Sunstein 2017), it can also reinforce the nationalization of elections and the division of topics between dominant parties.

Why Politicians Use Social Media

Several studies have explored why politicians increasingly use social media platforms like Twitter. Hong, Choi, and Kim (2019) found that extremists, underdogs, and opposing parties are more likely to use Twitter due to its ability to reach a wider audience without the gatekeeping barriers of traditional media. This particularly benefits those who need more visibility or want to challenge the mainstream narrative.

Moreover, politicians use social media to engage directly with their constituents, bypassing traditional media channels and providing real-time updates and responses. This direct engagement can foster a sense of connection and immediacy, which is particularly valuable in modern political campaigns (Hong, Choi, and Kim 2019).

While extensive research has been conducted on the role of social media in political communication, several limitations persist. Many studies focus on a single platform, primarily Twitter, which may need to fully capture the complexity of political communication across different social media channels. Additionally, longitudinal data is often needed, making it difficult to assess the long-term impact of social media on political outcomes (Kubin & von Sikorski 2021).

Methodology

Knowing that there is evidence showing that elections in the United States are getting more nationalized (Alemán and Kellam 2008; Carson and Sievert 2018; Hopkins 2018), and both Democratic and Republican parties tend to focus on different issues (Goggin, Henderson, and Theodoridis 2016), this study sought to prove that candidates who follow their party's topics win the election, being it the H2 Hypothesis. We believe that loose races have a more nationalized speech than tight ones, following the findings of Schürmann (2023), and formulated both H1a and H1b. The following study design framework was developed based on the three hypotheses, as demonstrated in Figure 1. It will guide the hypotheses testing and answer our main goal of creating a model to

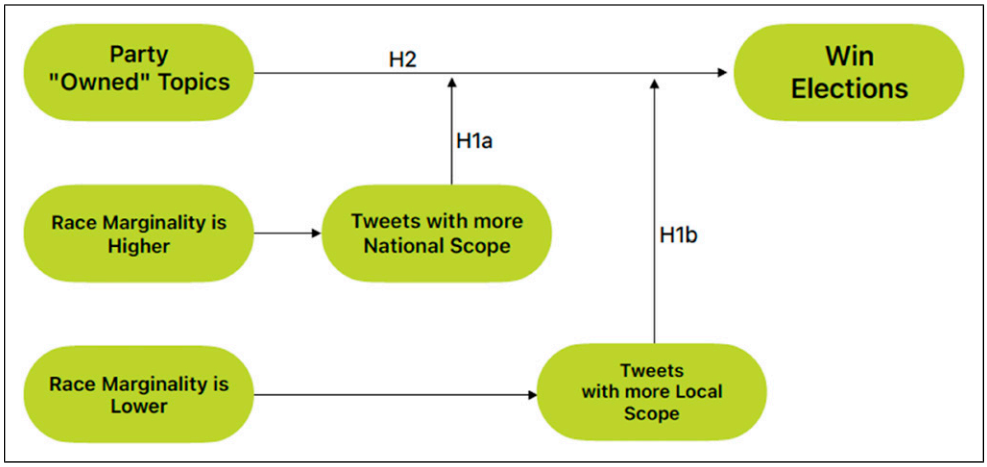

Figure 1. Study design framework.

predict election outcomes based on candidates' tweets during the campaign.

This research followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Chapman et al. 2000), a well-known and proven methodology for text mining and machine learning projects. The CRISP-DM framework consists of six steps that guide professionals and researchers through the whole project, keeping in mind that this kind of work follows an iteration approach, not a linear one.

This study uses secondary data from Twitter's public application programming interface (API) and other election-related sources. To study if candidates' tweets topics and scope can predict the 2022 midterm election result, three data sources were collected and constructed, as explained in the data understanding and data preparation parts of the methodology: one with candidates' tweets from March 1, 2022, till November 7, 2022, the date of the first primary election for this run ("2022 Midterm Election Calendar - 270toWin" 2022) and the day before of the polls, respectively. For example, although the elections for senator in the State of Georgia had a second round, only the tweets until November 7, 2022, were considered, not to introduce bias into the topic discovery; the second data source has election results, with name, state, percentage of votes, affiliated party and if the candidate won the election; the third data source is made of an extensive list of counties, cities and other places across all states, so it can be checked if a candidate makes tweets related with her State.

After data had been collected, it was pre-processed, and then different topic modeling models were applied, splitting this analysis between tweets containing local references and those not. Each candidate's average distribution of topics was calculated, and these variables were considered independent. Several classification and regression models were run to test the hypotheses and evaluate the results.

In analyzing the dependent variable, we focused on the distribution proportion of tweets across topics rather than the sheer number of tweets. This approach allows us to capture candidates' relative emphasis on different issues within their overall social media strategy. A candidate might tweet frequently about local or party-owned issues, but these could still constitute a smaller proportion of their tweets. By examining the proportion of tweets, we can better understand the strategic allocation of attention across various topics.

We also categorized tweets as "national" if they were not explicitly coded as "local," ensuring a clear distinction between the two categories. This method provides a consistent framework for analyzing the focus of candidates' social media content.

Project Understanding

The U.S. Congress comprises two chambers: The House of Representatives, composed of 435 congresspeople, and the Senate, composed of 100 Senators, two from each state ("U.S. Senate" 2023). Federal elections happen every 2 years, and all the 435 seats of the House of Representatives are up for election. In contrast, 33 or 34 of the Senate seats are generally disputed since some Senators serve a 6-year mandate, split into three different classes of seats, making up about a third of the Senate ballot every 2 years. In the 2022 Federal elections, due to the resignment of Oklahoma's senator Jim Inhofe (Ballotpedia 2023b) and the election of California's senator Kamala Harris as Vice President of the United States (Ballotpedia 2023a), there were 35 Senate seats up for election. Due to election rules, a second

round took place in the Georgia state senate election, which was held on the 6th of December.

Since there was one seat for election in each state for the Senate, this race can be seen as a direct dispute between candidates instead of multiple open seats for the House of Representatives in each state. For this reason, was studied the senate election where direct winners and losers for each race can be identified. An overview of all candidates' results can be seen in Appendix A.

Data Understanding

Twitter Data. Utilizing the advanced capabilities of Twitter's public API, data was extracted via a research account, allowing access to tweets from any period. Specific queries were designed to retrieve data based on the date range and the tweet author's handle, with a clear focus on original content generated by campaign pages.

Campaign Twitter handles were located on Ballotpedia, providing 81 valid accounts out of 98 candidates participating in 35 elections. Without a campaign account, personal or official Twitter accounts were utilized. Unfortunately, no Twitter account could be found for 17 candidates, often affiliated with parties other than Democrats or Republicans.

With the necessary handles identified, a specially developed script connected with Twitter's API to retrieve all the original content produced by these users from March 1, 2022, to November 7, 2022. March 1, 2022, was selected to ensure uniformity and a comprehensive analysis period. However, tweets during intraparty competition could confound the analysis by reflecting intraparty dynamics rather than general election strategies. These early tweets might include messaging tailored to party members rather than the broader electorate. Future research could enhance the precision of analysis by considering data from the end of the primary season or when candidates become presumptive nominees. Character limitations of Twitter's query led to a split extraction process. The extraction resulted in 61,425 tweets, with the Hawaiian candidate Brian Schatz not generating any content in the analyzed timeframe. This way, there is an initial data frame with 61,425 rows, one for each tweet, and 15 different columns, as seen in Table 1.

This data revealed some interesting findings. First, most accounts were created before 2022, so they were already used as personal accounts or for previous campaigns with an existing follower base. The same trend can be seen in Figure 2, with many accounts having a more significant share of tweets before March 2022, revealing they had activity before the campaign.

Focusing on the campaign timeframe, as is seen in Figure 3, of the 81 candidates with an existing Twitter account, 31 generated at most 530 tweets, and 30 generated between 530 and 1,060 tweets. There are also differences in the number of tweets generated by party, with the Democratic party generating 28,198 (46 percent) Tweets, the Republican party 19,751 (32 percent), the Liberal party 9,101 (15 percent), and independent candidates generating the rest 7 percent of tweets. This difference in Tweets generated by democrats is consistent with findings that the Twitter user base is left-leaned (Brito, Silva Filho, and Adeodato 2022), giving more incentives to Democrat candidates to share content. Finally, Figure 4 shows us that with the evolution of the campaign, candidates slowly increased communication volume, with the week before the election seeing a considerable increase in this volume.

