

A Public Dataset Tracking Social Media Discourse about the 2024 U.S. Presidential Election on Twitter/X

Ashwin Balasubramanian¹, Vito Zou¹, Hitesh Narayana¹, Christina You¹,
Luca Luceri^{1,2}, Emilio Ferrara^{1,2}

¹University of Southern California - Thomas Lord Department of Computer Science

²University of Southern California - Information Sciences Institute

ashwinba@usc.edu, vzou@usc.edu, hiteshna@usc.edu, youchris@usc.edu, lluceri@isi.edu, emiliofe@usc.edu

Abstract

In this paper, we introduce the first release of a large-scale dataset capturing discourse on X (a.k.a., Twitter) in the run-up to the 2024 U.S. Presidential Election. Our dataset comprises 46 million publicly available posts on X, collected from May 1, 2024, to November 30, 2024, using a custom-built scraper, which we describe in detail. By employing targeted keywords linked to key political figures, events, and emerging issues, we aligned data collection with the election cycle to capture evolving public sentiment and the dynamics of political engagement on social media. This dataset offers researchers a robust foundation to investigate critical questions about the influence of social media in shaping political discourse, the propagation of election-related narratives, and the spread of misinformation. We also present a preliminary analysis that highlights prominent hashtags and keywords within the dataset, offering initial insights into the dominant themes and conversations occurring in the lead-up to the election.

Dataset — <https://bit.ly/4jRjbZr>

Introduction

Social media has become an influential force in 21st-century politics globally (Woolley and Howard 2018; Allcott and Gentzkow 2017; Metaxas and Mustafaraj 2012). X (formerly Twitter) has been particularly significant in shaping political discourse and influencing public opinion, offering researchers a valuable resource to study the ideologies that are shared, the spread of misinformation, and online campaigns that support political movements and candidates (Abilov et al. 2021; Chen, Deb, and Ferrara 2022; Chen and Ferrara 2023; La Gatta et al. 2023b; Luceri et al. 2019a). With 586 million monthly active users,¹ X facilitates short-form, text-based interactions on a wide range of topics, making it a prime venue for public engagement in political discussions. It serves as a communication hub for prominent figures, including government officials and celebrities, who rely on X’s platform to reach and influence millions. For instance, 2024 presidential candidates Donald Trump and

Kamala Harris, who have garnered over 103 million² and 20 million³ followers respectively, used the platform to promote their campaigns and critique opponents. Such dynamics, where both political figures and their audiences play intertwined roles (Luceri, Cresci, and Giordano 2021), make X a unique platform for analyzing political sentiments and trends in the context of voting events (Deb et al. 2019; Mascaro, Agosto, and Goggins 2016; Bessi and Ferrara 2016), such as the 2024 U.S. Presidential Election.

In this article, we introduce the X 2024 U.S. Presidential Election dataset, which contains posts and metadata capturing this dynamic environment. We collected 22 million publicly available posts from May 1 to November 30, 2024, targeting key political figures and reflecting shifts in issue salience aligned with major electoral milestones. Through continuous data collection using targeted keywords and capturing significant events, our dataset provides researchers with an unprecedented view into how the discourse on X shapes and reflects public opinion around the election.

This dataset enables the detection of emerging trends and pivotal shifts in political discourse, also offering a valuable resource for analyses focused on identifying influence campaigns and foreign information operations (Minici et al. 2024; Cinus et al. 2025). Thereby, we release this dataset to the research community in the hope of enabling further studies on X in the context of voting events.

Related Work

In the run-up to the 2024 U.S. Presidential Election, several efforts have produced valuable datasets capturing political discourse across a variety of social media platforms.

The TikTok 2024 U.S. Presidential Election Dataset captures 3.14M videos published between November 2023 and November 2024, including videos and related transcripts (Pinto et al. 2025). By leveraging both the TikTok Research API and third-party tools, this dataset supports the analysis of linguistic and visual features—enabling investigations into user engagement across different modalities.

