

The 2024 Election Integrity Initiative

Tracking the 2024 US Presidential Election Chatter on Tiktok:
A Public Multimodal Dataset

Gabriela Pinto, Charles Bickham, Tanishq Salkar, Luca Luceri, Emilio Ferrara

University of Southern California
HUMANS Lab – Working Paper No. 2024.3

Tracking the 2024 US Presidential Election Chatter on Tiktok: A Public Multimodal Dataset

Gabriela Pinto, Charles Bickham, Tanishq Salkar, Luca Luceri, Emilio Ferrara
University of Southern California

Abstract

This paper documents our release of a large-scale data collection of TikTok posts related to the upcoming 2024 U.S. Presidential Election. Our current data comprises 1.8 million videos published between November 1, 2023, and May 26, 2024. Its exploratory analysis identifies the most common keywords, hashtags, and bigrams in both Spanish and English posts, focusing on the election and the two main Presidential candidates, President Joe Biden and Donald Trump.

We utilized the TikTok Research API, incorporating various election-related keywords and hashtags, to capture the full scope of relevant content. To address the limitations of the TikTok Research API, we also employed third-party scrapers to expand our dataset. The dataset is publicly available at <https://github.com/gabbypinto/US2024PresElectionTikToks>.

Introduction

Social media has profoundly transformed electoral politics, emerging as a critical platform for disseminating election-related information. This shift has garnered significant attention from researchers [1, 5–7, 9]. Twitter, in particular, has been instrumental in providing datasets that aid the study of global geopolitical events [2–4].

Meanwhile, TikTok, a short-form video app, has rapidly grown into a major platform for engaging and informing users on a variety of topics. With over a billion users worldwide,¹ TikTok is especially popular among adolescents [10]. Given its rising popularity, TikTok has the potential to become a central hub for disseminating information about the upcoming 2024 election.

As TikTok's influence expands, it has become a strategic platform for politicians aiming to reach young voters. For instance, Donald Trump joined TikTok in June 2024 and quickly amassed over 6 million followers,² while Joe Biden, with more than 373,000 followers, has posted over 200 videos since February 2024.³ The platform's growing popularity and diverse content formats provide a wealth of data that can reflect electoral sentiments and trends.

In this paper, we introduce the TikTok 2024 U.S. Presidential election dataset. This multimodal dataset, which includes both video and text data, aims to offer researchers a comprehensive view of TikTok's role in the election. By capturing political discourse on the platform and updating keywords as political campaigns progress, we hope this dataset will enable valuable insights into the evolving landscape of political communication and engagement on TikTok.

¹<https://whatsthebigdata.com/tiktok-statistics/>

²<https://cnn.com/2024/06/02/politics/donald-trump-joins-tiktok/index.html>

³<https://cnn.com/2024/06/12/tech/tiktok-pew-research-politics-x/index.html>

Method of Data Collection

TikTok API

The TikTok Research API⁴ facilitates the retrieval of detailed information regarding accounts and content on TikTok. Account-related data like user profiles, followers and following lists, liked videos, pinned videos, and reposted videos. Content details include comments, captions, subtitles, and the number of comments, shares, and likes that a video receives. Accessing data via this API requires an access token obtained using the user's credentials (i.e., client key and secret key). This bearer token, accompanied by specific parameters in the request body, allows the fetch of desired data.

Query

The TikTok API allows us to specify specific parameters to get desired data from the API. We define the parameters in the request body, the authentication token, and the format of the response expected from the API. These parameters include the range of dates of publication and the metadata fields of the videos needed in the response. The API allows us to get the desired data of videos published in specific geographical regions, videos made on particular topics, videos using different languages, and responses to the videos in the form of likes, shares, and comments.

- Start date: The starting date and time window from which to extract data.
- End date: The ending date and time window until which to extract data.
- Fields: The specific metadata of the videos required in the response, such as video identifier (id), description (video_description), creation time (create_time), region code (region_code), share count (share_count), view count (view_count), like count (like_count), comment count (comment_count), music identifier (music_id), hashtag names (hashtag_names), and username (username).
- Max_count: Specifies the number of records to be returned in a call, with a maximum of 100 records per request.
- Query Body: The TikTok API also allows the use of a query body to create more detailed queries, similar to SQL, by using logical operators such as AND and OR on specified fields like region code, hashtag names, likes etc. For example, to collect metadata of videos published in the United States with the hashtag "#elections2024" by specifying "US" in region_code field and #elections2024 in hashtag field, the necessary conditions can be specified in the query body.