The data retrieved from Twitter shows that candidates already used Twitter as a communication channel and that the communication volume increases when the campaign nears the end. This reflects that candidates use Twitter strategically to complement other online and offline communication channels (Grover et al. 2019). This data might be a good source to generalize candidates' behavior and policy concerns.

Election Results and United States Cities List. A data source of election results must be assembled to validate whether this Twitter data can predict election outcomes. To do this, Ballotpedia was consulted on February 21, 2023, and an Excel file with two different sheets was developed: "Results by State" and "Results by Candidate." "Results by State" comprises 35 rows, one for each senate election, as seen in Appendix B. "Results by Candidate" has 98 rows, one for each candidate with at least 1 percent of the voting share. Only the candidates and results from the last round for rank-choice voting systems with several rounds were considered.

Besides the results, there is a need for an extensive list of cities and counties in the United States. A list with 63,210 city aliases was available on GitHub,[1] which is comprehensive enough for the task. This data source comprises 63,210 rows, one for each "city alias," and five columns, with the State Name and Code, the County, City, and its alias.

Data Preparation – Original Tweets

Data preparation, modeling, and evaluation phases can be split into two parts: First, data preparation to model the topics and evaluate the resultant topics, and then the preparation of the consequent topics to model and test the hypotheses and research question and check those test results.

Table 1. Description of Data Extracted Through Twitter's API

Column Name	Description
Tweet_id	Internal id of each tweet
Text	Text contained in each tweet
Author_id	Internal id of the tweet's author
Created_at	Timestamp of tweet's creation
Public_metrics.retweet_count	How many retweets did each tweet have on February 26, 2023
Public_metrics.reply_count	How many replies did each tweet have on February 26, 2023
Public_metrics.like_count	How many likes did each tweet have on February 26, 2023
Public_metrics.quote_count	How many quotations did each tweet have on February 26, 2023
User.id	Internal id of the poster of the tweet. In this case, since all tweets are original, it is the same as author_id
User.username	Handle of each Twitter account
User.created_at	Timestamp of users' creation

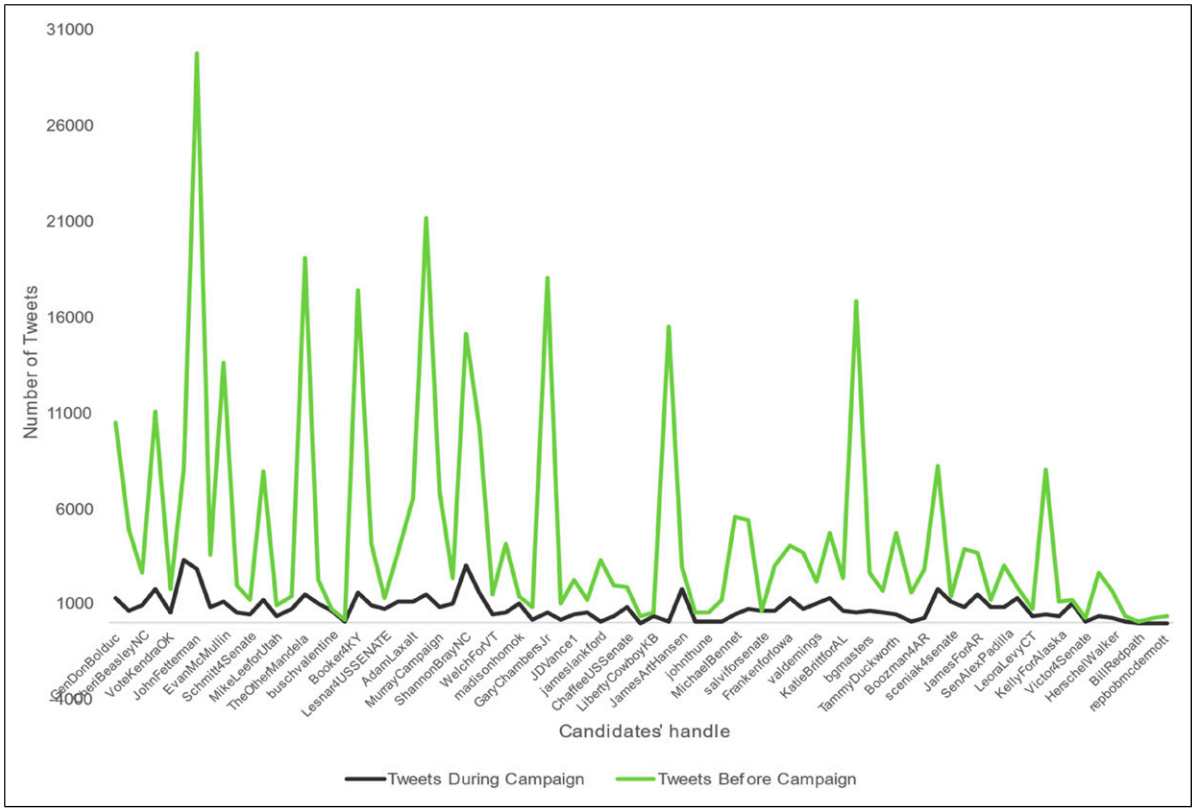

Figure 2. Distribution of tweets before and during campaign.

The data preparation step for the original Tweets dataset starts with Data Cleaning and then Data preprocessing.

Data Cleaning and Preprocessing. Each tweet was thoroughly cleaned, removing URLs, mentions, emojis, punctuation, numbers, line breaks, and other special characters. Tweets with less than two words were eliminated to reduce the noise in the dataset.

A specific tool from the polyglot.detect package (Al-Rfou 2015) was deployed to filter tweets by language, explicitly retaining those in English. However, the tool could not determine the language of a certain number of tweets (1,234 out of 58,447).

The next phase involved consolidating several data files, using common attributes to merge them into a comprehensive dataset. Ambiguity in city names

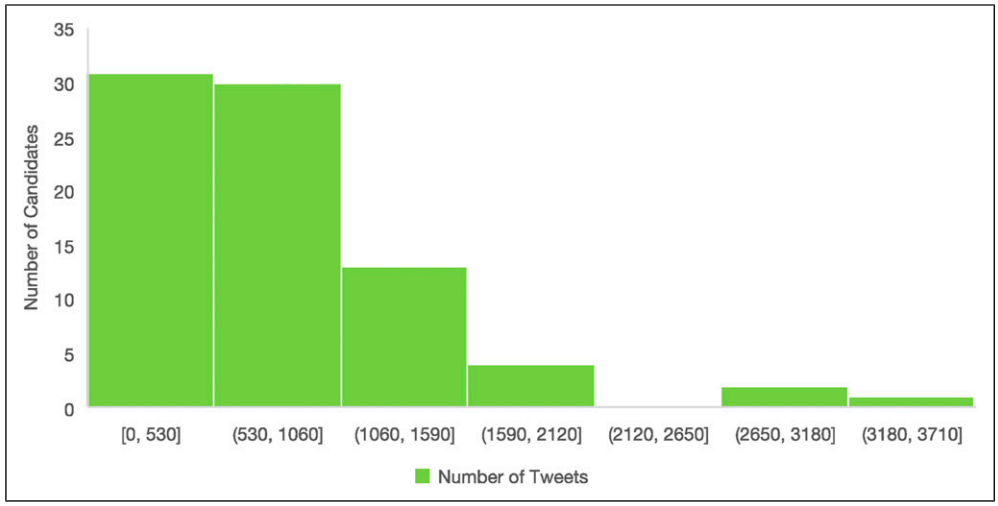

Figure 3. Generated tweets by candidate.