Recent efforts have focused on collecting data from X to investigate the platform’s algorithmic content recommendation mechanisms during the 2024 U.S. Presidential Election.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

²<https://x.com/realdonaldtrump>

³<https://x.com/KamalaHarris>

Specifically, one such study (Ye, Luceri, and Ferrara 2025) deployed 120 sock-puppet monitoring accounts to capture tweets from their personalized timelines. This approach offers insights into how content recommendation systems may contribute to the amplification of bias and the deepening of polarization among politically divided user groups.

In contrast, our dataset differs significantly in both scope and methodology. Rather than relying on synthetic monitoring accounts, we focus on organic discourse by collecting tweets from X users engaged in the election conversation. Our data was gathered by scraping X using a wide range of keywords specifically related to the U.S. Election. Importantly, our methodology does not involve tracking the complete activity of any single user; instead, it emphasizes capturing a diverse cross-section of election-related conversations without concentrating on individual behavioral patterns, thereby offering a broader and more representative understanding of the digital political landscape.

Beyond mainstream platforms such as TikTok and X, messaging apps like Telegram play an increasingly pivotal role in the digital political landscape (Cinus et al. 2025). Known for limited content moderation and used by a range of fringe and conspiratorial communities, Telegram served as a conduit for decentralized communication during previous election cycles (Luceri, Cresci, and Giordano 2021). A large-scale dataset comprising over 30,000 public chats and nearly half a billion messages was released to support research into encrypted or semi-encrypted political discourse (Blas, Luceri, and Ferrara 2025). The dataset includes structured chat networks and detailed message content, allowing for analysis of narrative diffusion, coordination strategies, and loosely connected ideological communities.

In this regard, alt-tech platforms like Truth Social offer a distinctive communication environment with minimal moderation and a user base often aligned with conservative ideologies. A publicly available dataset captures 1.5 million posts related to the 2024 election, spanning from February 2022 to November 2024 (Shah et al. 2024). It includes post metadata, user interactions, and multimedia content, enabling analyses of platform-specific dynamics of political engagement within alternative media ecosystems.

While these studies emphasize mainstream and alt-tech platforms, Discord remains comparatively underexplored in political discourse research. As a fast-growing, community-driven platform with decentralized moderation, Discord offers a unique environment to observe grassroots political engagement. A recent study (Buzelin et al. 2025) addresses this gap by analyzing over 30 million messages from political servers discussing the 2024 U.S. Election. Discord servers were classified as Republican-aligned, Democratic-aligned, or unaligned based on their descriptions. Using embedding analysis, the study tracked shifts in conversation during major campaign events and revealed distinct political valences and implicit biases. These findings highlight the role of Discord in shaping online political behavior and underscore its potential as a valuable lens for future research.

These collections provide complementary perspectives on how political narratives, public sentiment, and misinformation propagated throughout the election cycle, and serve as

essential resources for post-election analysis.

Data Collection Framework

X-Scraper Engine

The X-Scraper is a custom-built scraping engine that we developed to gather publicly available data from X with a specific focus on the 2024 U.S. Presidential Election. This tool is designed to collect rich, detailed insights into political discourse on the platform, focusing on posts related to election topics. We extract a variety of post-specific information, including tweet type, content, user engagement, media attachments, and user metadata. While this dataset captures a significant portion of the available data, it is important to note that it is not the entire dataset but rather the data that we were able to scrape from the user interface (UI), reflecting a sample of political discourse occurring on the platform. X-Scraper is built on a headless Chromium driver that automates navigation through the UI.

Additionally, we utilize targeted queries to scrape data from specific timeframes, ensuring that the content we collect aligns with particular periods of political activity. The keyword set was not static; it evolved in real time to reflect emerging events and shifts in public discourse. For example, the keyword “Assassination” was incorporated immediately after the attempted assassination of Donald Trump on July 13, 2024, ensuring relevant posts were captured promptly. Similarly, “JD Vance” was added to the list on July 15, 2024, following his selection as Trump’s vice-presidential running mate. This dynamic keyword adjustment strategy was essential in maintaining a corpus that is both temporally and contextually responsive to unfolding political events, particularly those surrounding the 2024 U.S. Presidential Election.

