Candidata: United States 2024 Elections Candidates and Social Media Handles

Megan A. Brown[1], Maggie Macdonald[2], Josephine Lukito[3], Cameron Hickey[4], Kaitlyn Dowling[4], Myra Miranda[4]

[1]University of Michigan
[2]University of Kentucky
[3]University of Southern Denmark
[4]National Conference on Citizenship, Algorithmic Transparency Institute
mgnbrown@umich.edu, maggie.macdonald@uky.edu, lukito@sam.sdu.dk, {cameron,kaitlyn,myra}@ncoc.org

Abstract

We introduce the Candidata dataset, a public archive of social media data for candidates for primary and general federal elections in the United States in 2024. The data collection spans multiple platforms, capturing the diverse ways candidates interact with potential voters, media, and other audiences. By documenting candidate behavior across platforms, this dataset provides an unprecedented resource for analyzing the full scope of social media use in political campaigns. The dataset includes the full list of candidates for federal elections in the United States in 2024, including candidate demographics, election information, and identifiers to link the candidates to other datasets. The social media data includes message content and metadata such as post timestamp and engagement metrics.

Dataset — https://doi.org/10.3886/sqhd-kv39

Introduction

A prominent challenge in the study of political behavior online is the lack of comprehensive data about candidates and their digital behaviors. Studies on candidates' communication tends to focus on a single platform (with some exceptions (Blum, Cormack, and Shoub 2023)), particularly because systematically identifying all candidates and their social media data can be time-intensive and cost-prohibitive for researchers.

We describe in this paper a dataset of social media handles for federal candidates including the U.S. House of Representatives, the Senate, and the Presidency. All social media handles for candidates from Facebook, Instagram, Threads, TikTok, X (Twitter), YouTube, Gettr, Rumble, Telegram, and Truth Social are included. We also make available candidate attributes such as FEC IDs, party affiliations, incumbency status, district competitiveness ratings for each contest, race and gender (where available). For incumbents, we also include birth year, Bioguide IDs, ICPSR IDs, and first-dimension DW-NOMINATE scores. In addition to the candidates dataset, we provide the candidates' organic posts on social media for analysis of the content.[1] This dataset provides the broader research community with a comprehensive view of the social media behavior of candidates for federal office. Using these data, researchers can link the politician data we provide to their own data; for example, given a survey of social media users with donated digital trace data, researchers can use these data to determine what politicians a survey participant engages with online.

The study of social media and candidates' behavior in particular is of interest across political science, communication, psychology, information science, and other disciplines (Lewandowsky, Jetter, and Ecker 2020; Farkas and Bene 2021). This dataset will facilitate the study of candidates' social media communications across a variety of social media platforms, each of which offers different audiences (Auxier and Anderson 2021) and affordances. This data allows researchers to examine the relationship between candidate communications and policy agendas (Barberá et al. 2019), the interplay between media and journalists and political campaigns (Kreiss, Lawrence, and McGregor 2018; Macdonald, Tucker, and Nagler 2025), partisan polarization and political hostility among elites (Ballard et al. 2022), and how politicians impact the quality of information online (Starbird, DiResta, and DeButts 2023; Mosleh and Rand 2022; Lasser et al. 2022). Moreover, the diverse array of platform data provided will allow researchers to study how candidates shape messaging for different audiences and mediums, whether candidate communications can garner support from donors and voters, and how audiences respond and engage with different messaging strategies across platforms. This dataset opens new avenues for more comprehensive, cross-platform analyses of candidate behavior on social media and the role of politicians in the information environment.

Data Collection Methods

We first outline the process by which we collected and validated social media handles for all candidates across ten social media platforms. We then describe how we collect the posts data from each social media platform.

Social Media Handle Collection

For the list of politicians, we start with a dataset of candidates compiled by the Center for Tech and Civic Life

[1]This means we include all posts that originate from the candidates' accounts, but we do not include posts that were created as advertisements on the platform.