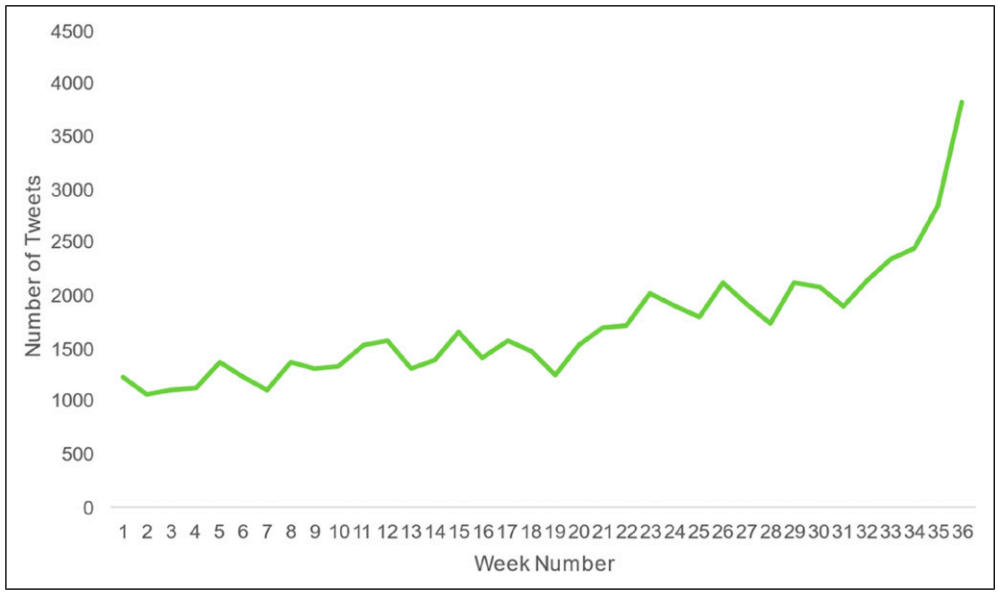

Figure 4. Evolution of Tweets on the 36 weeks of the campaign.

within the tweets necessitated additional disambiguation, which was accomplished using Named Entity Recognition (NER), a feature of the Spacy package (Honnibal and Montani 2021). This ensured that city names were indeed references to geographic locations.

Further refinements were conducted through manual checks to ensure accurate language labels and city classifications. This process uncovered several mislabeling, leading to a review and correction process. This included removing irrelevant words and refining city mentions. Lastly, the dataset underwent tokenization, lemmatization, and stop word removal using the NLTK Python package (Bird, Klein, and Loper 2009).

Topic Modeling – Original Tweets

Topic modeling, a machine learning technique, uncovers hidden thematic structures in document collections. Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) are common algorithms used for this task.

LDA is the most widely used, assuming each document is a mixture of latent topics with a particular word distribution (Blei, Ng, and Jordan 2003). For instance, Karami et al. (2022) and Das et al. (2022) employed LDA in their studies to analyze Twitter data related to political discussions.

CTM is a variant of LDA that captures topic correlations. It presupposes that each document is a blend of

latent issues and that each word within a document is generated based on a specific probability distribution (Blei and Lafferty 2007). Dybowski and Adämmer (2018) used this model to analyze data on U.S. Presidential speeches. According to Blei and Lafferty (2007), CTM has better predictability than LDA due to its probabilistic correlation between topics.

LDA. The LDA topic model, implemented via Python's gensim package (Řehůřek and Sojka 2010), was executed with varying numbers of expected topics to determine the optimal model. The Coherence Score metric, applicable to both LDA and CTM models, and a subjective evaluation of the resulting topics were used to assess outcomes. With the prerequisites met, an algorithm was executed to evaluate five to 16 subjects and their respective Coherence Scores, as illustrated in Figure 5.

The outcomes presented five coherence scores around 0.52, with the peak result slightly exceeding 0.54. Since nine topics represented the initial significant increase, and 13 topics produced the highest coherence score, further refinement was undertaken through hyper-tuning the algorithm for these topics. This process involved adjusting the alpha and beta values, two hyperparameters of the LDA model, to enhance the coherence score. The optimization yielded the top Coherence Score of 0.5639 for 13 topics, representing an advancement from the initial coherence score of 0.5412.

CTM. The CTM algorithm was developed using the tomotopy package (Heewon 2020), a resource gensim does not provide. This initial model used 13 topics, drawing from the optimal result of the LDA model.

To enable comparison with the LDA model, the topics produced by the CTM model were reconfigured into a format compatible with the gensim.CoherenceModel (Řehůřek and Sojka 2010). This allowed for the calculation of the Coherence Score for each topic cluster. This baseline model yielded a Coherence Score of 0.5978, surpassing the highest LDA result.

Further testing was conducted to find the optimal number of topics, exploring a range from five to 16. The results are presented in Figure 6. Like the LDA test, the best number of topics emerged as 13, with a coherence score of 0.5978.

Topic Modeling Evaluation

As illustrated in Figures 5 and 6, the CTM had the best results overall, and the optimal number of topics was 13. The 10 top words from each topic and the tweets with more weight from each topic were revised. Upon evaluation, it was determined that 13 topics resulted in excessive granularity, causing overlapping topics and words.

Given the research's focus on identifying similarities among candidates, this level of detail could lead to divergent, sparse topics. Consequently, the focus shifted to the topics derived from the nine-topic models. The CTM model did yield superior performance; however, the marginal coherence score difference of 0.0139 between the two models for nine topics, combined with the lower computational demand of the LDA model, suggested a

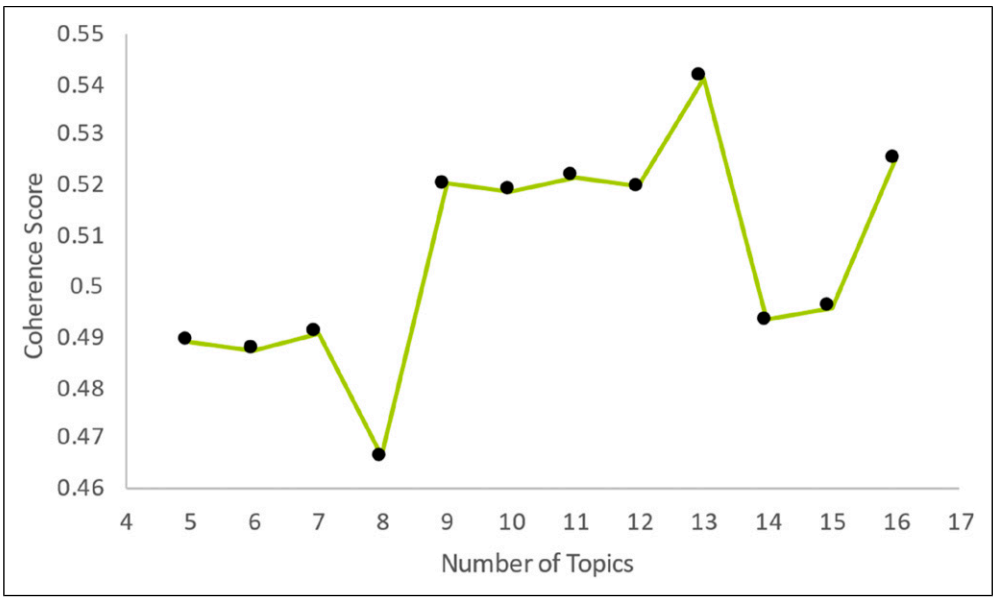

Figure 5. Coherence score for LDA, using 5 to 16 topics.

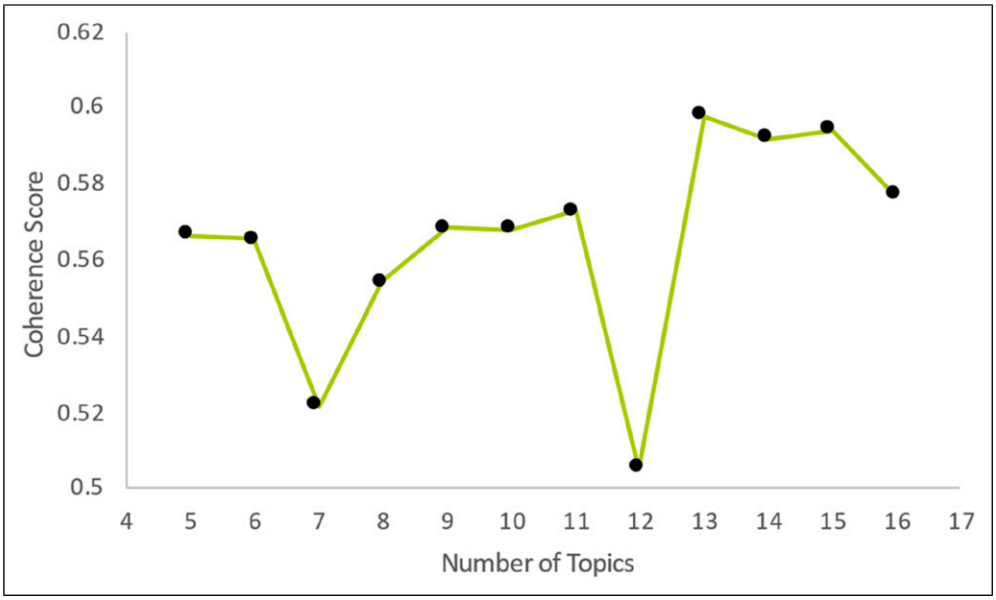

Figure 6. Coherence score for CTM, using 5 to 16 topics.

minimal practical difference, with benefits in speed becoming a factor for choosing LDA.