Finally, we capture metadata from user profiles, providing insights into user behavior, such as the number of followers, tweet frequency, and engagement levels. This user-level data allows us to gain a deeper understanding of how different individuals and communities participate in and contribute to the broader political conversation.

Query Structure

Our custom scraper uses a query structure similar to the X API, specifying parameters such as:

- **since:** Start date for collecting posts.
- **until:** End date for collecting posts.
- **Keywords:** Targeted election-related keywords, structured in quotations. Multiple keywords are included in comma-separated format, with OR logic applied.
- **from:** Filters to include only posts by specific users.
- **filter:** Specifies post type (e.g., retweets, quotes, replies), used to collect specific kinds of tweets.

Example query string:

```
("thedemocrats" OR "DNC" OR "Kamala Harris" OR "Dean Phillips" OR "williamson2024" OR "phillips2024" OR "Democratic party" OR "Republican party" OR "Third Party") until:2024-07-02_00:00:00_UTC since:2024-07-01_00:00:00_UTC
```

This query captures posts made between July 1 and July 2 in UTC that contain any of the specified keywords. For the list of tracked keywords, see the *Appendix*.

Methodology

To ensure a comprehensive and contextually rich dataset, our data collection process employed a temporally segmented approach. The overall observation period was divided into smaller, manageable intervals. Each segment was queried historically using a curated list of relevant keywords along with *from* and *until* parameters, enabling the identification of evolving discourse patterns and the timely capture of significant sociopolitical events.

To maintain data integrity and ensure consistent coverage across all intervals, we performed regular sanity checks. These included periodic audits of timestamp distributions, keyword match density, and manual inspections of collected samples to identify and rectify any potential data gaps or irregularities. These checks were critical in validating that the scraper was functioning as intended and that no significant segments of the timeline were inadvertently omitted.

Given that the scraping process was conducted via X’s UI, the dataset may be subject to the platform’s inherent UI limitations, such as rate-limiting, search visibility constraints, and algorithmic filtering of content. To mitigate this limitation, we utilize 72 distinct accounts to enhance the scalability of the data collection process. As a result, while we cannot claim absolute exhaustiveness, continuous manual monitoring throughout the data collection period established the quality of our pipeline throughout the data collection.

Exploratory Analysis

The dataset underlying this study is structured according to a schema designed to facilitate a comprehensive analysis of both the content and context of tweets. The dataset consists of around 46 million tweets; about 66.2% are reply tweets, 4.8% are retweets, 18.3% are original tweets, and 10.7% are quote tweets. A detailed description of the schema is provided in the *Appendix*. Each entry in the dataset corresponds to an individual tweet and includes a range of fields that collectively enable multifaceted investigation into patterns of communication, engagement, and influence across the platform. The dataset fields are summarized as follows:

- **id**: A unique identifier assigned to each tweet, enabling precise tracking and deduplication of entries across data processing stages.
- **text**: The main body of the tweet, containing the full textual content posted on X. This field serves as the primary source for linguistic, thematic, and sentiment analysis.
- **media**: This field captures metadata on attached multimedia content (e.g., images, videos, GIFs), which are increasingly central to platform engagement strategies.
- **epoch**: A machine-readable timestamp indicating the exact moment of tweet publication. This enables granular temporal analyses, such as detecting surges in discourse around key events, tracing the evolution of narratives, and modeling information diffusion over time.

In addition to content-level data, the schema also captures key engagement metrics that reflect the audience’s interaction with each tweet:

- **replyCount**: The number of direct replies received by a given tweet, indicating conversational depth.
- **retweetCount**: The number of times a tweet has been shared, providing a proxy for virality and amplification.
- **likeCount**: A measure of approval or support for the tweet.
- **quoteCount**: The number of times a tweet has been re-shared with commentary, offering insight into how content is reframed or debated.