Variable	Description	Options / Format	Source	H	S
Official Name	First and last (and sometimes middle) name of candidates	Name string	CTCL	T	T
Candidate Name	Candidate name; missing for some Senators not running in 2024	Name string	CTCL	T	T
FirstName	Candidate first name, extracted from Official Name and hand-verified	Name string	Authors/CTCL	T	T
LastName	Candidate last name, extracted from Official Name and hand-verified	Name string	Authors/CTCL	T	T
bioname	Full name for incumbents only	LAST, First Middle (Nickname)	VoteView	T	T
State	State where candidate is running	State name (e.g., Alabama)	CTCL	T	T
State.Abb	State abbreviation where candidate is running	AL, DE, etc.	Authors/CTCL	T	T
District	Congressional district	Numeric; at-large = 1	CTCL	T	F
Incumbent	Incumbency status in 2024, incl. retirement or office change	Categorical	Authors/FEC	T	T
SeatUp_2024	Whether Senate seat was up in 2024	Yes / No	Authors	F	T
Level	Level of office for which candidate is running	country, state, local	CTCL	T	T
Electoral District	District name (House) or state name (Senate)	Text string	CTCL	T	T
Office Name	Office sought by candidate	U.S. Rep. / U.S. Sen.	CTCL	T	T
Party ID	Party identifier	Democratic Party, Republican Party, etc.	CTCL	T	T
Party Standardized	Standardized party name	Democratic Party, Republican Party, etc.	—	T	T
DOB	Date of birth (incumbents only)	YY/MM/DD	VoteView	T	T
Race	Racial identity (incomplete)	White, Black or African-American, etc.	CTCL	T	T
Gender	Gender identity (incomplete)	Male, Female	CTCL	T	T
ICPSR	ICPSR ID (incumbents only)	5-digit code	VoteView	T	T
State_ICPSR	ICPSR state code (incumbents only)	2-digit code	VoteView	T	T
Bioguide ID	Congressional bioguide ID (incumbents only)	7 characters	VoteView	T	T
nominate_dim1	DW-NOMINATE ideology score (incumbents only)	Numeric	VoteView	T	T

Table 1: Demographic and identity variables for 2024 U.S. candidates.

(CTCL). This original dataset only contained Twitter, Facebook, YouTube, and Instagram handles for all primary candidates, incumbents and non-incumbents, running for election to the U.S. Congress in 2024. Between the House and Senate, this was 2,118 primary candidates. We deduplicated candidates that appeared more than once and several research assistants (RAs) validated and expanded this dataset in several ways. First, where politicians were reported by CTCL to not have an account on a particular platform, RAs validated this by either confirming the non-existent account or finding the account for the candidate on the platform. Second, RAs also identified TikTok and Threads handles for all candidates, expanding beyond the CTCL data. Third, RAs collected handles for all Republican candidates on Truth Social, Gettr, and Rumble, Telegram, further expanding beyond the CTCL data.[2]

For incumbents, we completed this task for their office accounts and campaign accounts separately; for non-incumbents, who do not have office accounts, we completed this task for their campaign accounts. Personal accounts for all politicians are excluded. We then conducted a secondary validation process for each account identified, where a second RA verified that the account found was indeed associated with the candidate listed. The secondary validation process also verified that no account existed for candidates where no account was found by RAs initially.

RAs identified and validated social media accounts for candidates following a structured process. First, RAs were assigned certain candidates and platforms for which they needed to identify accounts. RAs were instructed to focus on official office or campaign accounts, avoiding personal social media accounts for the candidates such as personal Facebook accounts or private profiles. RAs first searched official sources, such as the candidate's campaign or office website, which frequently provided the most reliable links to the candidates' social media profiles. If the social media profiles were not listed on these official websites, the research assistants expanded their search beyond these sources by searching the candidate's name on the platform's search tool or by searching for "[Candidate Name] [State] [Platform]" via a search engine.

After an account was located, a separate reviewer validated the account by cross-referencing the account with multiple sources including the candidate's official websites, other known social media profiles, and reputable media sources. If an account was initially marked as missing, but during validation an RA found the account, then it went for a third round of validation with a separate RA to ensure that

[2]We do not find handles for Democrats on Truth Social, Gettr, Rumble, or Telegram since these candidates are very rarely on alternative platforms (Lukito et al. 2025).

the sourced handle is in fact valid. During this process, we advised RAs to exercise caution, particularly on platforms with less stringent verification processes. For all search and validation steps, RAs assessed the legitimacy of the accounts by considering factors such as the presence of verification badges, consistent content related to the candidate's public or campaign activities, follower count, affiliation with campaign-related content, and the nature of interactions with other accounts. In cases where an account appeared suspicious or unverifiable, an author of the paper made the final decision on whether the account was officially affiliated with the candidate or not. If no account could be found after a thorough search, and a second RA agreed that an account couldn't be found, the candidate was marked as not having an account on that platform. To be in our final dataset, a handle was validated by at least two separate coders.

