Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

A Multi-Platform Collection of Social Media Posts about the 2022 U.S. Midterm Elections

Rachith Aiyappa*[1], Matthew R. DeVerna*[1], Manita Pote*[1], Bao Tran Truong*[1], Wanying Zhao*[2], David Axelrod[1], Aria Pessianzadeh[1], Zoher Kachwala[1], Munjung Kim[2], Ozgur Can Seckin[1], Minsuk Kim[2], Sunny Gandhi[2], Amrutha Manikonda[2], Francesco Pierri[3], Filippo Menczer[1], and Kai-Cheng Yang[1]

[1]Observatory on Social Media, Indiana University, Bloomington, USA
[2]Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, USA
[3]Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Italy
{racball, mdeverna, potem, baotruon, zhaowany, daaxelro, apessian, zkachwal, munjkim, oseckin, mk139, sugandhi, amanikon}@iu.edu, francesco.pierri@polimi.it, {fil, yangkc}@iu.edu

Abstract

Social media are utilized by millions of citizens to discuss important political issues. Politicians use these platforms to connect with the public and broadcast policy positions. Therefore, data from social media has enabled many studies of political discussion. While most analyses are limited to data from individual platforms, people are embedded in a larger information ecosystem spanning multiple social networks. Here we describe and provide access to the Indiana University 2022 U.S. Midterms Multi-Platform Social Media Dataset (*MEIU22*), a collection of social media posts from Twitter, Facebook, Instagram, Reddit, and 4chan. *MEIU22* links to posts about the midterm elections based on a comprehensive list of keywords and tracks the social media accounts of 1,011 candidates from October 1 to December 25, 2022. We also publish the source code of our pipeline to enable similar multi-platform research projects.

Introduction

As social interactions increasingly happen online, social media have become important for political discussions. Most Americans use at least one social media platform throughout the day (Auxier and Anderson 2021), with a sizable proportion of online discussion related to politics (Bestvater et al. 2022), especially during election seasons (Bestvater and Shah 2022). Government officials heavily leverage social media to interact with their constituencies (Shah and Grant 2021). These platforms, therefore, offer a fertile ground for studying public discourse. For instance, some research mines social media data to learn how misinformation spreads during elections (Shao et al. 2018a; Grinberg et al. 2019) and pandemics (Pierri et al. 2023a; Gallotti et al. 2020) and how polarization emerges in digital communities (Waller and Anderson 2021).

Most research on social media focuses on data collected from individual platforms—especially Twitter, owing to its openness in the past. However, this approach can only provide a limited understanding of the larger information ecosystem. As platforms and their user bases differ, results from one social network do not necessarily generalize to others. For instance, previous research suggests that polarization processes exhibit different patterns (Yarchi, Baden, and Kligler-Vilenchik 2021) and that audiences respond differently to similar campaign strategies by the same candidates (Bossetta and Schmøkel 2022) across platforms. Moreover, moderation efforts by individual platforms might not be sufficient to curb the spread of malicious content given the interconnection among social media (Johnson et al. 2019; Velasquez et al. 2021). Linking data from different social media sites can therefore reveal disinformation campaigns sharing content across platforms (Wilson and Starbird 2020; Golovchenko et al. 2020; Pierri et al. 2023b; Yang et al. 2021). As many people use multiple social media (Hardy and Castonguay 2018; Primack et al. 2017), they may be even more vulnerable to these cross-platform influence campaigns.

This evidence highlights the need to analyze multiple platforms simultaneously when studying social media. However, the lack of data hinders such efforts. In some cases, researchers may have to string together data collected at different times, for different purposes, and using different methods (Lukito 2020). Such data may not be comprehensive as it is retrieved retroactively. Further hurdles to multi-platform data analysis arise from the lack of unified data-sharing protocols (Pasquetto et al. 2020). We attempt to address these challenges by providing a topic-consistent dataset with broad coverage from multiple platforms during the same time period.

The *MEIU22* dataset is a collection of posts from Twitter, Facebook, Instagram, Reddit, and 4chan during the 2022 U.S. midterm election season. We focus on the first four because they are among the top ten most popular platforms used by Americans (Auxier and Anderson 2021). In particular, Twitter and Facebook are increasingly being used for political communication (Stier et al. 2018). Although not as mainstream as the others, 4chan underlies far-right extremist movements that might affect election results (Baele, Brace, and Coan 2020).

*These authors contributed equally to this work; ordered alphabetically

To maximize coverage, we first deploy a snowball sampling procedure to identify relevant keywords from multiple platforms in an iterative manner. In addition, we manually compile a list of social media handles for 1,011 candidates with social media presence from any U.S. state. Using this information, we build a data collection workflow that fetches posts continuously or through periodic searches, based on the functionalities of the application programming interface (API) endpoints available from different platforms.