The LDA model was run using the optimal hyperparameters identified for the nine-topic model. As expected, the resulting topics are like some of the 13 topics from the previous iteration but lost some granularity, with just two topics overlapping and space between most of them, as seen in Figure 7.

Data Preparation and Modeling – Identified Topics

After revising the word distribution and the tweets with more weight from each generated topic, the nine topics were kept and named as in Table 2.

Hypotheses H1a and H1b. The study examined hypotheses H1a and H1b using linear regression (LinR), aiming to understand if winners in tightly contested races emphasized more local issues while those winning by larger margins focused on national topics. Candidate-specific data was aggregated, with all local or national tweets summed and the average winning margin computed. The dataset included only race winners and the proportion of local and national tweets per candidate. It incorporated 34 races, except for candidate Brian Schatz, who had no tweets during the period.

In addition, another dataset was created to scrutinize whether the winning margin influences the focus on national and local issues for winners and losers. This set aggregated the data by state, reflecting each race's normalized distribution of local and national tweets. The dataset contained 35 rows, encompassing all races under consideration.

Hypothesis H2. Hypothesis H2 posits that election winners tend to align their Twitter sentiments with those of their party peers. To validate this claim, the Kruskal–Wallis test was chosen over a one-way analysis of variance (ANOVA) as the data distribution was not normal. This non-parametric test is preferred when the underlying assumptions of the ANOVA, especially normality, are not met.

This hypothesis was tested in two parts. The first part considered the original nine topics and their local or national scope. The second part disregarded the scope and concentrated solely on the nine topics. The testing involved creating topic-weighted columns concerning each tweet's local or national scope. The values were then aggregated by candidate names, calculating the average weight of each topic across their tweets, which would differentiate the winners from the rest.

Three specific Kruskal–Wallis tests were then implemented. The first test sought to identify statistical significance between the topic averages of winners and non-winners, considering the 18 scoped topics. This process was restricted to the Democrat and Republican parties, the only ones having election winners. The second test paralleled the first but focused only on the nine original topics, again considering just the Democrat and Republican parties due to their winners.

Lastly, a third test was conducted to determine whether the mean topic weights for Democratic and Republican winners were statistically significantly different. This analysis concentrated solely on the winners from both major parties, scrutinizing the mean topic weights for each group.

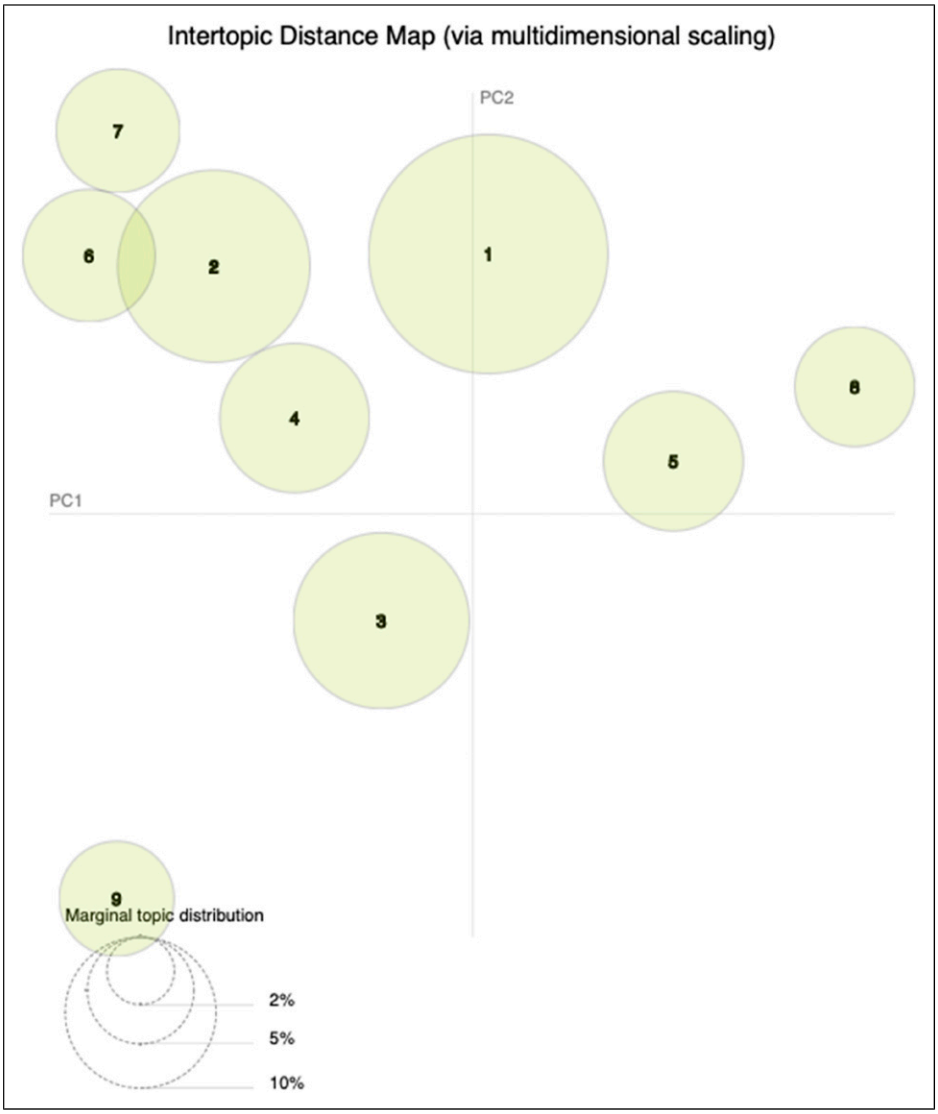

Figure 7. Intertopic distance map for the nine topics, generated by pyLDAvis.

Table 2. Nine Topics Naming.

Topic Number	Topic Name	Top Words
1	Campaign event	Thank, day, great, u, state, support, today, your, work, time
2	Social protection	Cost, $, family, need, job, care, working, for, child, make
3	General discussion	People, like, know, one, would, think, say, party, want, police
4	Abortion	Right, woman, abortion, protect, freedom, life, gun, law, senate, fight
5	Homeland security	Border, war, community, crisis, must, crime, country, secure, Ukraine, open
6	Candidate denigration	Big, oil, trump, interest, Ron, security, social, Johnson, hold, people
7	Vote instigation	Vote, election, day, get, make, ballot, voting, th, voter, November
8	Campaign contribution	Help, u, senate, $, win, race, November, *, today, campaign
9	Economy policies	Biden, American, inflation, energy, joe, democrat, gas, price, policy, %

Research Question. Two distinct approaches were explored to answer the Research Question. The first approach involved predicting the percentage of votes each candidate received, a task treated as a regression problem. The second approach aimed to identify the election winner, viewed as a classification task.

The dataset was divided into two distinct categories. One category encapsulated the 18 composed topics, nine with a national scope and nine with a local scope. The other category consisted of the nine original topics without scope considerations. This differentiation was made to probe whether the local or national coverage affected the predictive models.

Each candidate's average weight across their tweets was calculated for all topics in both categories. The party variable underwent a process known as "dummyfication" through one-hot encoding, which converted the categorical variable into a binary format, effectively creating a new binary variable for each category class. This transformation enables machine learning algorithms to process and understand categorical data efficiently. Finally, all numeric variables in the dataset were normalized.