Furthermore, we incorporate structural and relational fields such as:

- **conversationId**: A unique identifier linking tweets within the same conversation thread, essential for reconstructing dialogue sequences and analyzing interactive discourse.
- **mentionedUsers**: A list of users mentioned within a tweet, which helps map user interaction networks and identify patterns of attention, alliance, contention, and coordination among individuals and groups.

Collectively, these attributes enable a detailed reconstruction of user behavior on X in the context of the 2024 U.S. Election. The schema supports both micro-level investigations—such as the rhetorical strategies employed by individual users—and macro-level analyses, including the diffusion of political narratives, the emergence of new viewpoints, and the role of different classes of users (influencers, bots, coordinated actors, etc.) in shaping public discourse.

Top Keywords

Table 1 highlights the ten most frequent keywords, with “Biden” and “Trump” leading. This prominence reflects their centrality in U.S. political discourse, underscoring their roles as focal points of online conversations. The keyword “MAGA” (Make America Great Again) signals sustained engagement with the conservative base, suggesting that identity-based slogans maintain significant relevance in election debate. The inclusion of “GOP,” “Harris”, and “conservative” illustrates the prominent alignment of discourse along party and ideological lines.

Keyword	Frequency
Biden	16,526,765
Trump	11,138,357
MAGA	5,501,047
Harris	3,498,106
Donald	3,479,169
Kamala	3,170,842
People	2,923,784
President	2,797,223
Joe	2,486,667
GOP	2,343,743

Table 1: Top 10 Keywords

Top Social Media Domains

Table 2 lists the top domains referenced, highlighting the popularity of YouTube videos for multimedia content sharing, similar to previous election-related findings (La Gatta et al. 2023a). News sites like Fox News and Breitbart underscore how major political narratives are disseminated through recognized media outlets, shaping public discourse by providing prominent narratives to a wide audience.

Domain	Occurrences
youtube.com	647,999
X.com	348,742
foxnews.com	134,298
breitbart.com	97,964
share.newsbreak.com	80,954
msn.com	77,615
l.smartnews.com	72,687
rumble.com	49,472
apple.news	42,898
tiktok.com	38,478

Table 2: Top 10 Social Domains

Top Mentions

Table 3 displays the accounts most frequently mentioned, showing that public figures and political entities play a central role in driving discourse on X. Mentions of candidates and key figures such as Joe Biden, and Donald Trump illustrate the polarized nature of political conversations, where users consistently engage with or critique prominent personalities. Elon Musk’s high mention frequency reflects his influence over the platform’s direction, offering researchers an important perspective on how high-profile individuals impact public discourse (Ye, Luceri, and Ferrara 2025).

Mention	Occurrences
@GOP	423,427
@JoeBiden	358,955
@TheDemocrats	208,307
@POTUS	199,869
@elonmusk	146,573
@realDonaldTrump	139,868
@YouTube	127,044
@harryjsson	108,825
@KamalaHarris	103,781
@GuntherEagleman	101,248

Table 3: Top 10 Mentions

Top Languages

Table 4 summarizes the top languages in our dataset based on tweet metadata. The majority of tweets are in English, comprising over 39 million posts, followed by a significantly smaller proportion in Spanish, French, and Portuguese. The presence of multilingual content—ranging from Estonian to Tagalog—reflects the international visibility and reach of the

Language	Occurrences
English (en)	39,632,860
Spanish (es)	882,304
Undefined (und)	381,215
French (fr)	350,089
Estonian (et)	267,399
Portuguese (pt)	267,193
Japanese (ja)	211,754
German (de)	185,138
Tagalog (tl)	172,348