In this study, we are collecting metadata of videos published in the US using the TikTok Research API. The data collection encompasses all the fields mentioned previously, including share count and view count, such as count, comment count, music identifier, hashtag names, username, etc. Additionally, we are incorporating a list of keywords and hashtags, as detailed in the subsequent sections.

Keywords and Hashtags

The hashtags/keywords detailed in Table 5 (cf. Appendix) illustrate the inclusion and exclusion of keywords within the indicated phases. Table 1 details each phase's start and end dates. Those marked with ‘-’ means that their corresponding keywords/hashtags were excluded from the query;

⁴<https://developers.tiktok.com/doc/research-api-specs-query-videos/>

Table 1. Phases and the dates ranges.

Phase #	(MM/DD/YYYY)-(MM/DD/YYYY)
1	11/01/2023 - 01/15/2024
2	01/16/2024 - 02/07/2024
3	02/08/2024 - 02/27/2024
4	02/28/2024 - 03/05/2024
5	03/06/2024 - 05/26/2024

conversely, those marked with ‘+’ represent the inclusion. We created the list of keywords and phases based on significant political events that gained massive media attention.

We set the end of *Phase 1* on January 15, 2024, the day of the Iowa caucus.⁵ Thus, January 16, 2024, was the new start date for *Phase 2*. In *Phase 2* and for subsequent phases, we analyzed the most common keywords/hashtags within our data at the time of collection. The most frequent keywords and hashtags related to the US Presidential Elections were added to the query. On January 21, 2024, Republican nominee Ron DeSantis suspended his presidential campaign.⁶ Consequently, we removed Ron DeSantis in *Phase 2*. On February 7, 2024, Democratic nominee Marianne Williamson announced the end of her presidential campaign due to her loss at the Nevada Democratic Primary to President Joe Biden.⁷ Therefore, February 8, 2024, became the start date for *Phase 3*, where keywords/hashtags related to Marianne Williamson and Ron DeSantis were removed. *Phase 3* spanned from February 8, 2024, to February 27, 2024, since we wanted to reevaluate our data and extract the most popular keywords and hashtags. We updated the query by including the most reoccurring and relevant keywords and hashtags. *Phase 4* spanned between February 28, 2024 to March 5, 2024. The end date of March 5, 2024 was established due to the announcement of Democratic candidate Dean Phillips⁸ and Republican Nikki Haley⁹ announced their end to their presidential campaign. Given the lack of media attention on social media on Dean Phillips, we removed the related keywords/hashtags for *Phase 5*. March 6, 2024, was the start date of *Phase 5*; however, keywords and hashtags related to Nikki Haley were included until June 17, 2024, since she received more media attention than Dean Phillips.

Exploratory Analysis

Data Access

The dataset is publicly available on Github,¹⁰ where we provide the ID of each video collected. We will update the repository consistently for future collections as we collect more videos. In compliance with the Terms of Service of TikTok’s Research API, we can only publish the IDs since publishing data that risks users’ privacy is strictly prohibited.

Summary of Data

In Table 2, we present a summary of the statistics in the current version of the dataset presented in this paper. The proportion of the number of transcripts to the total number of videos is approximately 14.7% collected. The low proportion of TikTok-generated transcripts is one of the limitations of the

⁵<https://www.nytimes.com/interactive/2024/01/15/us/elections/results-iowa-caucus.html>

⁶<https://www.politico.com/news/2024/01/21/desantis-ends-presidential-campaign-00136839>

⁷<https://www.politico.com/news/2024/02/07/marianne-williamson-drops-out-2024-00140297>

⁸<https://www.politico.com/news/2024/03/06/dean-phillips-drops-out-00145403>

⁹<https://www.wsj.com/politics/elections/nikki-haley-drops-out-2024-presidential-election-625277ca>

¹⁰<https://github.com/gabbypinto/US2024PresElectionTikToks>

Table 2. Summary statistics of the dataset.

Number of videos	1,799,333
Number of transcripts	266,202
Number of comments	93,776,960
Number of views	3,535,119,560
Number of likes	871,474,292
Number of shares	140,038,704
Number of unique users	432,951