Social Media Data Collection

For each handle that we identified for candidates, we collect all posts from January 1, 2023 until December 31, 2024. Social media data was collected through Junkipedia, a tool created by the Algorithmic Transparency Institute (ATI) for collecting, monitoring, and analyzing social media content for the public interest. ATI collects all posts for candidates, not just posts submitted through ATI's misinformation ticketing system. For each social media site, we collect the posts for each candidate that holds an account on that site. After an initial collection throughout the 2024 primary period (spring-fall 2024), new posts were collected at least every 24 hours.[3] We outline the data collection methodology for each platform below. We standardize the metadata for each platform such that similar attributes (e.g., the publish date for the post) retain the same variable name. In total, our dataset includes about 3.1 million posts from 2024 primary candidates posted between January 1, 2023 and December 31, 2024. The posts in this dataset reflect the posts from January 1, 2023 and December 31, 2024 available in Junkipedia for the accounts identified in our social media handle collection process.

Facebook Facebook is a social media platform owned by Meta where candidates can create official "Pages" to share content, engage with followers, and organize their public presence distinct from their personal profiles. Facebook users can follow a candidate's page to see what a candidate posts in their news feed. Facebook allows users to post text, photos, videos, and URLs to external content. We collect the Facebook pages (rather than personal profiles) for the candidates in the dataset. For each candidate, we collect their posts via a proxy-based scraping API through Junkipedia for post data including the post content, publish date, and engagement.[4] Data is collected through Junkipedia every 4–24 hours. Collected post types include posts and videos.

Gettr Gettr is a micro-blogging service where candidates can create a profile to share posts and engage with other users and posts. Like other micro-blogging services, Gettr facilitates the posting of text, image, video, and URL content. For each account that candidates hold on Gettr, we collect posts via direct web scraping through Junkipedia every 4–24 hours. Collected post types include posts and videos.

Instagram Instagram is a social media platform owned by Meta where candidates can create official profiles to share content. Instagram is an image-based platform, so all posts contain image or video content alongside descriptions or captions for the content. Instagram users can follow a candidate's page to see what a candidate posts on their news feed. Instagram data is collected via direct web scraping through Junkipedia, facilitating the collection of public posts, captions, hashtags, and engagement metrics from candidate profiles.[5] Data is collected every 4–24 hours and collected post types include posts and videos.

Rumble Rumble is a YouTube-like video platform. Similar to YouTube, users create channels on which they can share videos, and users can view and comment on those videos. For each video, we collect the video URL, channel title, video description, and video publish date. Rumble data was collected via direct web scraping through Junkipedia; collected post types include videos only.

Telegram Telegram is a messaging application primarily used for direct messaging to individuals or groups. However, candidates can create "channels" which allow them to broadcast content to users who subscribe to their channel. Telegram allows channels to post text content, images, videos, and external URLs. For each candidate's Telegram channel, we collect content that the channel broadcasts via direct web scraping through Junkipedia every 4–24 hours. Collected post types include posts, replies, reposts, and videos.

Threads Threads is a Meta-owned micro-blogging service. On Threads, users can post content including text, images, videos, and URLs to external content. For each candidate Threads account, we collect relevant post metadata including the post content, publish date, and engagement via web scraping. Data is collected via direct web scraping through Junkipedia every 4–24 hours. Collected post types include posts, replies, reposts, and videos.

TikTok TikTok is a short-form video platform which allows users to post short videos to their followers. Users typically engage with the platform through the "For You Page," an algorithmically-generated feed of content. Users can follow candidate profiles, and those posts will be more

[3] As these are active accounts, posts and accounts that were available at the time of our collection may no longer be available today. For some states, we completed our first collection after its primary election; thus, some primary losers may have deactivated their accounts and are therefore missing from our post data. We further elaborate on the limitations of these data below.

[4] Initially Crowdtangle was used for Facebook collection. The switch was made to a proxy-based scraping API after Crowdtangle was shuttered in Fall 2024.

[5] Initially Crowdtangle was used for Instagram collection. The switch was made to a proxy-based scraping API after Crowdtangle was shuttered in 2024.