Table 4: Top Languages

2024 U.S. Presidential Election discourse, even on platforms primarily dominated by U.S.-based users.

Top Hashtags

Hashtag	Occurrences
#maga	1,057,632
#trump2024	957,478
#trump	355,499
#biden	257,772
#bidenharris2024	164,446
#donaldtrump	88,199
#kamalaharris	75,222
#usa	66,438
#maha	64,502
#gop	55,424
#trumpvance2024	49,755
#biden2024	32,761
#foxnews	32,263
#election2024	30,695
#joebiden	28,422
#harriswalz2024	26,818
#harris	25,235
#kamala	24,913
#maga2024	24,025
#smartnews	23,930

Table 5: Top 20 Hashtags

Table 4 reveals the top 20 hashtags, illustrating major candidates and slogans such as #maga, #trump2024, and #bidenharris2024. These hashtags underscore significant public engagement around key figures and movements, reflecting the electorate’s ideological divides. The frequent appearance of #trumpvance2024 and #harriswalz2024 further indicates active digital advocacy and long-term campaign strategies. Such recurring terms highlight how social media discourse aligns with, and at times shapes, campaign narratives and public sentiment during election cycles.

Conclusions

This article introduces a comprehensive dataset of posts from X related to the 2024 U.S. Presidential Election, collected between May 1, 2024, and November 30, 2024. Utilizing our custom-built X-Scraper Engine, which is designed

to adapt to real-time events, we gathered a broad range of election-related content, including posts, metadata, and user information. This dataset provides a unique resource for researchers to analyze trends in public opinion, investigate the spread of misinformation, and examine the influence of key figures on X. Given the extremely polarized nature of the 2024 election cycle and the impact of social media on shaping public perceptions, this dataset has significant potential to help understand how information and narratives are shared and propagated. While acknowledging limitations related to the representativeness of data from X, we believe this dataset offers a useful sample of election-related content, providing a strong foundation for ongoing research into the dynamics between social media and political processes in the context of the 2024 U.S. Election.

Limitations

While our data collection methodology is designed to capture a substantial and timely portion of public discourse on X, several inherent limitations must be acknowledged.

First, data acquisition is conducted through the user interface (UI) of X, rather than through direct API access. As a result, the scraper may not consistently represent the entirety of the platform’s activity across the observation period. Although continuous manual monitoring and interval-based queries help minimize data gaps, the reliance on UI scraping inherently carries risks of incomplete coverage—particularly in cases of rapid event surges or shifts in discourse.

Second, the dynamic and frequently updated nature of X’s front-end interface introduces additional challenges. Changes to UI components occasionally disrupt scraper performance, leading to intermittent delays in data collection. While such interruptions were typically brief and mitigated with timely updates to the scraper codebase, they may have temporarily affected data completeness during certain critical windows.

Third, the keyword-based search strategy imposes structural constraints. X’s native search engine supports only a limited number of keywords per query. To ensure broad coverage across themes, topics, and emerging events, the scraping process required multiple passes with distinct keyword groupings. While this iterative approach allows for adaptive data collection, it is inherently limited by the precision and recall of keyword matching, potentially excluding relevant content that falls outside defined query parameters.

Moreover, data collection is subject to a fixed retrieval ceiling imposed by X’s terms of service. Specifically, the scraper is rate-limited to approximately 2,300 posts per hour per account. To optimize throughput while remaining compliant, we employed multiple personal accounts in a distributed fashion. Nevertheless, this constraint may still limit real-time responsiveness, particularly during periods of high-volume activity (e.g., political debates, breaking news, or unexpected events).

Collectively, these limitations underscore the importance of interpreting the dataset as a representative—but not exhaustive—sample of public discourse on X during the 2024 U.S. election cycle.

Future Work

Future work will focus on longitudinal analysis of political discourse and trends related to the election. We also plan to explore deeper patterns by examining verified prominent users (Nogara et al. 2022), automated accounts (Luceri et al. 2019b), and state-sponsored human operators (Luceri, Giordano, and Ferrara 2020) to understand their roles in spreading information, shaping ideologies, and responding to real-time events. This will help study the behaviors and strategies behind both authentic and inauthentic activity, providing a holistic view of platform usage during election times.