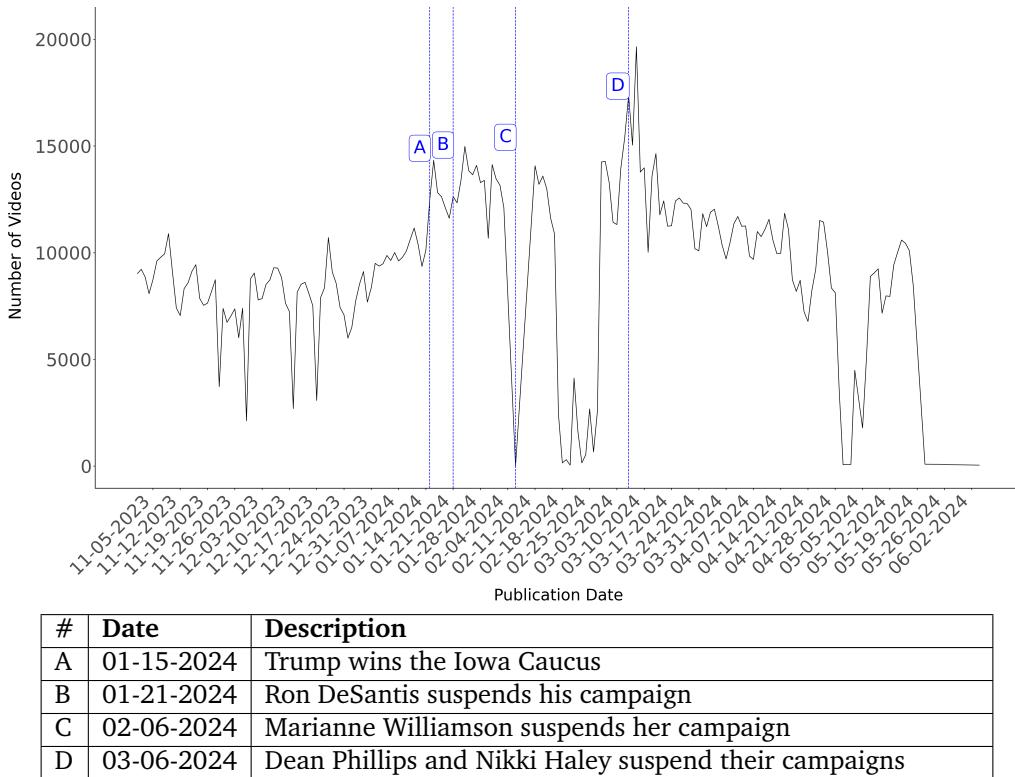


Fig. 1. Timeline of events and volume of TikTok posts.

TikTok Research API. Therefore, to provide more rich linguistic data on the published content, we will upload transcripts generated by Whisper¹¹ in our GitHub repository.

Number of Videos over time

In this dataset's first deployment, we collected video metadata posted between November 1, 2023, and May 26, 2024. Figure 1 shows the number of videos collected concerning its publication date. It is important to note that there are notable gaps within our data, which is content that we couldn't collect with respect to its publication date via the TikTok Research API. Table 3 contains each gap's start and end date inclusively. To solve this issue, we currently use third-party scrapers to collect video content published within the stated dates.

In addition, we also provided the dates of critical political events during the election cycle. We are currently using additional scrapers¹² to collect more data and fill in the gaps to obtain the complete discourse on TikTok from November 1, 2023, to January 1, 2025.

¹¹<https://github.com/openai/whisper>

¹²<https://github.com/davidteather/TikTok-Api>

Table 3. The range for each gap within our dataset.

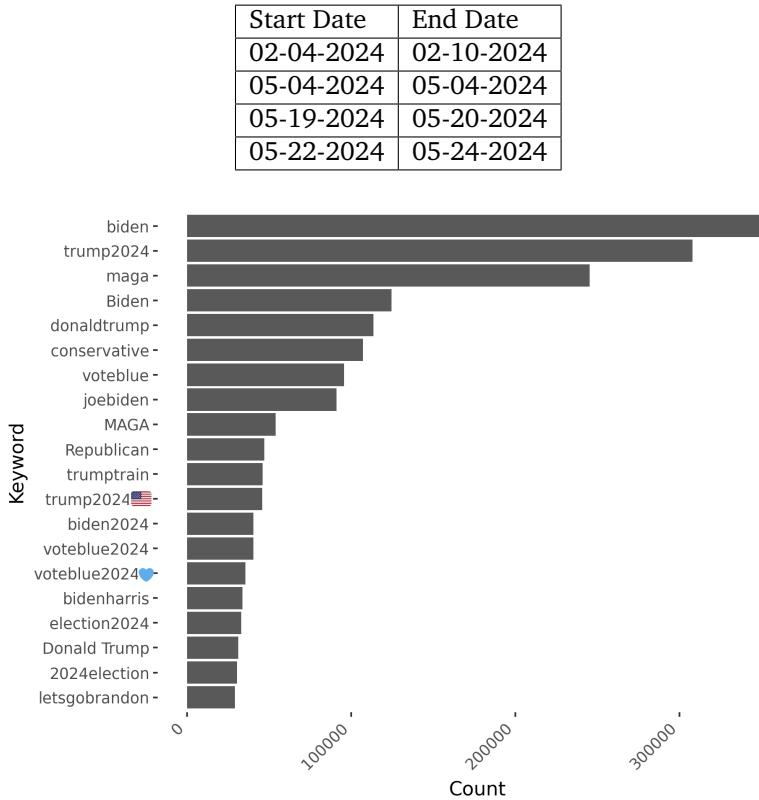


Fig. 2. Top Keywords within our query that appear in the 'video_description' attribute