Variable	Description	Format	Source	H	S
Official Facebook Page	Facebook account for congressional office	URL	CTCL/ATI	T	T
Official Instagram Account	Instagram account for congressional office	URL	CTCL/ATI	T	T
Official Twitter Account	Twitter/X account for congressional office	URL	CTCL/ATI	T	T
Official YouTube Account	YouTube account for congressional office	URL	CTCL/ATI	T	T
Official TikTok Account	TikTok account for congressional office	URL	CTCL/ATI	T	T
Campaign Facebook Account	Facebook account for campaign	URL	CTCL/ATI	T	T
Campaign Instagram Account	Instagram account for campaign	URL	CTCL/ATI	T	T
Campaign Twitter Account	Twitter/X account for campaign	URL	CTCL/ATI	T	T
Campaign YouTube Account	YouTube account for campaign	URL	CTCL/ATI	T	T
Campaign TikTok Account	TikTok account for campaign	URL	CTCL/ATI	T	T
Campaign Threads Account	Threads account for campaign	URL	CTCL/ATI	T	T
Office Threads Account	Threads account for congressional office	URL	CTCL/ATI	T	T
Office TruthSocial Account	Truth Social account for office (Republicans only)	URL	CTCL/ATI	T	T
Campaign TruthSocial Account	Truth Social account for campaign (Republicans only)	URL	CTCL/ATI	T	T
Office Telegram Account	Telegram account for office (Republicans only)	URL	CTCL/ATI	T	T
Campaign Telegram Account	Telegram account for campaign (Republicans only)	URL	CTCL/ATI	T	T
Official Gettr Account	Gettr account for office (Republicans only)	URL	CTCL/ATI	T	T
Campaign Gettr Account	Gettr account for campaign (Republicans only)	URL	CTCL/ATI	T	T
Official Rumble Account	Rumble account for office (Republicans only)	URL	CTCL/ATI	T	T
Campaign Rumble Account	Rumble account for campaign (Republicans only)	URL	CTCL/ATI	T	T

Table 2: Social media variables for 2024 U.S. candidates.

likely to appear in their "For You Page" and will appear in the separately-curated "Following" page which compiles all posts by the individuals a user follows. For each candidate TikTok account, we collect all videos posted to their account through a third-party API via Junkipedia every 4–24 hours. Collected post types include posts and videos as well as metadata such as engagement metrics, captions, and sounds used.

Truth Social Truth Social is a micro-blogging service owned by former president Donald Trump's media company. Like other micro-blogging services, candidates can post text and multi-media content which users can engage with. Data is collected via direct web scraping through Junkipedia every 4–24 hours. Collected post types include posts, replies, reposts, and videos and post metadata including the publish date and engagement metrics.

Twitter/X Twitter is a micro-blogging service which allows candidates to post text and multi-media content. For each candidate profile, we collect all posts and metadata including the post content, publish date, and engagement metrics. Data is collected via a third-party API through Junkipedia every 5 hours. Collected post types include posts, replies, reposts, comments, and videos.

YouTube YouTube is a video-based platform organized into "channels." Channels post videos, which users can view and comment on. For YouTube, we first convert all candidate channel URLs to their respective channel IDs, the unique identifier assigned to each channel by YouTube. For each video from a channel, we collect the video URL, channel title, video description, and video publish date. Using the channel ID, data is collected via the official YouTube Data API and RSS feeds through Junkipedia every 4–24 hours. Video is the sole collected post type, with full collection history available.

Variables

The complete database of candidate social media handles is available through the Social Media Archive (SOMAR) at the Inter-university Consortium for Political and Social Research[6]. The dataset includes both candidate attributes and social media handles and a dataset of the candidates' posts on the platforms collected. To facilitate analysis of the social media data, we provide a dashboard at https://www.candidata24.org/ to analyze the data in aggregated formats. For access to the full data, researchers can apply for access through the Social Media Archive at ICPSR (see

[6]See https://socialmediaarchive.org

Variable	Definition
id	Unique identifier for the post on Junkipedia platform
channel	Name or identifier of the channel/account that published the post on Junkipedia platform
channel_url	Direct URL of the channel or account that published the post
post_url	Direct URL to the original post on the platform
likes	Number of likes/reactions the post has received
views	Number of times the post has been viewed
replies	Number of replies/comments on the post
shares	Number of times the post has been shared/reposted
thumbnail_url	URL of the thumbnail image associated with the post
published_at	Timestamp indicating when the post was published
audio_text	Text extracted from audio associated with the post (using Whisper speech-to-text).
image_text	Text extracted from images in the post (e.g., OCR output)
complete_post_text	Full textual content of the post, combining all available text sources
transcript_text	Transcript text associated with the post, if available
image_text_search	Search-optimized text derived from image-extracted text
text_search	Search-optimized text derived from post text content
all_text	Aggregated text combining all textual fields for analysis or search
junkipedia_url	URL linking to the post's page on Junkipedia (to access further metadata where necessary)
platform	Platform/domain on which the post was published

Table 3: Variable names and definitions for post metadata

https://doi.org/10.3886/sqhd-kv39).