The remainder of this paper presents our system architecture, data collection, sources, and processing. Lastly, we discuss the limitations and potential applications of the data. For example, our dataset will allow researchers to study public discourse around the 2022 U.S. midterm elections. The data can also be used to analyze the information diffusion process and potential manipulation on multiple social networks at once. Furthermore, the social media handles and activities of the candidates included in this dataset allow for in-depth analyses of their public communication strategies.

We provide public access to the dataset as well as the source code of our data collection framework to facilitate replication in other contexts (github.com/osome-iu/MEIU22).

System Architecture

The architecture of our data collection system is illustrated in Figure 1. This infrastructure is hosted by Extreme Science and Engineering Discovery Environment (XSEDE) Jetstream virtual machines (VMs) (Towns et al. 2014; Stewart et al. 2015), which can be replaced with any server that has access to the internet and enough storage space.

We distribute the tasks to two VMs. The data *collection machine* is responsible for collecting the raw data and copying it to the *analysis machine* periodically. The collected data is also backed up to Indiana University's Scholarly Data Archive, a fast tape storage system, to ensure robustness (not illustrated). Depending on the functionalities of the API endpoints, we use different strategies to collect data from different platforms, as explained in detail in the following section. Computationally expensive tasks such as data cleaning and analysis are executed on the *analysis machine*. This machine is also responsible for hosting a dashboard[1] to share insights obtained from the data with the public.

Data Collection

Our dataset consists of two parts: (1) general social media posts discussing election-related issues, and (2) posts published by U.S. congressional candidates. For the first part, we employ a keyword-based data collection approach, using the same keyword list for different platforms to ensure that the data is comparable. For the second part, we compile a list of the social media handles of all the candidates and use it to track their social media activity. In the following, we describe the procedure we adopt to obtain relevant keywords and the midterm candidate list. Additionally, we provide details on how the data is collected on each social media platform.

[1] osome.iu.edu/tools/midterm22

Date of inclusion	Keywords
2022-09-16*	midterm election, 2022 midterm, 2022 election, midterm 2022
2022-09-20*	vote 2022, vote midterm, vote november, midterm november, vote republicans, vote democrat, voteblue, votered
2022-09-30*	november republicans, november democrats, absentee vote, absentee ballot, mail in vote, mail in ballot
2022-10-07	mail ballot, october surprise
2022-11-04	ivoted, red wave, blue wave

Table 1: List of keywords used to collect data from different platforms. Only the basic form of each keyword is shown here. For the full list containing all the variants, please refer to the GitHub repository. Asterisks indicate the initial phase. The term "ivoted" is removed on November 11, our only keyword removal.

Keyword List

We build the keyword list during an initial curation phase between September 16 and September 30, 2022. We employ a snowball sampling procedure (DeVerna et al. 2021; Di Giovanni et al. 2022), as detailed below. This leads to the collection of phrases appearing most often in discussions about the midterm elections. Such a collection is performed separately on each platform (Twitter, Facebook, Instagram, and Reddit), but all the phrases are merged into a single list.

We perform three rounds of the snowball procedure to iteratively build the list of keywords/phrases (see Table 1). We start from a short list containing a number of unambiguous *seed* phrases related to the midterm elections. These seeds are used as queries for the APIs of different platforms. Each snowball round consists of collecting data for multiple days and then identifying the top 50 unigrams and top 50 bigrams (not necessarily consecutively) that co-occur with any of the phrases in the current list, for each platform.

Before being added to the list, potential phrases are manually reviewed by three of the authors (R.A., M.R.D., and K.-C.Y.) for inclusion in the list for the next round. They are included only after considering relevance and precision with respect to capturing discussion related to the U.S. midterm elections. For example, it is common for issue-related phrases to appear within the top-ranked uni/bigrams. Some of these phrases, e.g., "abortion," or "abortion ban," capture general discussions about abortion that are not necessarily connected to the elections, and are therefore excluded. Other issue-related phrases, such as "mail-in vote," lead to very few false positives, and are therefore included. Note that due to the nature of the streaming API, newly added keywords only affect subsequent matching.

Once a phrase enters the keyword list, different variants of it are also added to ensure complete coverage of the semantic meaning of this phrase and because different platforms differ in the way their APIs match keywords (see individual platform-related sections for details). For example, af-