Top Keywords mentioned in the video descriptions

The intersection between the words in the 'video_descriptions' attribute (the caption the user has posted along with the video) and the keywords shown in Table 5. Figure 2, shows the 20 most reoccurring keywords and their count. The frequency reflects that the video content is mainly related to President Joe Biden and Donald Trump.

Top Hashtags

Table 4 presents the 20 most frequent hashtags for each video. Each video collected through the TikTok Research API provides a list of hashtags labeled 'hashtag_names.' Overall, the hashtags reflect that the videos labeled by the creator were related to Donald Trump, President Joe Biden, and the right-leaning ideology through keywords such as "maga" ("Make America Great Again") and "republican."

Language Detection

We applied LangDetect¹³ on the transcripts for language detection. In Figure 3, we show the languages within the transcripts generated by the TikTok Research API, excluding English. In our current dataset, we classified the 252,155 transcripts as English-based content. The second is Spanish-based, with 8,698 transcripts. In future work, we aim to provide an analysis of Spanish-based content to gain a more complete overview of the content published during the election cycle.

Most common bigrams in English and Spanish TikTok-generated transcripts.

We generated the bigrams for 252,155 of the English TikTok-generated transcripts and 8,698 of Spanish Tik-Tok generated transcripts, shown in Figure 4 and Figure 5, respectively. Presidents Biden and Donald Trump are frequently mentioned based on the 200 most common bigrams in

¹³<https://pypi.org/project/langdetect/>

Table 4. Top 20 Most Frequently Occurred Hashtags.

Tag	Count
trump	407,541
trump2024	285,840
biden	198,247
maga	170,354
duet	146,676
republican	144,067
usa	130,607
donaldtrump	113,514
politics	107,345
democrat	99,747
news	95,608
america	92,740
joebiden	86,267
trending	85,231
conservative	76,203
fjb	69,401
capcut	62,360
voteblue	61,309
democrats	53,799
election	52,119

Spanish and English transcripts. Other common bigrams include Supreme Court ("corte suprema" in Spanish), New York ("nueva york" in Spanish), and social media ("redes sociales") in Spanish. In Spanish-based and English-based content, it is clear that the main topic discussed is related to the U.S. Presidential Elections.

Conclusions

This paper presents the current process of collecting TikTok content related to the upcoming U.S. Presidential Elections. The primary method for data collection used in this study was the TikTok Research API; however, due to the limitations mentioned earlier, we plan to use third-party scrapers for a thorough and complete analysis of the discourse. Based on our exploratory study on the 1,799,333 videos, the most frequent hashtags and keywords within the 'video_description' attributes are related to President Joe Biden and Donald Trump. Within our dataset, 266,202 videos included a transcript generated by the TikTok API. Of these, 252,155 of the transcripts were detected to be written in English and 8,698 in Spanish. Within those transcripts, in addition to Donald Trump and President Joe Biden, the Supreme Court and New York were frequent bigrams mentioned in both transcripts.

Limitations

The used API has certain limitations. The bearer token, used for authentication, remains valid for only two hours and must be regenerated after expiration to continue fetching the data. Each API call retrieves metadata for only 100 videos at a time. Additionally, when extracting more data for a specific date, a wait time must be included. If the search ID from a previous call is used to fetch subsequent records for the same day without this wait time, an error indicating an invalid search ID is thrown. Furthermore, a wait time must be observed after each request to comply with the API's rate limiting. Also, the API has a limit of 1,000 calls per day.

Future Work

To address the limitations in this paper and the gaps within our dataset, we will utilize third-party scrapers to collect more video metadata. To conduct a more in-depth analysis, we also plan to collect