Various models were utilized to address regression and classification problems, each with advantages and potential limitations (Kursuncu et al. 2019). For regression problems, the target variable was the percentage of votes each candidate received. Initially, the Spearman correlation test was conducted to identify strong correlations between the 18 variables. As no such correlations were identified, all variables were retained. Regression models implemented included linear regression (LinR), decision trees (DT), and neural networks (NN). These models were evaluated based on metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Max Error. The dataset was split throughout this process, with a 75 percent allocation for training and 25 percent for testing. Wherever possible, hyperparameter tuning methods were applied to seek optimal model performance.

In the case of the classification problem, the target variable was the winner of the election. Only the nine original topics were used here due to dimensionality concerns when applying the 18 composed topics. A range of models, including Logistic Regression (LogR), Decision Trees, Support Vector Machines (SVM), K-Nearest Neighbor (K-NN), Neural Networks, and Naïve Bayes (NB), were trained, tested, and evaluated. To balance the number of observations for the binary target variable, the Synthetic Minority Over-sampling Technique (SMOTE) algorithm was applied to tackle the imbalance issue in the dataset. The performance of these models was assessed based on Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC).

Despite a lack of detailed algorithmic disclosure in most election prediction studies using Twitter data (Gayo-Avello 2013), Liu et al.'s (2021) study served as a helpful benchmark. They tested numerous regression and classification models against Twitter data, reflecting Georgia state counties' support of Hillary Clinton in the 2016 Presidential election. Their work highlighted the variable performance of models and the impact of data processing and variable selection. Their classification models, excluding Linear SVC, achieved accuracy rates between 78.72 percent and 82.88 percent. Decision Trees stood out with a higher precision rate. Regarding regression models, K-NN displayed the best RMSE (0.1526), while other models presented comparable outcomes. However, these results can vary depending on the data processing and chosen variables, underscoring the models as benchmarks rather than universally superior solutions.

Results and Discussion

This research divided tests and models based on the hypotheses or research questions, presenting results accordingly.

Focusing on Hypotheses H1a and H1b, an interesting pattern emerged in Table 3's regression analysis. As the race margin shrank, the proportion of local tweets amplified, whereas a boost in national topics correlated with wider margins. However, with a *p*-value of .067 for both tests, these observations did not quite meet the statistical significance threshold of a 5 percent confidence level.

This analysis hints that the shifts in race margin values only capture about 10.1 percent of the changes in tweet focus from local to national and vice versa. Essentially, the tweet's scope seems influenced by other significant factors not considered in this model.

A broader regression analysis was conducted, incorporating all candidates' tweets. The findings in Table 4 showed a *p*-value of .042, meeting the 5 percent significance level. It indicates a significant correlation between the race's competitiveness and a candidate's inclination to focus on local issues.

The results suggest that as races tighten, candidates focus more on local issues in their Twitter communication. This aligns with Schürmann's (2023) research and better explains the dynamics between electoral competitiveness and campaign messaging strategies.

Based on the nationalization trend noted by Das et al. (2022), the second hypothesis proposed was that election winners' tweets would align with those of their party counterparts. Three Kruskal–Wallis tests examined this, which revealed various outcomes.

The first test investigated the variance in tweets across 18 topics. When local tweets were included, both parties

showed significant results on Candidate Denigration and Economy Policies. Excluding local tweets, significance remained only on Economic Policies for both parties. The Democratic party showed significance in more topics, including Abortion, Economic Policies, General Discussion, Social Protection, and Homeland Security. The Republicans showed significance only in the Campaign Event and Vote Instigation topics.

The second Kruskal–Wallis test examined variances within the nine original topics. Differences were evident between both parties on various topics, but no significant disparities were observed in the Campaign Contribution topic for either party.

The tests showed significant differences in winners' and losers' tweets, suggesting varying communication patterns across topics, which contradicts hypothesis H2. The details in Table 5 show that while Democrats emphasized Abortion, their internal tweeting styles varied significantly. The divergence is more prominent among Democrats than Republicans, who display fewer internal disparities. For instance, the Economy Policy topic, traditionally "owned" by the party, shows no significant variation within the party. These outcomes align with the findings of Ausubel (2019) and Goggin, Henderson, and Theodoridis (2016), confirming some degree of topic ownership during the 2022 Senate Elections.

In the context of this study's primary research question, the results illustrated in Table 6 provide compelling insights into the potential predictability of election outcomes using content from both local and national tweets. These results reflect the performance of three distinct predictive models: Linear Regression, Decision Tree, and Neural Network, on both training and testing datasets.

The R^2 values in Table 6 suggest the predictive potential of all models on training data, with Linear Regression achieving the highest value. However, lower R^2 values on test data indicate overfitting. Error metrics reveal a pattern of higher training errors that significantly reduce test data for both Linear Regression and Decision Tree models. Although showing high training errors, Neural Network displays better prediction accuracy on

Table 3. Regression Results for Impact of Tight Races on Local or National Focus From Winners.

	National Tweets (1)	Local Tweets (2)
Intercept	−0.473 (0.299)	0.473 (0.299)
Winner margin for runners-up	2.492* (1.314)	−2.492* (1.314)

Note. *p < .1; **p < .05; ***p < .01.

Table 4. Regression Results for the Impact of Tight Races on Local or National Focus From Every Candidate.

	National Tweets (1)	Local Tweets (2)
Intercept	−0.502 (0.286)	0.502 (0.286)
Winner margin for runners-up	2.517* (1.187)	−2.517* (1.187)

Note. *p < .1; **p < .05; ***p < .01.

Table 5. Kruskal–Wallis Results for Democrats and Republican Winners for the Nine Original Topics.

Topic	p-Value	Democrats Mean	Republicans Mean
Abortion	<.001***	0.144	0.070
Campaign contribution	.086*	0.061	0.042
Campaign event	.036**	0.275	0.352
Candidate denigration	.327	0.055	0.047
Economy policies	<.001***	0.040	0.133
General discussion	.027**	0.060	0.090
Homeland security	.007***	0.049	0.077
Social protection	<.001***	0.246	0.119
Vote instigation	.972	0.071	0.072

Note. *p < .1; **p < .05; ***p < .01.

Table 6. Results for Regression With 18 Topics.

	R^2	MAX Error	RMSE	MAE
Train_LinR	0.843	0.316	0.120	0.097
Test_LinR	0.397	0.353	0.158	0.122
Train_DT	0.753	0.378	0.143	0.108
Test_DT	0.322	0.686	0.219	0.151
Train_NN	0.805	0.463	0.134	0.104
Train_NN	0.530	0.299	0.139	0.110

Table 7. Results for Regression With Nine Topics.

	R^2	MAX Error	RMSE	MAE
Train_LinR	0.796	0.411	0.137	0.106
Test_LinR	0.533	0.295	0.139	0.112
Train_DT	0.823	0.354	0.121	0.089
Test_DT	0.340	0.631	0.216	0.154
Train_NN	0.776	0.486	0.144	0.115
Train_NN	0.415	0.392	0.156	0.118

unseen data, demonstrating the most consistent performance despite signs of overfitting.

Table 7 reveals differences in regression results between the original topics and the mix of local and national tweets. Across three models—Linear Regression, Decision Tree, and Neural Network—all showed substantial performance drops from training to test data, indicating a prevalent issue of overfitting. This pattern was slightly less prominent in Linear Regression, suggesting that incorporating diverse tweets might increase overfitting. Despite this, prediction accuracy declined substantially on unseen data for all models. Error measures indicated similar prediction error patterns across training and test sets.

Compared with Liu et al. (2021), these models performed similarly but showed more overfitting. The best model (LinR with nine topics) displayed slightly better RMSE in training, yet it fell short in the testing set, revealing an inability to generalize to new data.

In the classification tasks, the NN and DT models were the top performers, with an accuracy of about 0.8 on the test set and fewer false positives, as seen in Table 8. The LogR and SVM models had a high recall on the training set, but the SVM performance dropped on the test data, implying lower generalization. The classification models outperformed the regression models, suggesting that predicting the winning candidate might be easier than estimating the exact vote share. These results align with the ones from Burnap et al.'s (2016) study, where they could predict the top parties in voting share but could not correctly predict the distribution of seats in the English parliament.