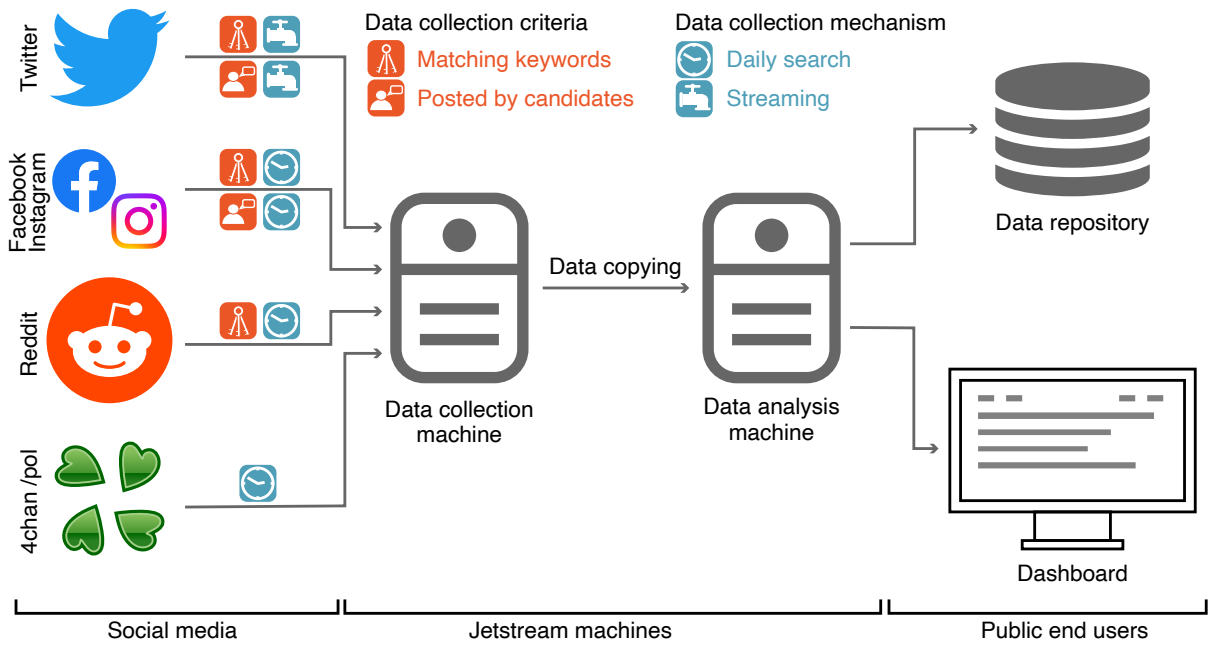

Figure 1: Architecture of the *MEIU22* data collection and analysis system. Data flows in the direction of the arrows.

ter identifying the phrase "november midterm," we also include "midterm november." Similarly, we add plurals of the phrases when they have a proper semantic meaning (e.g., "mail-in vote" is complemented with "mail-in votes").

Data collection from different platforms occurs between October 1 and December 25 using the keyword list. During this period, we repeat the snowball sampling procedure every seven days but with stricter criteria for adding new phrases to keep the list stable. For example, new phrases are only included if they capture stand-out "viral" events that dominate political discussion at the time.

Candidate List

We obtain the social media handles of the U.S. Senate and House election candidates from ballotpedia.org.[2] For each candidate, we collect the following information: name, party affiliation, election type (Senate or House), house candidate district, Twitter handle, Facebook and/or Instagram pages, YouTube channel, and links to the candidate's campaign/official websites. The personal pages of the candidates are not added to the list as we assume they do not contain election-related content. Since Twitter users are allowed to change their usernames, we also extract the unique numerical ID for each handle. We exclude candidates who have already lost in their primary elections, keeping information for those that have advanced to the general elections in November 2022. In total, the list contains information about 4,508 social platform handles from as many as 1,011 candidates running for the 2022 U.S. House and Senate elections. The list is also shared on the GitHub repository.

[2]ballotpedia.org/List_of_congressional_candidates_in_the_2022_elections

Twitter

Twitter data is collected using the `tweepy` Python library, which uses the Twitter V1 filter streaming API endpoint.[3] This endpoint allows us to collect all public tweets containing our keywords. In this process, the texts of the tweets and certain entity fields are considered for matches. These entity fields include hashtags, the expanded and display URLs, and the screen name for user mentions.[4]

We also use the filter streaming API endpoint to collect tweets from the candidates. This collection process did not start until October 24, 2022, so we additionally fetch all posts by these candidates since June 1, 2022 with the user timeline endpoint.[5]

To abide by Twitter's terms of service, we are only allowed to publicly share the tweet IDs of the retrieved tweets. The dataset can be re-hydrated by querying the Twitter API directly or using tools like Hydrator[6] or twarc.[7] Although we use the Twitter V1 API to collect the data, one can alternatively employ the V2 API[8] to re-hydrate the data.

Facebook and Instagram

Facebook and Instagram data is collected from CrowdTangle, a public insights tool owned and operated by Facebook (CrowdTangle Team 2022). Specifically, we query the

[3]developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

[4]developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/basic-stream-parameters

[5]developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline

[6]github.com/DocNow/hydrator

[7]github.com/DocNow/twarc

[8]developer.twitter.com/en/docs/twitter-api

/posts/search endpoint,[9] retrieving posts from *both* Facebook and Instagram simultaneously. These searches are not case-sensitive. As this platform does not provide access to real-time data, we retrieve posts for the previous day every morning (all dates based on Coordinated Universal Time, or UTC).