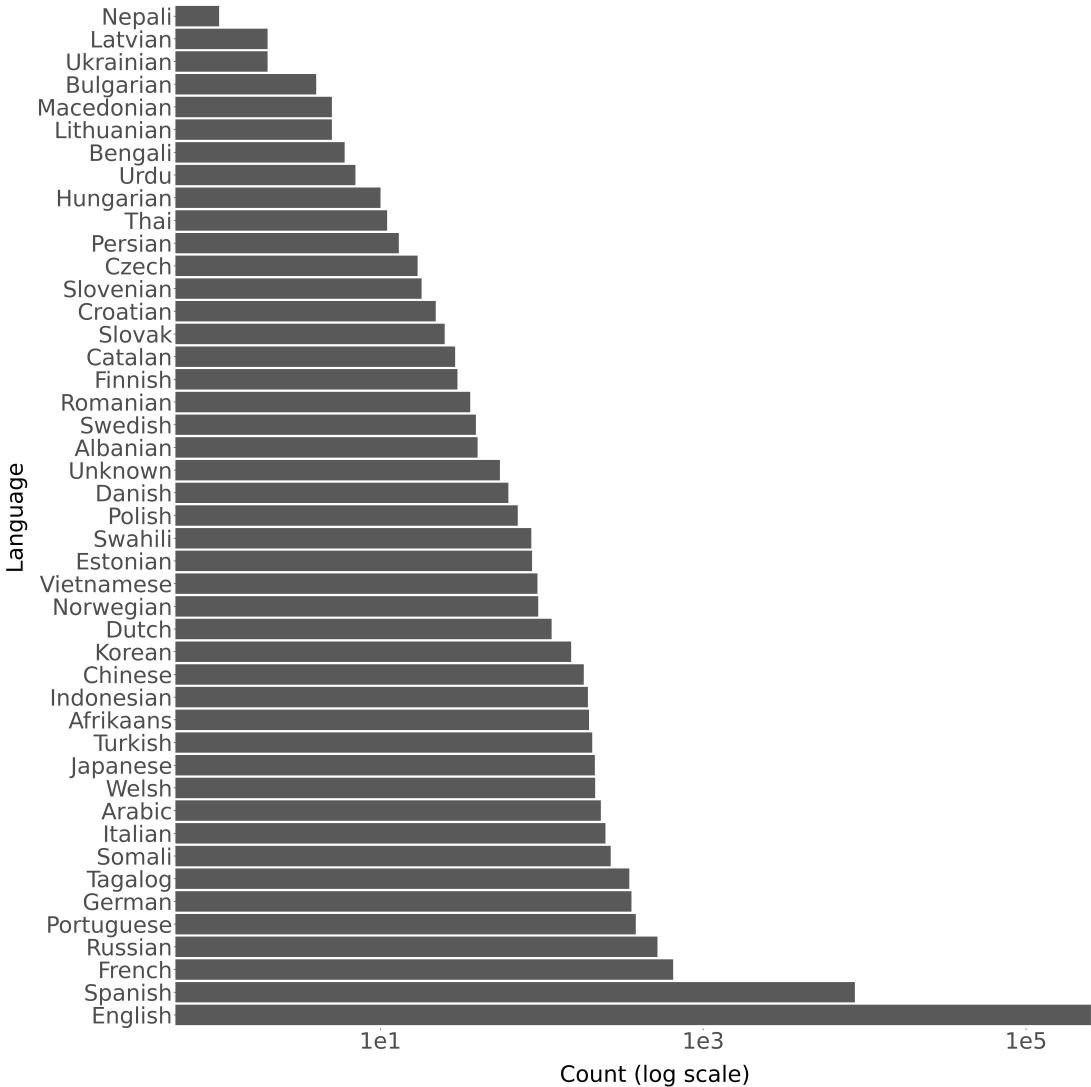


Fig. 3. Classification of languages used within the transcripts (log scale)

other attributes within a video, as shown in Figure 6. While collecting more metadata, we are in the process of collecting videos to analyze the content using VideoLLMs such as Video-LLaMa[11] or LLaVa-NeXT-Video[8]. We will collect the comments and comment replies for each video using a third-party scraper to study the discussions between users within the comment section. Due to the low presence of transcripts for each video, we will extract the audio and use Whisper to generate the transcript for each video.

Data Availability

Our data collection will continue uninterrupted for the foreseeable future. As the election approaches, we anticipate that the amount of data will grow significantly. The data set available on GitHub is released in compliance with the Tiktok's Terms and Conditions, under which we are unable to publicly release the videos of the collected posts. We are, therefore, releasing the Video IDs, which are unique identifiers tied to specific posts. The Video IDs can be used by researchers to query Tiktok's API and obtain the complete video objects, including multimedia and metadata information as depicted in Figure 6. A publicly accessible GitHub repository that we will continue to routinely update is available at <https://github.com/gabbypinto/US2024PresElectionTikToks>