The models perform robustly when pitted against Liu et al.'s research benchmarks. Liu et al.'s study yielded an accuracy of 58.53 for the Linear SVC, and their other classification models registered accuracy values ranging between 78.72 and 82.88. The accuracy rates for LR, NN, and DT models on the testing set fall comfortably within this range.

The study's findings indicate that the predictive power of local versus national tweets is modest, reflecting the complexity of voter behavior and the multifaceted nature of electoral decisions. Voter behavior is influenced by

Table 8. Results for Classification With Nine Topics.

	Accuracy	Precision	Recall	F1 Score	AUC
Train_LogR	0.85	0.79	0.97	0.87	0.85
Test_LogR	0.80	0.73	0.89	0.8	0.81
Train_SVM	0.85	0.79	0.97	0.87	0.85
Test_SVM	0.65	0.57	0.89	0.7	0.67
Train_NN	0.87	0.80	0.97	0.88	0.87
Test_NN	0.80	0.78	0.78	0.78	0.80
Train_NB	0.65	0.59	1.00	0.74	0.65
Test_NB	0.60	0.53	1.00	0.69	0.64
Train_DT	0.81	0.92	0.68	0.78	0.81
Test_DT	0.80	0.78	0.78	0.78	0.80

various factors beyond social media content, including personal beliefs, economic conditions, historical voting patterns, and offline campaign activities. While tweets provide a snapshot of candidates' communication strategies, they only capture part of the spectrum of voter considerations, which may explain the limited predictive power observed.

Campaigns are keenly aware of their initial standing—whether they are underdogs, favorites, or in tight races—and they tailor their messaging accordingly. This preexisting knowledge influences the content and tone of their Twitter feeds, creating a bidirectional relationship between social media strategy and electoral outcomes. For instance, campaigns that start as underdogs might adopt more aggressive or innovative social media strategies to gain traction and visibility. At the same time, favorites might focus on maintaining their lead and reinforcing their core messages. This correlation between prior electoral performance and current strategies underscores the complexity of causality in political communication.

Another factor contributing to the modest predictive power is the presence of echo chambers and filter bubbles on social media platforms like Twitter. These platforms often expose users to content that reinforces their preexisting views, potentially diluting the impact of national or local issues discussed in tweets. As individuals within these echo chambers are likely to have entrenched opinions, the predictive power of tweets may be

constrained by the homogeneous nature of the audience engaging with them.

Furthermore, the interaction effects between local and national issues are likely more complex than a simple dichotomy. Candidates often discuss national issues within a local context or vice versa, and understanding these interaction effects and their resonance with different voter segments could provide deeper insights into the predictive power of tweets.

Conclusions and Future Work

This research explored the predictive potential of U.S. Senate race outcomes based on the scope and content of candidates' tweets, guided by three core hypotheses: H1a and H1b postulated a correlation between the local versus national nature of political speech and the competitiveness of the election, and H2 proposed that candidates focusing on their party's primary issues had increased chances of winning.

Partial support was found for H1a and H1b, as the scope of political speech explained approximately 11 percent of the variance in election outcomes. However, H2, suggesting a candidate's alignment with the party's primary issues as a key predictor, did not find significant backing.

Further exploration of the predictive power of regression and classification models applied to the content extracted from candidates' local and national tweets yielded intriguing results. While the models showed considerable predictive capabilities on training data, their performance faced challenges on unseen test data due to overfitting. A similar pattern was seen when models were built on the original nine topics, though with a less pronounced performance drop, especially in the Linear Regression model.

The comparison with previous research, such as Liu et al. (2021), revealed comparable performance with regression models but a higher propensity for overfitting. Despite their substantial predictive potential on training data, these models' effectiveness on new, unseen data could have been more robust.

In contrast, classification models demonstrated stronger resilience, displaying a superior ability to generalize from training to test data, thus amplifying their potential to discern election winners based on party and topic weights. Notably, the Neural Network model exhibited consistent, strong performance across training and testing datasets, affirming its capability to identify a candidate's winning potential, even if it did not predict precise vote percentages.

Surprisingly, no additional predictive power was discerned from segmenting tweets into local and national scopes. These findings contribute new perspectives to academic discourse by incorporating topical analysis in local and national contexts, enriching previous research primarily on sentiment analysis and relational aspects.

Research Contributions and Managerial Implications

This study contributes to the existing body of literature by addressing several unexplored aspects. While previous research has delved into the role of sentiments and relationships in election predictions, this work broadens this by incorporating topic analysis within local and national contexts. Significantly, the findings unveiled that the scope of topics did not substantially sway the predictive capacity of the models, be it national or local, contrary to the initial expectations. Moreover, it serves as a valuable benchmark for future investigations into the predictive power of social media content in political elections, especially given the scarcity of existing benchmarks. By demonstrating the ability of topical analysis in classification models and debunking the assumed importance of discourse scope in predicting election outcomes, this study paves the way for further innovative research at the intersection of social media and electoral politics.

The findings from this study carry important practical implications, providing critical insights for candidates and campaign managers. They indicate that a strategy solely focused on party-based issues or centered predominantly around national topics does not necessarily secure electoral victory. These results illuminate the complex role social media platforms play within political dialogues, illustrating the intricacies of the digital political landscape. For future research, the interpretability of complex models, such as the Neural Network, could be a focal point. Understanding how these models function and make predictions may help practitioners formulate more effective communication strategies, further optimizing the role of social media in political campaigns.

Limitations and Future Research

Despite its contributions, this research was met with some limitations. A key aspect is the volume of tweet data: most users generated fewer than 1,060 tweets during the 36-week campaign period. Given the continuous nature of discourse in today's digital world, more granularities may be needed to create robust topics for analysis. Thus, future studies might benefit from access to larger or more diverse data sets.

One other limitation of our study is the exclusion of incumbency as an independent variable. Incumbency can significantly influence campaign strategies and electoral outcomes, as incumbents often possess greater visibility,

resources, and a stronger focus on national issues. Future research should incorporate incumbency to provide a more nuanced understanding of how it impacts social media strategies and election results. By considering incumbency, subsequent studies can better capture the dynamics between established candidates and challengers, enriching the analysis of partisan issue ownership and campaign communication strategies.

The timing of tweets and the temporal dynamics of topics discussed can significantly impact their influence on voter behavior. For instance, tweets about local issues might gain traction closer to election day, while national issues could dominate earlier in the campaign. Future studies could benefit from incorporating temporal analysis to understand how the timing of tweets influences their predictive power and voter engagement.

Enhancing the predictive accuracy of models analyzing social media data by integrating additional data sources, such as campaign finance records, media coverage, public opinion polls, and offline campaign activities, could provide a more comprehensive understanding of electoral dynamics. Combining these data sources with social media data could lead to more accurate predictions.

Employing more other machine learning techniques, such as deep learning models, and transformers, could also improve predictive accuracy. These models could capture complex patterns and interactions within the data, potentially leading to better performance. Incorporating sentiment analysis and emotion detection in tweets could provide additional layers of information about the electorate's response to candidates' messages, refining predictions further.

Geospatial analysis, which analyzes the geographic distribution of tweet engagement and sentiment, could reveal regional variations in voter behavior, helping tailor predictive models to account for local dynamics more accurately. Conducting longitudinal studies that track social media activity and electoral outcomes over multiple election cycles could provide deeper insights into the evolving role of social media in politics, helping identify trends and changes in voter behavior over time.

While our analysis primarily identifies correlations between social media content and electoral outcomes, it is essential to acknowledge the potential for reciprocal causality. The high correlation between dependent and independent variables suggests that past election outcomes could significantly influence current social media strategies. This dynamic complicates the analysis but cannot be overlooked. To better understand this, future research should examine how previous election results, especially those of candidates from the same party, predict social media content. This approach can provide insights into how campaigns assess their environment, develop strategies, and ultimately succeed or fail. Incorporating this perspective will enrich our understanding of the interplay between social media strategies and electoral performance.

The current research only considered Senate races, typically involving fewer candidates and less tweet volume than House of Representatives races. This limitation inherently introduces complexities due to the distinct structure of Senate races, where clear winners and losers emerge. A broader approach incorporating different race types could offer a richer perspective on the predictive power of social media discourse. Data limitations due to the lack of some Twitter handles for candidates from parties other than Republicans and Democrats may have impacted topic generation and, thus, the depth of analysis.