Facebook posts from candidates are collected using the /posts endpoint[10] that downloads all public candidate pages and groups in a CrowdTangle list. This list, which we manually curate, manages access to all posts from pages and groups with a web-based interface. We again adopt the practice of downloading all posts from the previous day, each morning beginning on October 31. Earlier data since June 6, 2022 (inclusive) was fetched between October 24 and October 29.

Note that CrowdTangle only tracks data from *public* Facebook and Instagram pages and groups, so we do not have complete visibility into the platform activity. This also means that we are unable to track some candidate accounts not indexed by CrowdTangle. We are only allowed to share the URL of collected posts, which can be used to access the data publicly.

Facebook and Instagram Advertisements

We collect information about advertisements on Facebook and Instagram using the Meta Ad library API.[11] This API provides a single endpoint ads_archive to search all ads stored in the Ad library. Queries are made using the same keyword list described above. The Meta Ad library captures text in various data fields, such as text, image, audio, video, and the "call-to-action," of an advertisement. Thus, if any of an advertisement's fields contain any of the matching phrases, data for that advertisement is captured within our dataset. Each day, we collect all the advertisements that are labeled as political or issue-related, and that are delivered in the U.S. Meta implemented a restriction on election-related advertisements between November 1 and 8,[12] leading to a drastic decrease in the data volume (see Figure 2).

According to platform policy, we are allowed to share the raw advertisement data only with researchers or journalists who have a Meta developer account and agree to Meta's platform policy. Here, we share the list of page IDs in which advertisements are displayed. Users can retrieve the advertisements by querying the API with these IDs.

Reddit

Reddit submissions and comments are collected using the Pushshift API (Baumgartner et al. 2020). The API provides two endpoints for retrieving submissions (/reddit/search/submission) and comments (/reddit/search/comment), respectively.[13] We search for all posts that match at least one keyword. Given that Pushshift only provides historical data, we apply the same strategy as for CrowdTangle to retrieve posts from Reddit. There are no sharing restrictions for Reddit data, which is publicly available, and we provide access to the entire collection of submissions and comments.

4chan

4chan is an image-based bulletin board where users can create a thread by posting an image and a message to a board and others can reply to it. For our data collection, we focus on the "Politically Incorrect" board /pol (Papasavva et al. 2020). One of 4chan's features is the ephemerality of the content. Specifically, threads that receive recent replies are bumped to the top of the board, pushing older threads down. The /pol board has limited space (21 pages), and once a thread is pushed out of the board, it enters the archive. Archived threads are static, and can no longer be replied to. They are deleted after a certain amount of time that depends on the speed at which new threads are archived.

We adopt two approaches to collect the data from 4chan's /pol board. In the first case, we hope to understand what people are discussing on this board. So we simply retrieve all the threads from the catalog endpoint[14] every five minutes, starting on October 1, 2022. Based on some preliminary analysis, it takes at least 15 minutes for a thread to be archived, therefore we should have obtained all the original posts this way. But one problem with this approach is that we might miss some replies. For each original post obtained from the catalog endpoint, its five most recent replies are attached as well. However, it is very common for people to reply to this thread more than five times since the last snapshot, so some replies are missed.

Therefore we deploy a second approach to obtain the full threads, including all replies starting October 11, 2022. Here, we leverage the archive endpoint[15] and the thread endpoint.[16] The archive endpoint returns a list of threads in the archive at the moment. By comparing the snapshot of the archive at the current moment with the snapshot at a previous moment (we use 10 minutes in our collection), we can find the threads that have been archived recently. Next, we query the thread endpoint to fetch the whole tree, i.e., the original post and all its replies. Since the archived threads can no longer be updated, the result is final. We make the data collected with both methods available for the public to use.

Data Processing

Cleaning

We notice that the posts obtained through the keyword-matching approach contain a lot of false positives. The issue is mainly due to various award events involving voting (e.g., the American Music Awards) and elections in other countries (e.g., the 2022 Gujarat Legislative Assembly election in India). Twitter suffers more from the first issue, whereas the

[9] github.com/CrowdTangle/API/wiki/Search

[10] github.com/CrowdTangle/API/wiki/Post

[11] facebook.com/ads/library/api

[12] developers.facebook.com/blog/post/2022/09/28/upcoming-restriction-period-for-us-ads

[13] github.com/pushshift/api

[14] a.4cdn.org/pol/catalog.json

[15] a.4cdn.org/pol/archive.json

[16] a.4cdn.org/pol/thread

Platform	Overall	Before Nov. 8	After Nov. 8
Twitter	0.89	0.94	0.85
Meta	0.88	0.96	0.82
Reddit	0.92	0.93	0.92

Table 2: Data quality evaluation results. We report the precision for the whole time period (October 1 – December 25), before election day (October 1 – November 8), and after election day (November 8 – December 25).

second one is more pronounced on Meta and Reddit. Many posts pertaining to these events use the keywords "vote," "election," and "2022," and thus end up in our collection. On some days, the false positives make up over 50% of the posts in the raw collection and we accordingly apply extra data-cleaning procedures.