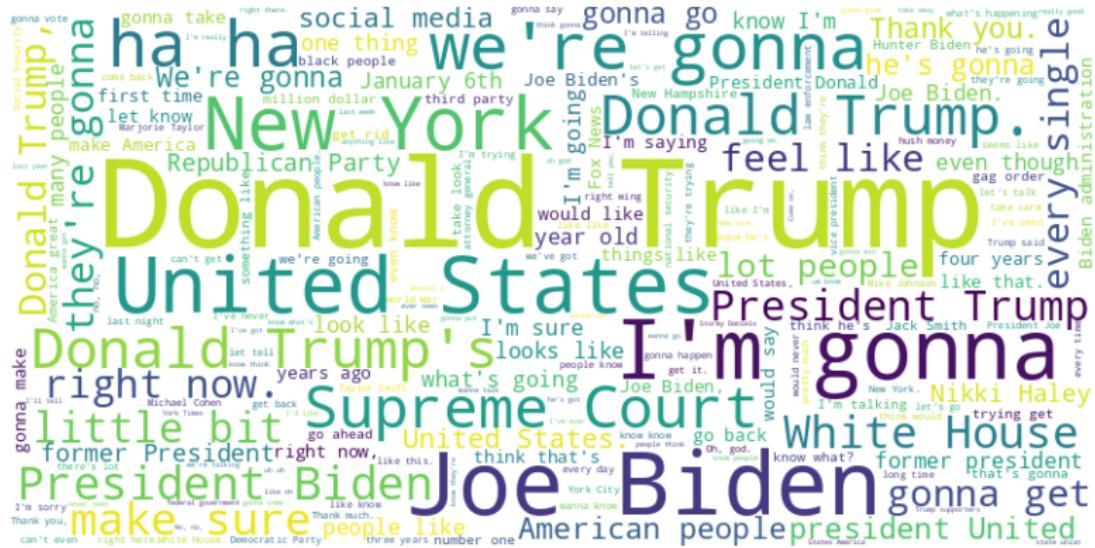


Fig. 4. 200 of the most frequent bigrams in the English Transcripts

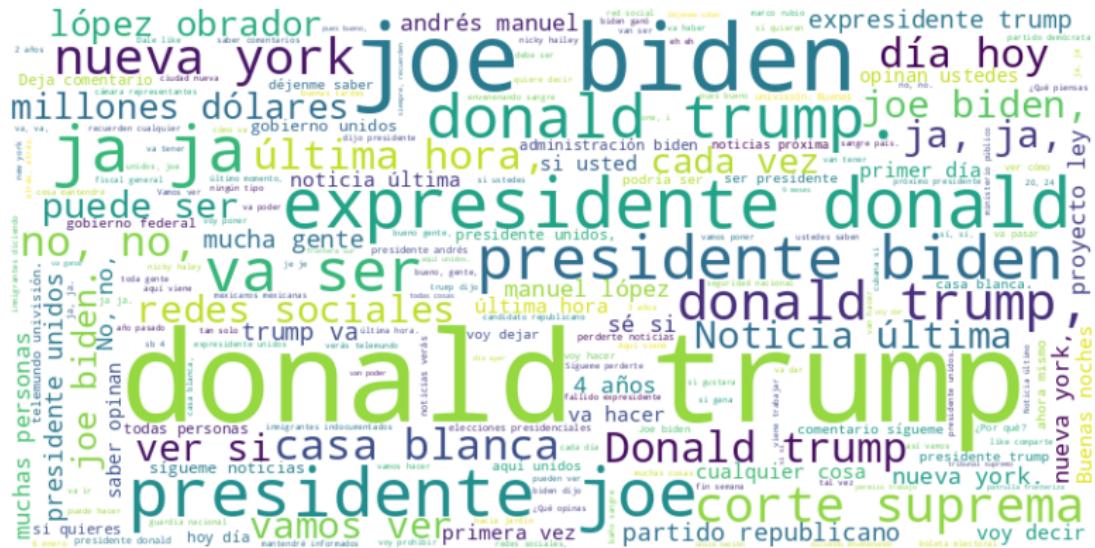


Fig. 5. 200 of the most frequent bigrams in the Spanish Transcripts

About the Team

The 2024 Election Integrity Initiative is carried out by a collective of USC students and volunteers whose contributions are instrumental to enable these studies. The authors are grateful to the following HUMANS Lab's members for their tireless efforts on this project: Ashwin Balasubramanian, Leonardo Blas, Keith Burghardt, Sneha Chawan, Vishal Reddy Chintham, Eun Cheol Choi, Srilatha Dama, Priyanka Dey, Isabel Epistelomogi, Saborni Kundu, Grace Li, Richard Peng, Jinhui Qi, Ameen Qureshi, Namratha Sairam, Srivarshan Selvaraj, Kashish Atit Shah, Gokulraj Varatharajan, Reuben Varghese, Siyi Zhou, and Vito Zou.