In terms of the modeling process, employing different topic modeling algorithms, such as the Structural Topic Model (STM), could provide a more nuanced understanding of discourse. Exploring a wider variety of regression algorithms might enhance prediction models, and a mix of variables like socio-economic or demographic attributes could contribute to a more holistic view of political dynamics.

Moreover, the scope of topics deserves further investigation. This study primarily examined the difference between local and national scope. Future work might delve into whether specific issues within these scopes have a more significant impact on election outcomes. Furthermore, integrating network analysis with topic modeling could provide unique insights. Studying how the electorate interacts with each topic and how these interactions shape their voting behavior may predict electoral success.

On a similar note, although this study contributes to establishing a benchmark for future research, the scarcity of comparative benchmarks still needs to be improved. Consequently, future studies should continue to expand this field, perhaps focusing on incorporating data from platforms beyond Twitter, especially those featuring visual content like images or videos.

Finally, improving the interpretability of complex models is a critical step for future work. It would facilitate a better understanding of model predictions and provide more actionable insights for practitioners, aiding them in developing effective communication strategies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UIDB/04152/2020 (DOI: 10.54499/UIDB/04152/2020) - Centro de Investigação em Gestão de Informação (MagIC)/ NOVA IMS).

ORCID iDs

Francisco Afonso https://orcid.org/0009-0007-0836-5724
Paulo Rita https://orcid.org/0000-0001-6050-9958

Data Availability Statement

The data and code supporting the findings of this study are publicly available in Zenodo at https://doi.org/10.5281/zenodo.13738284.

Supplemental Material

Supplemental material for this article is available online.

Note

1. Grammakov, "USA Cities and States," *GitHub*, accessed January 21, 2023, https://github.com/grammakov/USA-cities-and-states.

References

2022 Midterm Election Calendar - 270toWin. 2022. "270toWin. https://www.270towin.com/2022-election-calendar/. Accessed January 12, 2023.

Abramowitz, Alan I., and Kyle L. Saunders. 1998. "Ideological Realignment in the U.S. Electorate." *The Journal of Politics* 60 (3): 634–52. doi:10.2307/2647642.

Al-Rfou, Rami. 2015. "Polyglot: Distributed Computing for NLP. Python Package." *Polyglot*. Accessed February 20, 2023. https://polyglot.readthedocs.io/en/latest/

Alemán, Eduardo, and Marisa Kellam. 2008. "The Nationalization of Electoral Change in the Americas." *Electoral Studies* 27 (2): 193–212. doi:10.1016/j.electstud.2007.10.005.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36. doi:10.1257/jep.31.2.211.

Ausubel, Jacob R. 2019. "Social Media and the 2018 Midterms: An Analysis of House Candidates' Twitter Posts." *College Undergraduate Research Electronic Journal, University of Pennsylvania*, March.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, Michael B. F. Hunzaker, Jaemin Lee, Marcus Mann, Floe Merhout, and Volfovsky Alexander. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences* 115 (37): 9216–21. doi:10.1073/pnas.1804840115.

Ballotpedia. 2023a. "United States Senate Special Election in California, 2022." *Ballotpedia*. Accessed February 20, 2023. https://ballotpedia.org/United_States_Senate_special_election_in_California,_2022

Ballotpedia. 2023b. "United States Senate Special Election in Oklahoma, 2022." *Ballotpedia*. Accessed February 20, 2023. https://ballotpedia.org/United_States_Senate_special_election_in_Oklahoma,_2022

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26 (10): 1531–42. doi:10.1177/0956797615594620.

Benton, Joshua. 2022. "Most People on Twitter Don't Live in Political Echo Chambers—But Mostly Because They Don't Care Enough to Bother Building One." *Nieman Lab*; October 5, 2022. https://www.niemanlab.org/2022/10/most-people-on-twitter-dont-live-in-political-echo-chambers-but-mostly-because-they-dont-care-enough-to-bother-building-one/

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. USA: O'Reilly Media. https://www.nltk.org/book/

Blei, David M., and John D. Lafferty. 2007. "A Correlated Topic Model of Science." *The Annals of Applied Statistics* 1 (1): 17–35. doi:10.1214/07-AOAS114.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Bode, Leticia, and Kajsa E. Dalrymple. 2016. "Politics in 140 Characters or Less: Campaign Communication, Network Interaction, and Political Participation on Twitter." *Journal of Political Marketing* 15 (4): 311–32. doi:10.1080/15377857.2014.959686.

Brito, Kellyton Dos Santos, Rogério Luiz Cardoso Silva Filho, and Paulo Jorge Leitão Adeodato. 2021. "A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions." *IEEE Transactions on Computational Social Systems* 8 (4): 819–43. doi:10.1109/TCSS.2021.3063660.

Brito, Kellyton, Rogério Luiz Cardoso Silva Filho, and Paulo Adeodato. 2022. "Please Stop Trying to Predict Elections Only with Twitter." In Proceedings of the 23rd Annual International Conference on Digital Government Research, Virtual Event Republic of Korea, 15–17 June, 2022, 88–95. doi:10.1145/3543434.3543648.

Burnap, Pete, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. "140 Characters to Victory?: Using Twitter to Predict the UK 2015 General Election." *Electoral Studies* 41: 230–33. doi:10.1016/j.electstud.2015.11.017.

Carson, Jamie, and Joel Sievert. 2018. *Electoral Incentives in Congress*. Ann Arbor, MI: University of Michigan Press. doi:10.3998/mpub.9235728.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. 2000. "CRISP-DM." *The CRISP-DM Consortium*.

Conover, Michael D., Goncalves Bruno, Ratkiewicz Jacob, Alessandro Flammini, and Filippo Menczer. 2011. "Predicting the Political Alignment of Twitter Users." In 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011, 192–99. doi:10.1109/PASSAT/SocialCom.2011.34.

Das, Sanmay, Betsy Sinclair, Steven W. Webster, and Hao Yan. 2022. "All (Mayoral) Politics Is Local?" *The Journal of Politics* 84: 1021-34. doi:10.1086/716945.

Data Reportal. 2022. "The Global State of Digital in October 2022." *DataReportal – Global Digital Insights*. Accessed February 20, 2023. https://datareportal.com/reports/digital-2022-october-global-statshot

Del Vicario, Michela, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. "The Spreading of Misinformation Online." *Proceedings of the National Academy of Sciences* 113 (3): 554–59. doi:10.1073/pnas.1517441113.

Dimitrova, Daniela V., Shehata Adam, Jesper Strömbäck, and Lars W. Nord. 2014. "The Effects of Digital Media on Political Knowledge and Participation in Election Campaigns: Evidence from Panel Data." *Communication Research* 41 (1): 95–118. doi:10.1177/0093650211426004.

Dybowski, Thomas P., and Philipp Adämmer. 2018. "The Economic Effects of U.S. Presidential Tax Communication: Evidence from a Correlated Topic Model." *Journal of Economic Behavior & Organization* 148: 1–19.

Effing, Robin, Jos Van Hillegersberg, and Theo Huibers. 2011. "Social Media and Political Participation: Are Facebook, Twitter and YouTube Democratizing Our Political Systems?" In *Electronic Participation*, edited by Efthimios Tambouris, Ann Macintosh, and Hans De Bruijn, 25–35. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23333-3_3.

Feezell, Jessica T. 2018. "Agenda Setting through Social Media: The Importance of Incidental News Exposure and Social Filtering in the Digital Era." *Political Research Quarterly* 71 (2): 482–94. doi:10.1177/1065912917744895.

Gayo-Avello, Daniel. 2013. "A Meta-Analysis of State-of-the-Art Electoral Prediction from Twitter Data." *Social Science Computer Review* 31 (6): 649–79. doi:10.1177/0894439313493979.

Goggin, Stephen, John A. Henderson, and Alexander Theodoridis. 2016. "What Goes with Red and Blue? Assessing Partisan Cognition through Conjoint Classification Experiments." *SSRN Scholarly Paper*. Rochester, NY. doi:10.2139/ssrn.2852786.