The first dataset consists of the candidate attributes and social media handles table. Each observation in this table is a single candidate, which is unique at the person-contest level. We provide candidate attributes including name, Bioguide ID (a commonly-used identifier for members of Congress), political party, and other attributes. The full list of variables can be found in Tables 1 and 2. Table 1 contains the demographic variables we collect for candidates. Table 2 contains the social media variables we collect for candidates.[7] We include the variable name, description, and format of the variable. We also include a column denoting the source of the data. The "H" column represents variables that are available for candidates for House. The "S" column represents variables that are available for candidates for Senate.

Finally, Table 3 contains the variable list for the social media data. We normalize common fields across platforms so that fields that are the same (e.g., the date that the post was published) have the same column name. We also select fields that are commonly used in research including channel information, engagement information, and content. However, recognizing that there are individual platform differences in metadata that may be important to researchers, we also provide a URL in the "junkipedia_url" column where researchers can access and download the full post metadata. This structure balances making data across many platforms more usable both in terms of data complexity while still providing access to expanded metadata for researchers who are interested further analysis with those data.

Data Validation

As described in the previous section, we undergo a thorough validation process where each candidate's profile is verified by multiple research assistants. For each candidate profile surfaced, a second research assistant validates that the account is indeed affiliated with the candidate by verifying through official sources such as the candidate's campaign website, office websites, platform verification symbols, and news media references to the candidate's social media account. For each candidate that we do not find a profile on a given platform, a second research assistant validates that the profile does not exist by searching using the platform's native search tools, a search engine, and the candidate's website. In cases where the second research assistant finds a profile associated with the candidate, that profile is validated like other profiles.

To validate the identifiers included in the dataset, we merged on candidate state, district (for House candidates), last name, and first initial of first name. We then manually checked every candidate that did not merge every time we merged in a new dataset. Where applicable, we manually fixed inconsistencies– often variations in how names appear across databases, such as 'Robert' vs. 'Bob' as a first name.

Data Limitations

Researchers should note that the data provided is from social media, which can frequently change. Users can delete (or in some cases edit) their posts, meaning if a post was created and deleted in between our collection times, we would have no record of that post. Moreover, candidates may delete their accounts after losing a primary or general election, meaning links back to the original posts may not be valid after a certain period. Additionally, engagement metrics are reported for the timestamp that the data was collected, meaning that posts may have garnered more engagement after we collected the post. Researchers should carefully caveat their claims to account for potential bias due to the temporal instability of platform data.

Demonstrating Data Utility

To demonstrate the utility of the data, we show the number of posts on each platform over time. Figure 1 shows

[7] These are shared as a single dataset, but we show the variable descriptions separately for legibility.