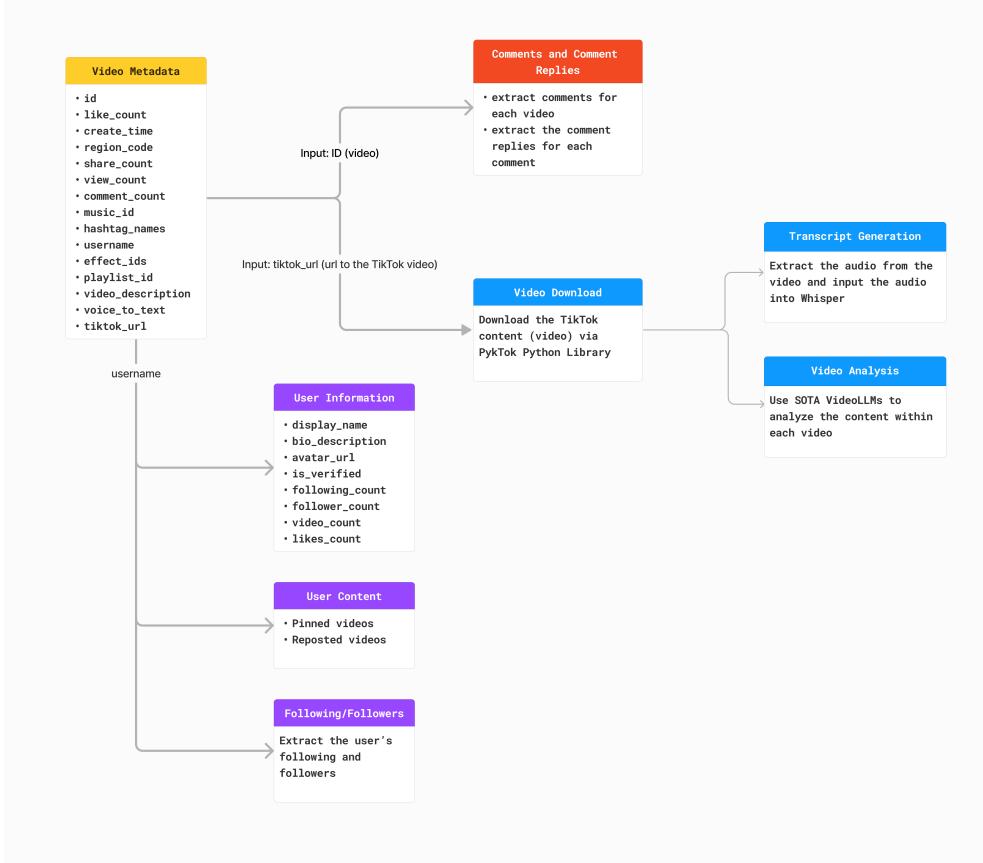


Fig. 6. Data Collection Schema

References

- [1] Anton Abilov, Yiqing Hua, Hana Matatov, Ofra Amir, and Mor Naaman. 2021. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. In *Proc. Int. AAAI Conference on Web & Social Media*. 901–912.
- [2] Emily Chen, Ashok Deb, and Emilio Ferrara. 2022. #Election2020: the first public Twitter dataset on the 2020 US Presidential election. *Journal of Computational Social Science* (2022), 1–18.
- [3] Emily Chen and Emilio Ferrara. 2023. Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia. In *Proc. 17th Int. AAAI Conference on Web & Social Media*. 1006–1013.
- [4] Clayton A Davis, Giovanni Luca Ciampaglia, Luca Maria Aiello, Keychul Chung, Michael D Conover, Emilio Ferrara, Alessandro Flammini, Geoffrey C Fox, Xiaoming Gao, Bruno Gonçalves, et al. 2016. OSoMe: the IUNI observatory on social media. *PeerJ Computer Science* 2 (2016), e87.
- [5] Ashok Deb, Luca Luceri, Adam Badawy, and Emilio Ferrara. 2019. Perils and Challenges of Social Media and Election Manipulation Analysis: The 2018 US Midterms. In *Proc. 2019 World Wide Web Conference*. 237–247.
- [6] Andreas Jungherr. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics* 13, 1 (2016), 72–91.
- [7] Nane Kratzke. 2017. The #btw17 Twitter dataset—recorded tweets of the federal election campaigns of 2017 for the 19th German Bundestag. *Data* 2, 4 (2017), 34.
- [8] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. (January 2024). <https://llava-vl.github.io/blog/2024-01-30-llava-next/>
- [9] Luca Luceri, Ashok Deb, Silvia Giordano, and Emilio Ferrara. 2019. Evolution of bot and human behavior during elections. *First Monday* 24, 9 (2019).
- [10] Christian Montag, Haibo Yang, and Jon D Elhai. 2021. On the psychology of TikTok use: A first glimpse from empirical findings. *Frontiers in public health* 9 (2021), 641673.
- [11] Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLAMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* (2023). <https://arxiv.org/abs/2306.02858>

Appendix

Table 5: Keywords and hashtags applied in the query with respect to its publication date.