Granberg-Rademacker, J. Scott, and Kevin Parsneau. 2021. "Let's Get Ready to Tweet! An Analysis of Twitter Use by 2018 Senate Candidates." *Congress & the Presidency* 48 (1): 78–100. doi:10.1080/07343469.2020.1728425.

Grover, Purva, Arpan Kumar Kar, Yogesh K. Dwivedi, and Marijn Janssen. 2019. "Polarization and Acculturation in US Election 2016 Outcomes – Can Twitter Analytics Predict Changes in Voting Preferences?" *Technological Forecasting and Social Change* 145 (August): 438–60. doi:10.1016/j.techfore.2018.09.009.

Gruzd, A., and J. Roy. 2014. "Investigating Political Polarization on Twitter: A Canadian Perspective." *Policy & Internet* 6 (1): 28–45. doi:10.1002/1944-2866.POI354.

Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020. "Exposure to Untrustworthy Websites in the 2016 US Election." *Nature Human Behaviour* 4: 472–80. doi:10.1038/s41562-020-0833-x.

Hanna, Alexander, Ben Sayre, Leticia Bode, JungHwan Yang, and Dhavan Shah. 2011. "Mapping the Political Twitterverse: Candidates and Their Followers in the Midterms." *Proceedings of the International AAAI Conference on Web and Social Media* 5 (1): 510–13. doi:10.1609/icwsm.v5i1.14179.

Heewon, Cho. 2020. "Tomotopy. https://bab2min.github.io/tomotopy/v/en/. Accessed April 7, 2023.

Hong, Sounman, and Sun Hyoung Kim. 2016. "Political Polarization on Twitter: Implications for the Use of Social Media in Digital Governments." *SSRN Scholarly Paper*. Rochester, NY. https://papers.ssrn.com/abstract=2771015

Hong, Sounman, Haneul Choi, and Taek Kyu Kim. 2019. "Why Do Politicians Tweet? Extremists, Underdogs, and Opposing Parties as Political Tweeters." *Policy & Internet* 11 (3): 305–23.

Honnibal, Mathew, and Ines Montani. 2021. "spaCy: Industrial-Strength Natural Language Processing in Python. https://spacy.io/. Accessed April 7, 2023.

Hopkins, Daniel J. 2018. *The Increasingly United States: How and Why American Political Behavior Nationalized. Chicago Studies in American Politics*. Chicago: The University of Chicago Press.

Howard, Philip N., Samuel Woolley, and Calo. Ryan. 2018. "Algorithms, Bots, and Political Communication in the US 2016 Election: The Challenge of Automated Political Communication for Election Law and Administration." *Journal of Information Technology & Politics* 15 (2): 81–93. doi:10.1080/19331681.2018.1448735.

Iyengar, Shanto, and Sean J. Westwood. 2015. "Fear and Loathing across Party Lines: New Evidence on Group Polarization." *American Journal of Political Science* 59 (3): 690–707. doi:10.1111/ajps.12152.

Jost, John T., Pablo Barberá, Richard Bonneau, Megan Langer, Michael Metzger, Joshua Nagler, Jonathan Sterling, and Joshua A. Tucker. 2018. "How Social Media Facilitates Political Protest: Information, Motivation, and Social

Networks." *Political Psychology* 39 (S1): 85–118. doi:10.1111/pops.12478.

Jungherr, Andreas. 2016. "Twitter Use in Election Campaigns: A Systematic Literature Review." *Journal of Information Technology & Politics* 13 (1): 72–91. doi:10.1080/19331681.2015.1132401.

Karami, Amir, Spring B. Clark, Anderson Mackenzie, Dorathea Lee, Michael Zhu, Hannah R. Boyajieff, and Bailey Goldschmidt. 2022. :2020 U.S Presidential Election in Swing States: Gender Differences in Twitter Conversations." *Journal of Information Management and Engineering* 4 (1): 40–55.

Kubin, Emily, and Christian von Sikorski. 2021. "The Role of (Social) Media in Political Polarization: A Systematic Review." *Annals of the International Communication Association* 45 (3): 188–206. doi:10.1080/23808985.2021.1976070.

Kursuncu, Ugur, Manas Gaur, Usha Lokala, Krishnaprasad Thirunarayan, Amit Sheth, and I. Budak Arpinar. 2019. "Predictive Analysis on Twitter: Techniques and Applications." In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, edited by Nitin Agarwal, Nima Dokoohaki, and Serpil Tokdemir, 67–104. Lecture Notes in Social Networks. Cham: Springer International Publishing. doi:10.1007/978-3-319-94105-9_4.

Lee, Eun-Ju. 2007. "Deindividuation Effects on Group Polarization in Computer-Mediated Communication: The Role of Group Identification, Public-Self-Awareness, and Perceived Argument Quality." *Journal of Communication* 57 (2): 385–403. doi:10.1111/j.1460-2466.2007.00348.x.

Liu, Ruowei, Xiaobai Yao, Chenxiao Guo, and Wei Xuebin. 2021. "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach." *Annals of GIS* 27 (1): 43–56. doi:10.1080/19475683.2020.1829704.

Loader, Brian D., and Dan Mercea. 2011. "Networking Democracy? Social Media Innovations and Participatory Politics." *Information, Communication & Society* 14 (6): 757–69. doi:10.1080/1369118X.2011.592648.

Mitchell, Travis. 2022. "Survey Findings on Twitter Users' Political Attitudes and Experiences." *Pew Research Center - U.S. Politics & Policy (blog)*; June 16, 2022. Accessed February 20, 2023. https://www.pewresearch.org/politics/2022/06/16/survey-findings-on-twitter-users-political-attitudes-and-experiences/

Odabaş, Meltem. 2022a. "5 Facts About Twitter 'Lurkers'." *Pew Research Center (blog)*. Accessed March 17, 2023. https://www.pewresearch.org/fact-tank/2022/03/16/5-facts-about-twitter-lurkers/

Odabaş, Meltem. 2022b. "10 Facts About Americans and Twitter." *Pew Research Center (blog)*. Accessed March 17, 2023. https://www.pewresearch.org/fact-tank/2022/05/05/10-facts-about-americans-and-twitter/

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. https://radimrehurek.com/gensim/. Accessed April 7, 2023.

Rita, Paulo, Nuno António, and Ana Patrícia Afonso. 2023. "Social Media Discourse and Voting Decisions Influence: Sentiment Analysis in Tweets during an Electoral Period." *Social Network Analysis and Mining* 13 (1): 46. doi:10.1007/s13278-023-01048-1.

Schürmann, L. 2023. "Do Competitive Districts Get More Political Attention? Strategic Use of Geographic Representation during Campaign and Non-Campaign Periods." *Electoral Studies* 81: 102575. doi:10.1016/j.electstud.2022.102575.

Shah, Sono, and Samuel Bestvater. 2022. "As the 2022 Campaign Draws to a Close, Here's How Federal, State and Local Candidates Have Used Twitter." *Pew Research Center (blog)*. https://www.pewresearch.org/fact-tank/2022/11/02/as-the-2022-campaign-draws-to-a-close-heres-how-federal-state-and-local-candidates-have-used-twitter/. Accessed March 17, 2023.

Stieglitz, Stefan, and Linh Dang-Xuan. 2013. "Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior." *Journal of Management Information Systems* 29 (4): 217–48. doi:10.2753/MIS0742-1222290408.

Sunstein, Cass R. 2017. *#Republic: Divided Democracy in the Age of Social Media*. Princeton: Princeton University Press.

U.S. Senate. 2023. "U.S. Senate. https://www.senate.gov/about/origins-foundations/senate-and-constitution.htm. Accessed February 20, 2023.

Vaisey, Stephen, and Omar Lizardo. 2010. "Can Cultural Worldviews Influence Network Composition?" *Social Forces88* 88 (4): 1595–1618. doi:10.1353/sof.2010.0009.

Yardi, Sarita, and Danah Boyd. 2010. "Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter." *Bulletin of Science, Technology & Society* 30 (5): 316–27. doi:10.1177/0270467610380011.

Zingher, Joshua N., and Jesse Richman. 2019. "Polarization and the Nationalization of State Legislative Elections." *American Politics Research* 47 (5): 1036–54. doi:10.1177/1532673X18788050.