Additionally, we aim to incorporate data from complementary sources, such as cross-platform interactions and broader demographic metrics, to contextualize X discourse within the wider digital media landscape. This expanded approach will allow researchers to assess the platform’s unique role and interaction with other social media channels, leading to a richer understanding of election-related discourse across the broad online ecosystem.

Ethical Statement

Scraping of publicly available data is in compliance with the guidelines provided by the US Supreme Court.⁴ X’s Terms of Service (ToS) allow data redistribution for research purposes. Researchers should handle the data responsibly, ensure compliance with ethical guidelines, and download the datasets in small segments to comply with X’s ToS. PII is not disclosed to protect privacy. Researchers should not attempt to de-anonymize or reconstruct sensitive information.

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*, 901–912.
- Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.
- Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. *First monday*, 21(11-7).
- Blas, L.; Luceri, L.; and Ferrara, E. 2025. Unearthing a Billion Telegram Posts about the 2024 US Presidential Election: Development of a Public Dataset. In *ACM Web Conference (WWW’25)*.
- Buzelin, A.; Dutenhefner, P. R.; Locatelli, M. S.; Malaquias, S.; Bento, P.; Aquino, Y.; Dayrell, L.; Estanislau, V.; Santana, C.; Alzamora, P.; Vasconcelos, M.; Jr., W. M.; and Almeida, V. 2025. Analyzing Political Discourse on Discord during the 2024 U.S. Presidential Election. arXiv:2502.03433.
- 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*, 1–18.

⁴<https://techcrunch.com/2022/04/18/web-scraping-legal-court>

- Chen, E.; and Ferrara, E. 2023. Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia. In *Proceedings of the 17th International AAAI Conference on Web and Social Media*, 1006–1013.
- Cinus, F.; Minici, M.; Luceri, L.; and Ferrara, E. 2025. Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 US Election. In *ACM Web Conference (WWW'25)*.
- Deb, A.; Luceri, L.; Badawy, A.; and Ferrara, E. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Proceedings of the 2019 World Wide Web Conference*, 237–247.
- La Gatta, V.; Luceri, L.; Fabbri, F.; and Ferrara, E. 2023a. The interconnected nature of online harm and moderation: Investigating the cross-platform spread of harmful content between youtube and twitter. In *Proceedings of the 34th ACM conference on hypertext and social media*, 1–10.
- La Gatta, V.; Wei, C.; Luceri, L.; Pierri, F.; and Ferrara, E. 2023b. Retrieving false claims on Twitter during the Russia-Ukraine conflict. In *Companion Proceedings of the ACM Web Conference 2023*, 1317–1323.
- Luceri, L.; Cresci, S.; and Giordano, S. 2021. Social media against society. *The Internet and the 2020 Campaign*, 1.
- Luceri, L.; Deb, A.; Badawy, A.; and Ferrara, E. 2019a. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion Proceedings of the 2019 World Wide Web Conference*, 1007–1012.
- Luceri, L.; Deb, A.; Giordano, S.; and Ferrara, E. 2019b. Evolution of Bot and Human Behavior During Elections. *First Monday*, 24(9).
- Luceri, L.; Giordano, S.; and Ferrara, E. 2020. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 417–427.
- Mascaro, C.; Agosto, D.; and Goggins, S. P. 2016. The method to the madness: The 2012 United States presidential election Twitter corpus. In *Proceedings of the 7th 2016 International Conference on Social Media & Society*, 1–9.
- Metaxas, P. T.; and Mustafaraj, E. 2012. Social media and the elections. *Science*, 338(6106): 472–473.
- Minici, M.; Luceri, L.; Cinus, F.; and Ferrara, E. 2024. Uncovering Coordinated Cross-Platform Information Operations Threatening the Integrity of the 2024 US Presidential Election Online Discussion. *First Monday*.
- Nogara, G.; Vishnuprasad, P. S.; Cardoso, F.; Ayoub, O.; Giordano, S.; and Luceri, L. 2022. The disinformation dozen: An exploratory analysis of covid-19 disinformation proliferation on twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, 348–358.
- Pinto, G.; Bickham, C.; Salkar, T.; Luceri, L.; and Ferrara, E. 2025. Tracking the 2024 US Presidential Election Chatter on Tiktok: A Public Multimodal Dataset. In *ACM Web Conference (WWW'25)*.
- Shah, K.; Gerard, P.; Luceri, L.; and Ferrara, E. 2024. Unfiltered Conversations: A Dataset of 2024 U.S. Presidential Election Discourse on Truth Social. arXiv:2411.01330.
- Woolley, S. C.; and Howard, P. N. 2018. *Computational propaganda: Political parties, politicians, and political manipulation on social media*. Oxford University Press.
- Ye, J.; Luceri, L.; and Ferrara, E. 2025. Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X Approaching the 2024 US Presidential Election. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT'25)*.