Keywords/Hashtags	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
election2024	-	-	-	+	+
Election2024	-	-	-	+	+
US Elections	+	+	+	+	+
USElections	+	+	+	+	+
us elections	+	+	+	+	+
uselections	+	+	+	+	+
2024Elections	+	+	+	+	+
2024 Elections	+	+	+	+	+
2024 elections	+	+	+	+	+
2024elections	+	+	+	+	+
2024election	-	+	+	+	+
2024PresidentialElections	+	+	+	+	+
2024 Presidential Elections	+	+	+	+	+
2024presidentialelections	+	+	+	+	+
2024 presidential elections	+	+	+	+	+
saveamerica2024	-	-	+	+	+
presidentbiden	-	-	+	+	+
Biden	+	+	+	+	+
biden	+	+	+	+	+
bidenharris	-	+	+	+	+
JoeBiden	+	+	+	+	+
Joe Biden	+	+	+	+	+
joebiden	+	+	+	+	+
joe biden	+	+	+	+	+
joseph biden	+	+	+	+	+
Joseph Biden	+	+	+	+	+
Biden2024	+	+	+	+	+
biden2024	+	+	+	+	+
bidenharris2024	-	-	+	+	+
Donald Trump	+	+	+	+	+
donald trump	+	+	+	+	+
DonaldTrump	+	+	+	+	+
donaldtrump	+	+	+	+	+
donaldtrump2024	-	+	+	+	+
Trump2024	+	+	+	+	+
trump2024	+	+	+	+	+
trumpsupporters	+	+	+	+	+
trumprtrain	+	+	+	+	+
republicansofttiktok	+	+	+	+	+
conservative	+	+	+	+	+
MAGA	+	+	+	+	+
maga	+	+	+	-	-
makeamericagreatagain	-	+	+	+	+
ultramaga	-	+	+	+	+
KAG	+	+	+	+	+
Republican	+	+	+	+	+
trump2024	-	+	+	+	+
presidenttrump	-	-	+	+	+
trumpismypresident	-	-	+	+	+
letsgobrandon	-	+	+	+	+
GOP	+	+	+	+	+
CPAC	+	+	+	+	+
NikkiHaley	+	+	+	+	-

Continued on next page

Table 5 – continued from previous page

Keywords/Hashtags	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
nikkihaley	+	+	+	+	-
DeSantis	+	-	-	-	-
RonDeSantis	+	-	-	-	-
desantis	+	-	-	-	-
rondesantis	+	-	-	-	-
RNC	+	+	+	+	+
democratsoftiktok	+	+	+	+	+
democratsarehot	+	+	+	+	+
thedemocrats	+	+	+	+	+
voteblue2024❤️	-	-	+	+	+
voteblue2024	-	-	+	+	+
vote blue	-	-	+	+	+
DNC	+	+	+	+	+
dnc	+	+	+	+	+
kamalaharris	+	+	+	+	+
kamala harris	-	-	-	+	+
mariannewilliamson	+	+	-	-	-
deanphillips	+	+	+	-	-
williamson2024	+	+	+	-	-
phillips2024	+	+	+	-	-
democratic party	+	+	+	+	+
Democratic party	+	+	+	+	+
republican party	+	+	+	+	+
Republican party	+	+	+	+	+
Third party	+	+	+	+	+
third party	+	+	+	+	+
Green party	+	+	+	+	+
green party	+	+	+	+	+
Independent party	+	+	+	+	+
independent party	+	+	+	+	+
No Labels	+	+	+	+	-
RFKJr	+	+	+	+	+
RFK Jr.	+	+	+	+	+
RFK Jr	+	+	+	+	+
rfkjr	+	+	+	+	+
rfkj.r.	+	+	+	+	+
rfk	+	+	+	+	+
Robert F. Kennedy Jr.	+	+	+	+	+
robert f. kennedy Jr.	+	+	+	+	+
jill stein	+	+	+	+	+
jillstein	+	+	+	+	+
Jill Stein	+	+	+	+	+
JillStein	+	+	+	+	+
CornellWest	+	+	+	+	+
cornellwest	+	+	+	+	+