By analyzing the false positives on Twitter, we find that these tweets often contain hashtags referring to specific events. Take the American Music Awards as an example: most tweets use the hashtag *#AMAs* together with the names of some artists, such as *#BTS*. Therefore, we curate a negative list of hashtags that refer to events irrelevant to the U.S. midterm elections. For tweets from a given day, we extract the 50 most popular hashtags and manually inspect them to identify the irrelevant ones. We then add these hashtags to our negative list and exclude the tweets containing any of the hashtags in it. The procedure is repeated until all 50 most popular hashtags are related to the midterm elections. We repeat this procedure for all days from October 1 to December 25, 2022. In the end, we use the full negative list to filter tweets in the entire collection.

Unlike Twitter, posts from Facebook, Instagram, and Reddit tend to contain fewer hashtags. Therefore, we extract the unigrams (stop words excluded) from the posts for each day, and rank them by their TF-IDF (term frequency–inverse document frequency). We manually inspect the top 100 to identify irrelevant ones and add them to the negative list. Finally, we use the resulting negative list to filter the posts in the entire collection.

Quality Evaluation

To evaluate the quality of the data collected using the keyword-matching approach after cleaning, we perform manual annotation of sampled data. For Twitter, Meta (we mix the Facebook and Instagram posts together since they are obtained through the same API endpoint), and Reddit (we combine the submissions and comments together), we sampled ten posts from each day's collection between October 1 to December 25 (i.e., 860 posts in total for each platform). The authors then manually label each post as a true positive or false positive. The true positives include posts clearly discussing the U.S. midterm elections and those referring to other events but using midterm-related keywords to gain attention. The false positives mainly consist of the posts dedicated to the elections and politics in other countries and those voting artists, movies, or games for awards.

With the annotations, we are able to calculate the precision (the number of false positives divided by the total number of posts) for each platform. The results can be found in Table 2. We also report the precision before and after the election day as references. The results suggest that there is more noise in the post-election period.

Platform	# posts (key.)	# posts (cand.)	# handles
Twitter	6,242,412	638,448	1,237
Facebook	304,106	259,385	1,209
Instagram	35,314	N/A	N/A
Reddit	160,773	N/A	N/A
Meta Ads	5,352	N/A	N/A

Table 3: Summary statistics. We report the number of posts collected using the keyword list, the number of posts from candidates, and the number of candidate handles on different platforms. Note that some candidates have multiple social media handles on a single platform.

Data Volume

In this section, we briefly characterize the data collected. Table 3 summarizes the total number of midterm-related posts from each platform and the number of candidate handles for Twitter and Facebook.

Figure 2 shows the daily volume of posts collected by matching keywords on different platforms. We observe volumes ranging from a peak of around 500 thousand posts per day on Twitter to a few thousand on Instagram (recall that CrowdTangle only provides limited data about public Instagram accounts). Despite the difference in volume, these platforms share similar temporal patterns, with the highest number observed around election day. The case for advertisements on Facebook and Instagram is different, as there are barely any election-related ones after November 1 due to the aforementioned platform policy.

Figure 3 shows the volume of congressional candidates' tweets and Facebook posts, respectively. With retrospective search, the data collection covers the time period from June 1 to December 25, 2022. The time series of data from the two platforms share very similar temporal patterns.

Discussion

Limitations

Despite our efforts to cover as many social media platforms as possible, our dataset does not include platforms with substantial user bases that might play important roles in the current information ecosystem, such as YouTube and TikTok (Auxier and Anderson 2021). Emerging platforms such as Parler and Truth Social, which are alternative destinations when users are banned from major platforms, are also not covered (Stocking et al. 2022). This is mainly because these platforms offer no programmatic methods with which we can obtain data. Even for some of the data we are able to obtain, we are not allowed to share the content directly. To retrieve the information, users will need to go through review processes on different social media platforms to obtain access first. This still poses an obstacle for efforts to study discussions on social media and impedes the progress

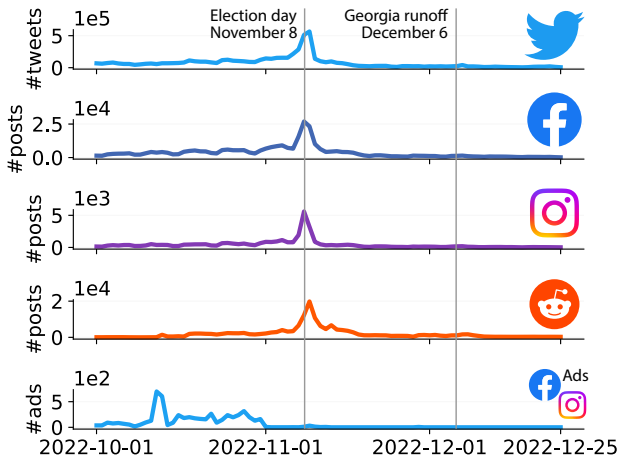

Figure 2: Daily volume of midterm-related posts collected through the keyword-matching approach from each platform. For Reddit, we combine the number of submissions and comments together. For the advertisements, we combine the number on Facebook and Instagram. We annotate the election day, i.e., November 8, and the day of the Georgia runoff, i.e., December 6.