Appendix

Keywords
2024 Elections
2024 Presidential Election
Biden
Biden2024
conservative
CPAC
Donald Trump
GOP
Joe Biden and Kamala Harris
Joe Biden
Joseph Biden
KAG
MAGA
Nikki Haley
RNC
Ron DeSantis
Snowballing
Trump2024
trumpsupporters
trumptrain
US Elections
thedemocrats
DNC
Kamala Harris
Marianne Williamson
Dean Phillips
williamson2024
phillips2024
Democratic Party
Republican Party
Third Party
Green Party
Independent Party
No Labels
RFK Jr
Robert F. Kennedy Jr
Jill Stein
Cornel West
utramaga
voteblue2024
letsgobrandon
bidenharris2024
makeamericagreatagain
Vivek Ramaswamy

Table 6: Keywords used in the data collection

Field Name	Data Type	Description
id	object	Unique identifier for each entry.
text	object	Text content of the tweet.
url	object	URL associated with the tweet or content.
epoch	object	Epoch timestamp when the tweet was created.
media	object	Media content included in the tweet (images, videos, etc.).
retweetedTweet	object	Content of the retweeted tweet, if applicable.
retweetedTweetID	object	ID of the retweeted tweet.
retweetedUserID	object	ID of the user who originally tweeted the retweeted content.
id_str	object	ID of the tweet as a string (alternative format).
lang	object	Language of the tweet content.
rawContent	object	Raw unprocessed text of the tweet.
replyCount	object	Number of replies to the tweet.
retweetCount	object	Number of retweets.
likeCount	object	Number of likes.
quoteCount	object	Number of quotes.
conversationId	object	ID of the conversation the tweet is part of.
conversationIdStr	object	Conversation ID as a string.
hashtags	object	Hashtags included in the tweet.
mentionedUsers	object	Users mentioned in the tweet.
links	object	External links included in the tweet.
viewCount	object	View count of the tweet.
quotedTweet	object	Content of the quoted tweet, if applicable.
in_reply_to_screen_name	object	Screen name of the user being replied to.
in_reply_to_status_id_str	object	ID of the tweet being replied to as a string.
in_reply_to_user_id_str	object	User ID of the user being replied to as a string.
location	object	Location information of the tweet or user.
cash_app_handle	object	Cash App handle mentioned in the tweet, if applicable.
user	object	User information or metadata.
date	object	Date of the tweet.
_type	object	Type of tweet (e.g., original, reply, retweet).
user_id	float64	ID of the user as a float.

Table 7: Description of the Data Schema

Paper 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**
- (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**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes**
- (e) Did you describe the limitations of your work? **Yes**
- (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**
- (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)? **Yes**, we release source code via an URL.

- (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, **without compromising anonymity**...

- (a) If your work uses existing assets, did you cite the creators? **NA**
- (b) Did you mention the license of the assets? **NA**
- (c) Did you include any new assets in the supplemental material or as a URL? **Yes**, we are releasing code and data.
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **NA**
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **Yes**
- (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? **NA**
- (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? **NA**

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

- (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 de-identified? **NA**