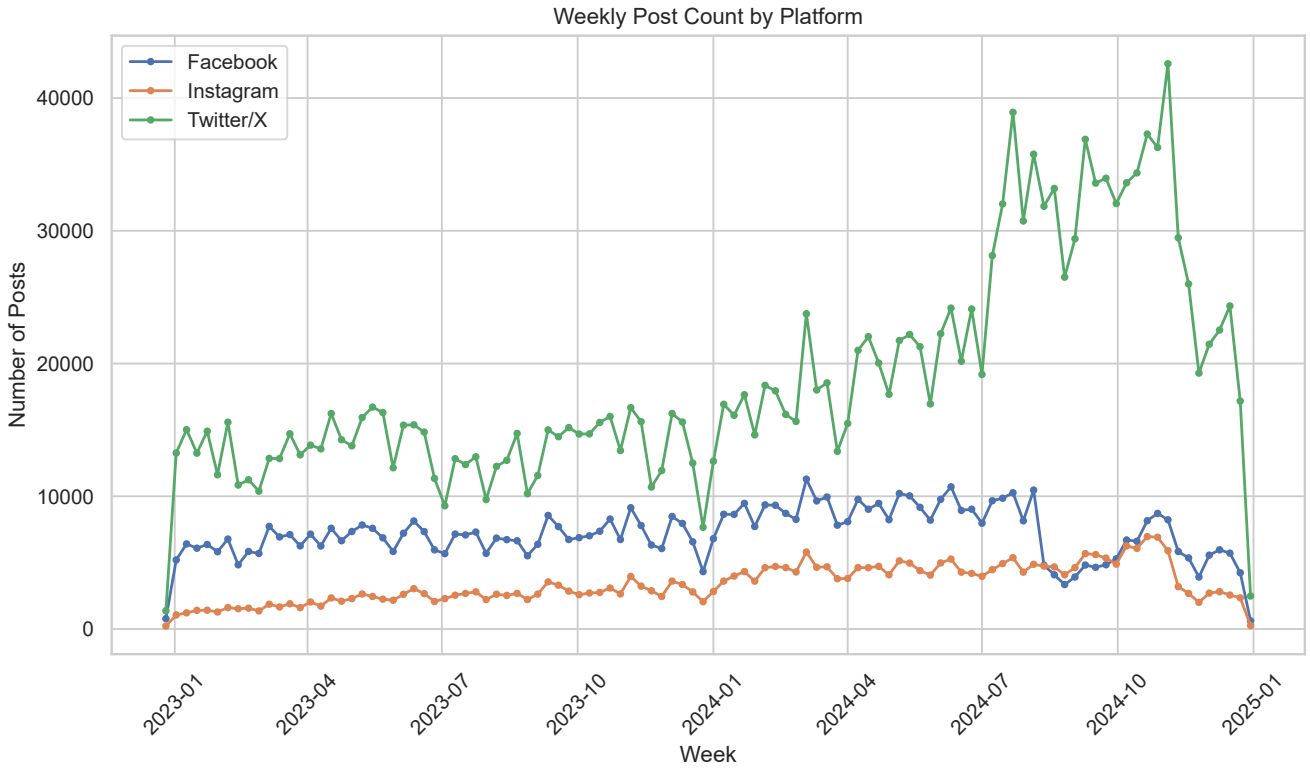

Figure 1: Weekly post count by platform (Twitter, Facebook, Instagram).

the number of weekly posts across each platform for Twitter, Facebook, and Instagram. Over 2023, posts across all platforms are generally constant. In early 2024, the weekly count of posts increases on Twitter, and it increases more sharply in late 2024, finally decreasing after the election period. Unlike Twitter, posts on Facebook and Instagram do not increase. This may be because candidates shift their focus to other platforms (such as Twitter), or candidates may shift their posting on Facebook and Instagram to paid advertisements, which are not captured in this data collection.

Figure 2 shows the number of weekly posts for the remainder of the platforms in our data. YouTube (pink) has the highest variability, but it appears to generally increase over the course of the campaign. The volume of TikTok posts (purple) also increased over the course of the campaign cycle before dropping after the elections. Threads appears to have minimal posting over time. This is likely because the uptake of Threads was low across candidates. The remaining platforms (Gettr, Rumble, Telegram, and Truth Social) appear to have generally consistent posting amounts over time. Notably, these are all alt-tech platforms. This suggests that alt-tech platforms may serve different communicative purposes for campaigns than the mainstream platforms.

Other Research Applications

Beyond the application shown here, we propose several lines of research that could be pursued using this dataset. First, video platforms remain understudied, and this dataset contains the same actors linked across both video and non-video platforms. Researchers could analyze differences in messaging strategies, audience engagement, and emotional appeals across video platforms compared to text-based ones. This would help fill in a gap in the literature, which has largely focused on platforms like Twitter/X and Facebook due to the accessibility of those data.

Second, misinformation remains prevalent, and recent analyses shows that popular political influencers, including politicians, are key spreaders of misinformation online. The dataset could be used to identify, track, and compare instances of misinformation shared by political candidates across multiple social media platforms. By linking posts to individual politicians and campaigns, researchers could examine when misinformation is most likely to be deployed and how it spreads across platforms. This would support deeper understanding of elite-driven misinformation and its potential impact on democratic processes.

Third, this dataset spans multiple modes of content, including text, images, and video. It enables multimodal analyses of how actors combine visual and linguistic elements to persuade voters. Researchers could study how visual cues, tone of voice, captions, and imagery reinforce or contradict textual messages. Such analyses would advance computational social science by capturing the full complexity of modern political communication rather than focusing on text