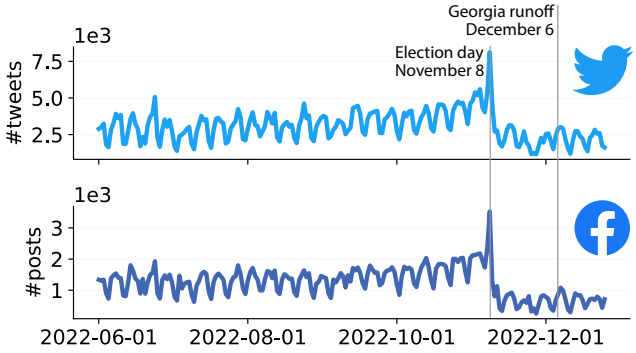

Figure 3: Daily volume of tweets and posts generated by the congressional candidates on Twitter and Facebook.

of open science. A case in point is the recent decision by Elon Musk to discontinue free access to Twitter data for research. As a result, researchers may no longer be able to re-hydrate tweets from the IDs in our dataset. In this case, we will make the raw data available to researchers upon reasonable request.

In the future, we hope to expand our framework to support more social media platforms. For instance, TikTok has recently started new programs to increase academic access to the platform.[17] We also need to adapt the system to the changing API specifications. For example, Twitter retires its V1 API endpoints on November 23, 2021, but migration to the V2 API endpoints is not trivial.

Another known issue with our dataset is noise stemming from the keyword selection process (King, Lam, and Roberts 2017). Keyword-matching approaches inevitably introduce false positives since some keywords capture both relevant and irrelevant posts. Even after the cleaning procedures, we still observe a considerable number of irrelevant posts. In addition, our snowball-sampling procedure might miss some relevant keywords or exclude keywords leading to many irrelevant posts. As a result, the list is comprehensive but not necessarily exhaustive.

Because of the library used to query the Twitter filter streaming API endpoint, a tweet is matched if all of the terms in any keyword phrase are present in the tweet, regardless of order and case. For example, for the phrase "midterms 2022," tweet objects containing both "midterms" *and* "2022," not necessarily consecutively or in that order, are returned. This might lead to a higher number of both false and true positives compared to other platforms, creating a potential inconsistency in coverage. An alternative approach would be to reformulate queries on Twitter to retrieve only tweets where the phrase terms appear consecutively and in the right order.

These sources of noise could affect downstream analyses. We recommend researchers focus more on the pre-election period when the data quality is higher. Our dataset includes both the raw data collection and the cleaned version so that researchers can deploy their own data-cleaning processes using more advanced methods, such as natural language processing techniques if needed.

Related Datasets

There are a number of datasets in the literature related to the context of U.S. elections. Here we briefly describe them.

2016 U.S. presidential election: Shao et al. provide access to the IDs of over 29M tweets linking to low-credibility and fact-checking websites shared in 2016 and 2017 (Hui et al. 2018; Shao et al. 2018b). They also open source the data collection platform, i.e., Hoaxy.[18] Similarly, Bovet et al. collect over 250M tweets by querying Twitter's streaming API over a period of 6 months (from June 1 to November 8, 2016), using a list of keywords regarding Donald Trump and Hillary Clinton (Bovet and Makse 2019).

2018 U.S. midterm elections: Deb et al. collect two tweet datasets through Twitter's API. One dataset contains over 250k tweets sharing the hashtag *#ivoted* posted on November 6, 2018, the election day. The other one includes over 2M tweets containing election-related hashtags over a period of six weeks (Deb et al. 2019). Similarly, Yang et al. create a dataset of over 60M tweets using a list of 143 relevant hashtags constructed through a snowball sampling approach (Yang, Hui, and Menczer 2022).

2020 U.S. presidential election: Chen et al. provide a continuous collection of over 1B tweets starting May 2019 using Twitter's streaming API. They include discussions related to presidential candidates and tweets containing keywords and hashtags in a manually-complied list (Chen, Deb, and Ferrara 2022). Abilov et al. focus on election fraud claims and curate a dataset containing over 30M tweets and retweets matching a manually created set of keywords, along

[17] newsroom.tiktok.com/en-us/strengthening-our-commitment-to-transparency

[18] hoaxy.osome.iu.edu

with links and metadata of YouTube videos and information of images shared in the tweets (Abilov et al. 2021). Kennedy et al., on the other hand, leverage real-time reports of over 400 distinct misinformation stories and use keyword-based searches to collect almost 50M related tweets (Kennedy et al. 2022).

Compared with those datasets, our collection is unique in that it covers the 2022 U.S. midterm elections and spans multiple social media platforms. The data allows researchers from different disciplines to better understand the online discourse in the most recent U.S. election cycle. The multi-platform nature of the dataset offers new opportunities to study the whole information ecosystem. In addition to the dataset, we also make available the source code of our collection framework, which can be used in different contexts.

Potential Applications

Our dataset provides a fertile ground for many studies that might focus on a specific platform or consider multiple platforms simultaneously. For instance, researchers can compare the spreading patterns of mis/disinformation across different platforms, as has been done comparing Facebook and Twitter (Yang et al. 2021), and analyze how malicious content migrates from one social network to another. The analyses will also shed light on the role of "superspreaders" of misinformation, who are often active on multiple platforms simultaneously (Pierri et al. 2023a; DeVerna et al. 2022) For those working on detecting and characterizing inauthentic coordinated behaviors, our dataset provides a new test bed for them to implement their methods. More importantly, researchers can investigate the presence and the impact on influence operations taking place across different platforms.

By combining the general discussion and the posts from the candidates, researchers can better understand the online interactions between these candidates and their constituencies. The advertisement data can be used to study politicians' campaigns (Islam, Roy, and Goldwasser 2023; Pierri 2023). Alternatively, the focus could be the political communication strategies put in place by different candidates on different platforms. One can also investigate the correlation between online speech and electoral outcomes.

Ethical Statement

This study has been granted exemption from Institutional Review Board review (Indiana University protocol 17036). The collection and release of the dataset are in compliance with the platforms' terms of service.

Acknowledgments

This work was supported in part by the Knight Foundation, Craig Newmark Philanthropies, Volkswagen Foundation, Italian Ministry of Education (PRIN project HOPE), and the National Science Foundation (grant ACI-1548562). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. The authors thank the technical support from Pasan Kamburugamuwa and Jacob J. Shaw at Indiana University Network Science Institute.

References

Abilov, A.; Hua, Y.; Matatov, H.; Amir, O.; and Naaman, M. 2021. VoterFraud2020: a Multi-modal Dataset of Election Fraud Claims on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 901–912.

Auxier, B.; and Anderson, M. 2021. Social Media Use in 2021. Pew Research Center.

Baele, S. J.; Brace, L.; and Coan, T. G. 2020. Uncovering the far-right online ecosystem: An analytical framework and research agenda. *Studies in Conflict & Terrorism*, 1–21.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 830–839.

Bestvater, S.; and Shah, S. 2022. 5 facts about political tweets shared by U.S. adults. Pew Research Center.

Bestvater, S.; Shah, S.; Rivero, G.; and Smith, A. 2022. Politics on Twitter: One-Third of Tweets From U.S. Adults Are Political. Pew Research Center.

Bossetta, M.; and Schmøkel, R. 2022. Cross-Platform Emotions and Audience Engagement in Social Media Political Campaigning: Comparing Candidates' Facebook and Instagram Images in the 2020 US Election. *Political Communication*, 0(0): 1–21.

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, 10(1): 1–14.

Chen, E.; Deb, A.; and Ferrara, E. 2022. # Election2020: The first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science*, 5(1): 1–18.

CrowdTangle Team. 2022. CrowdTangle. https://crowdtangle.com. Accessed: 2022-12-25.

Deb, A.; Luceri, L.; Badaway, A.; and Ferrara, E. 2019. Perils and challenges of social media and election manipulation analysis: The 2018 us midterms. In *Companion Proceedings of the World Wide Web Conference*, 237–247.

DeVerna, M. R.; Aiyappa, R.; Pacheco, D.; Bryden, J.; and Menczer, F. 2022. Identification and characterization of misinformation superspreaders on social media. *Preprint arXiv:2207.09524*.

DeVerna, M. R.; Pierri, F.; Truong, B. T.; Bollenbacher, J.; Axelrod, D.; Loynes, N.; Torres-Lugo, C.; Yang, K.-C.; Menczer, F.; and Bryden, J. 2021. CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 992–999.

Di Giovanni, M.; Pierri, F.; Torres-Lugo, C.; and Brambilla, M. 2022. VaccinEU: COVID-19 vaccine conversations on Twitter in French, German and Italian. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1236–1244.

Gallotti, R.; Valle, F.; Castaldo, N.; Sacco, P.; and De Domenico, M. 2020. Assessing the risks of 'infodemics' in response to COVID-19 epidemics. *Nature Human Behaviour*, 4(12): 1285–1293.

Golovchenko, Y.; Buntain, C.; Eady, G.; Brown, M. A.; and Tucker, J. A. 2020. Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election. *The International Journal of Press/Politics*, 25(3): 357–389.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.

Hardy, B. W.; and Castonguay, J. 2018. The moderating role of age in the relationship between social media use and mental well-being: An analysis of the 2016 General Social Survey. *Computers in Human Behavior*, 85: 282–290.

Hui, P.-M.; Shao, C.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018. The Hoaxy misinformation and fact-checking diffusion network. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Islam, T.; Roy, S.; and Goldwasser, D. 2023. Weakly Supervised Learning for Analyzing Political Campaigns on Facebook. In *Proceedings of the International AAAI Conference on Web and Social Media*. Preprint arXiv:2210.10669.

Johnson, N. F.; Leahy, R.; Restrepo, N. J.; Velásquez, N.; Zheng, M.; Manrique, P.; Devkota, P.; and Wuchty, S. 2019. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773): 261–265.

Kennedy, I.; Wack, M.; Beers, A.; Schafer, J. S.; Garcia-Camargo, I.; Spiro, E. S.; and Starbird, K. 2022. Repeat Spreaders and Election Delegitimization: A Comprehensive Dataset of Misinformation Tweets from the 2020 US Election. *Journal of Quantitative Description: Digital Media*, 2.

King, G.; Lam, P.; and Roberts, M. E. 2017. Computer-Assisted Keyword and Document Set Discovery from Unstructured Text. *American Journal of Political Science*, 61(4): 971–988.

Lukito, J. 2020. Coordinating a multi-platform disinformation campaign: Internet Research Agency activity on three US social media platforms, 2015 to 2017. *Political Communication*, 37(2): 238–255.

Papasavva, A.; Zannettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 885–894.

Pasquetto, I. V.; Swire-Thompson, B.; Amazeen, M. A.; Benevenuto, F.; Brashier, N. M.; Bond, R. M.; Bozarth, L. C.; Budak, C.; Ecker, U. K.; Fazio, L. K.; et al. 2020. Tackling misinformation: What researchers could do with social media data. *The Harvard Kennedy School Misinformation Review*, 1(8).

Pierri, F. 2023. Political advertisement on Facebook and Instagram in the run-up to 2022 Italian general election. In *WebSci'23 – 15th ACM Web Science Conference*.

Pierri, F.; DeVerna, M. R.; Yang, K.-C.; Axelrod, D.; Bryden, J.; and Menczer, F. 2023a. One Year of COVID-19 Vaccine Misinformation on Twitter: Longitudinal Study. *Journal of Medical Internet Research*, 25: e42227.

Pierri, F.; Luceri, L.; Jindal, N.; and Ferrara, E. 2023b. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. In *15th ACM Web Science Conference 2023*.

Primack, B. A.; Shensa, A.; Escobar-Viera, C. G.; Barrett, E. L.; Sidani, J. E.; Colditz, J. B.; and James, A. E. 2017. Use of multiple social media platforms and symptoms of depression and anxiety: A nationally-representative study among US young adults. *Computers in Human Behavior*, 69: 1–9.

Shah, S.; and Grant, A. 2021. Charting Congress on Social Media in the 2016 and 2020 Elections. Pew Research Center.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018a. The spread of low-credibility content by social bots. *Nature Communications*, 9(1): 1–9.

Shao, C.; Hui, P.-M.; Wang, L.; Jiang, X.; Flammini, A.; Menczer, F.; and Ciampaglia, G. L. 2018b. Anatomy of an online misinformation network. *Plos One*, 13(4): e0196087.

Stewart, C. A.; Cockerill, T. M.; Foster, I.; Hancock, D.; Merchant, N.; Skidmore, E.; Stanzione, D.; Taylor, J.; Tuecke, S.; Turner, G.; Vaughn, M.; and Gaffney, N. I. 2015. Jetstream: A Self-Provisioned, Scalable Science and Engineering Cloud Environment. In *Proceedings of the XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15, 1–8. Association for Computing Machinery.

Stier, S.; Bleier, A.; Lietz, H.; and Strohmaier, M. 2018. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on Facebook and Twitter. *Political Communication*, 35(1): 50–74.

Stocking, G.; Mitchell, A.; Matsa, K. E.; Widjaya, R.; Jurkowitz, M.; Ghosh, S.; Smith, A.; Naseer, S.; and St. Aubin, C. 2022. The Role of Alternative Social Media in the News and Information Environment. Pew Research Center.

Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; and Wilkins-Diehr, N. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering*, 16(5): 62–74.

Velasquez, N.; Leahy, R.; Restrepo, N. J.; Lupu, Y.; Sear, R.; Gabriel, N.; Jha, O.; Goldberg, B.; and Johnson, N. 2021. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. *Scientific Reports*, 11(1): 1–8.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888): 264–268.

Wilson, T.; and Starbird, K. 2020. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 1(1).

Yang, K.-C.; Hui, P.-M.; and Menczer, F. 2022. How Twitter data sampling biases US voter behavior characterizations. *PeerJ Computer Science*, 8: e1025.

Yang, K.-C.; Pierri, F.; Hui, P.-M.; Axelrod, D.; Torres-Lugo, C.; Bryden, J.; and Menczer, F. 2021. The COVID-19 Infodemic: Twitter versus Facebook. *Big Data & Society*, 8(1): 1–16.

Yarchi, M.; Baden, C.; and Kligler-Vilenchik, N. 2021. Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media. *Political Communication*, 38(1-2): 98–139.