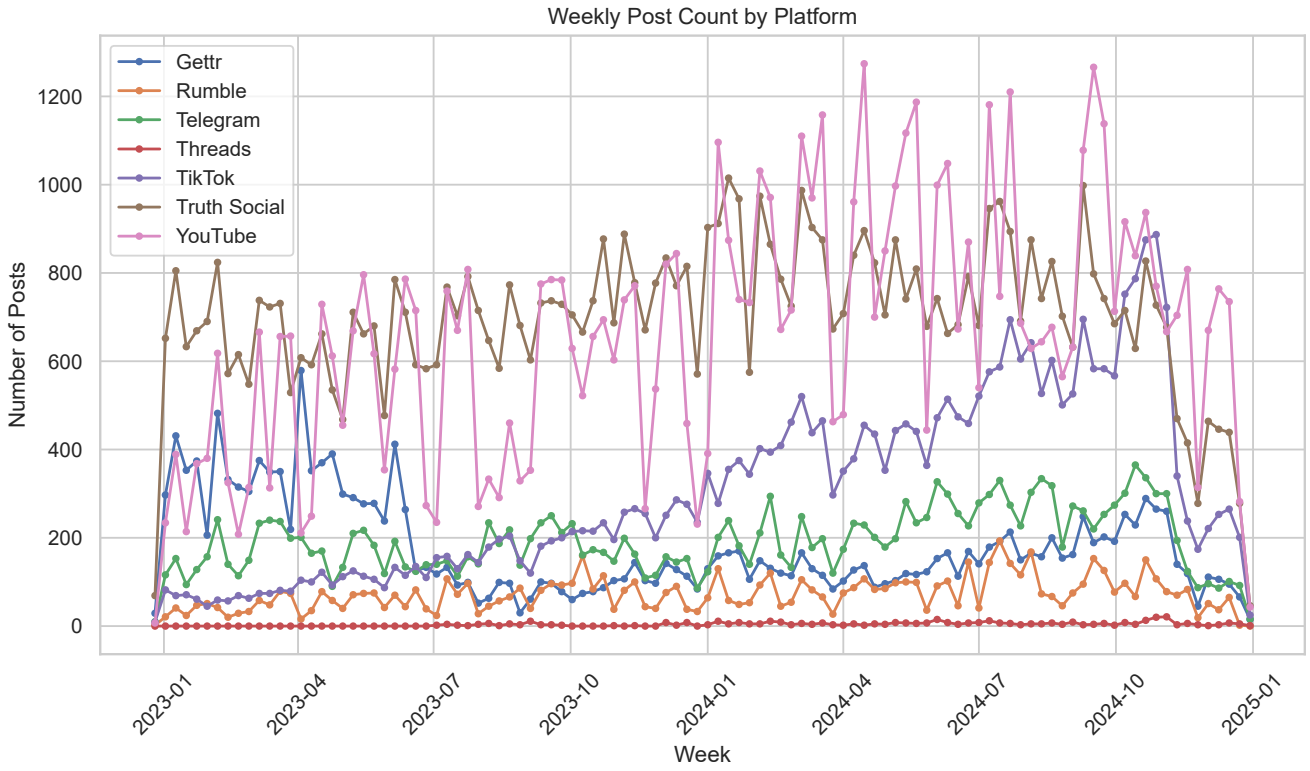

Figure 2: Weekly post count by platform (Gettr, Rumble, Telegram, Threads, TikTok, Truth Social, YouTube).

alone.

Finally, there are many applications beyond the scope of those suggested in this paper. While the inclusion of solely politicians may seem restrictive, we argue that they are an ideal case for studying how platform audiences, affordances, and modes impact the spread of information. In cross-platform studies, it is often difficult to match users across platforms since many platforms are pseudononymous (e.g., Reddit and Twitter/X), and even users with the same username might be different individuals. Thus, when examining how platforms differ in aggregate, it is difficult to disentangle how features of the platform shape behavior from how selection of users into a platform shapes behavior. Holding fixed the same users across platforms makes analyses of platform effects easier. We hope this dataset allows for a more expansive analysis of both how elites shape the political sphere online and how platform effects shape user behavior.

FAIR Principles

Our dataset uses FAIR principles. Our dataset is:

- **Findable**: Our dataset is available online with a persistent identifier.
- **Accessible**: The dataset is accessible through HTTP protocols.
- **Interoperable**: The dataset is interoperable. It is available in CSV and JSON formats, which are readable regardless of the programming language the researcher chooses to use. The dataset also contains common identifiers (e.g., ICPSR and FEC IDs) that can be used to combine the data with other datasets of interest to researchers.
- **Reusable**: The data has thorough documentation and codebooks available to ensure the data can be reused by other researchers.

Ethical Considerations

This research makes identifying, individual-level data available for researchers. While standard practices in these cases is often anonymization or aggregation to ensure that subjects cannot be reidentified, such steps are not appropriate in this context. The politicians in this dataset are public figures who make public statements across a variety of platforms, including social media platforms, press releases, their websites, public debates, news media, and other public events. Social media data are a part of this repertoire of communication forums. Furthermore, archiving and sharing this data does not violate the expected privacy boundaries of these data, as politicians often expect that their posts will be analyzed, discussed, and archived in the public (Russell and Macdonald 2024). We argue that the potential harm to subjects is minimal, as these posts are already widely shared, analyzed, and archived, and the politicians (and their campaign staff who run their accounts) are aware of this when they decide to post.

Conclusion

We introduce the Candidata dataset, which offers a comprehensive, publicly available archive of social media activity from candidates in the 2024 U.S. federal primary and general elections. Spanning multiple platforms, it captures the varied ways candidates engage with voters, media, and broader audiences. By systematically documenting candidate behavior across platforms, the dataset provides an unprecedented foundation for studying the scope and dynamics of social media use in contemporary political campaigns. In addition to message content and engagement metadata, Candidata includes a complete roster of federal candidates with demographic and election-related information, along with identifiers that facilitate linkage to other datasets, enabling a wide range of future research applications.

References

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

Ballard, A.; DeTamble, R.; Dorsey, S.; Heseltine, M.; and Johnson, M. 2022. Dynamics of Polarizing Rhetoric in Congressional Tweets. *Legislative Studies Quarterly*.

Barberá, P.; Casas, A.; Nagler, J.; Egan, P. J.; Bonneau, R.; Jost, J. T.; and Tucker, J. A. 2019. Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *American Political Science Review*, 113(4): 883–901.

Blum, R.; Cormack, L.; and Shoub, K. 2023. Conditional Congressional communication: how elite speech varies across medium. *Political Science Research and Methods*, 11(2): 394–401.

Farkas, X.; and Bene, M. 2021. Images, politicians, and social media: Patterns and effects of politicians' image-based political communication strategies on social media. *The International Journal of Press/Politics*, 26(1): 119–142.

Kreiss, D.; Lawrence, R.; and McGregor, S. 2018. In their own words: Political practitioner accounts of candidates, audiences, affordances, genres, and timing in strategic social media use. *Political Communication*, 35(1): 8–31.

Lasser, J.; Aroyehun, S. T.; Simchon, A.; Carrella, F.; Garcia, D.; and Lewandowsky, S. 2022. Social media sharing of low-quality news sources by political elites. *PNAS nexus*, 1(4): pgac186.

Lewandowsky, S.; Jetter, M.; and Ecker, U. K. 2020. Using the president's tweets to understand political diversion in the age of social media. *Nature Communications*, 11(1): 5764.

Lukito, J.; Macdonald, M.; Chen, B.; Brown, M. A.; Prochaska, S.; Yang, Y.; Greenfield, J.; Suk, J.; Zhong, W.; Dahlke, R.; et al. 2025. Candidates Be Posting: Multi-Platform Strategies and Partisan Preferences in the 2022 US Midterm Elections. *Social Media+ Society*, 11(2): 20563051251337541.

Macdonald, M.; Tucker, J. A.; and Nagler, J. 2025. The Democratizing and Polarizing Impact of Fundraising on Twitter: Viral Incentives and the Catalyzing Role of Mainstream Media.

Mosleh, M.; and Rand, D. G. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, 13(1): 7144.

Russell, A.; and Macdonald, M. 2024. Congressional Communications in a Digital Era. In Dodd, L. C.; Oppenheimer, B.; Rubin, R. B.; and Evans, C. L., eds., *Congress Reconsidered, 13th Edition*. CQ Press.

Starbird, K.; DiResta, R.; and DeButts, M. 2023. Influence and improvisation: Participatory disinformation during the 2020 US election. *Social Media+ Society*, 9(2): 20563051231177943.

Generative AI Disclosure

Generative AI was used to refine grammar and style written by the authors. Generative AI was also used to assist in generating code for data cleaning and figure generation. The authors take responsibility for all errors in the manuscript.

Ethics Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, because this dataset allows researchers to understand how candidates for office behave online and their role in the information environment. Posting individual-level data about those running for political office is a common practice. Public and journalistic archives already routinely cover and collect this content to discuss in public communication.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we clearly describe the validation process for the data and any assumptions we made during this process.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we fully describe the limitations of the dataset proposed.

   (e) Did you describe the limitations of your work? Yes, we fully describe the limitations of the dataset proposed.

   (f) Did you discuss any potential negative societal impacts of your work? Yes

   (g) Did you discuss any potential misuse of your work? Yes

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we provide thorough documentation of the data and its limitations. We also provide information about where the data is published.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, we cite sources of data included in our analysis.
   (b) Did you mention the license of the assets? The license of the data is that provided by SOMAR.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, we provide a URL to our dataset in the main text.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, we do not discuss how consent was obtained for the candidates whose data we collect. However, we discuss the ethical considerations of making candidate social media information public.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we discuss the personally identifying information and the presence of offensive content.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, we discuss how we make the dataset FAIR.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? No, we do not make available a Datasheet for the dataset. Because the Datasheet for the Dataset is designed for machine learning datasets, we did not find the majority of the questions to be usable or applicable for this application. We provide answers to the relevant questions regarding our ethical considerations for sharing the data, the data collection methodology, and limitations of the data in the corresponding sections of this manuscript.
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA