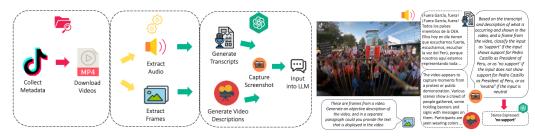
GET-Tok: A GenAl-Enriched Multimodal TikTok Dataset Documenting the 2022 Attempted Coup in Peru

GABRIELA PINTO, KEITH BURGHARDT, KRISTINA LERMAN, EMILIO FERRARA, University of Southern California, Los Angeles, CA, 90007, USA



- (a) GET-Tok: Our framework for Generative Al-Enriched TikTok
- (b) Example of classification given transcript, video description and frame.

Fig. 1. (a) Outline of the proposed GET-Tok framework and (b) its application to classify an actual TikTok video

TikTok is one of the largest and fastest-growing social media sites in the world. TikTok features, however, such as voice transcripts, are often missing and other important features, such as OCR or video descriptions, do not exist. We introduce the Generative AI Enriched TikTok (GET-Tok) data, a pipeline for collecting TikTok videos and enriched data by augmenting the TikTok Research API with generative AI models. As a case study, we collect videos about the attempted coup in Peru initiated by its former President, Pedro Castillo, and its accompanying protests. The data includes information on 43,697 videos published from November 20, 2022 to March 1, 2023 (102 days). Generative AI augments the collected data via transcripts of TikTok videos, text descriptions of what is shown in the videos, what text is displayed within the video, and the stances expressed in the video. Overall, this pipeline will contribute to a better understanding of online discussion in a multimodal setting with applications of Generative AI, especially outlining the utility of this pipeline in non-English-language social media. Our code used to produce the pipeline is in a public Github repository: https://github.com/gabbypinto/GET-Tok-Peru.

1 INTRODUCTION

Open data initiatives have enabled research on political turmoil and social unrest, predominantly using Twitter streams [10]. Research on TikTok has been understudied, but has included research into U.S. political communication [15, 16], the spread of COVID-19 information [4, 12, 21], the Brazilian Presidential Election [13], disinformation campaigns [3, 11], or climate change [5].

These studies, however, have highlighted limitations of TikTok data, questioning if the samples are representative, emphasizing the scarcity of non-English content during political crises in non-English speaking countries, or highlighting limitations in the data or the auto-generated speech transcripts that TikTok APIs produce [11, 12, 15, 17].

These shortcomings motivated us to design the proposed data collection pipeline, shown in Fig. 1, which includes using generative AI to augment the collected data, providing potential avenues for in-depth analysis of TikTok content on a larger scale. Possible research avenues that could be addressed from the presented work include analyzing content in a multilingual environment, multimodal analysis, potentially enabling answers to social science questions, such as the offline effects of social media, for countries that are traditionally understudied by our research community. We

1

Keywords	Hashtags
Castillo, Presidente Castillo, Pedro	PedroCastillo, Castillo, Boluarte, pedro-
Castillo, Dina Boluarte, Presidente	castillo, pedrocastilloperu, pedrocastil-
Boluarte, Boluarte	loperú, golpedeestado, GolpeDeEstado,
	golpedeestadoperu, crisispoliticaenperu,
	crisispoliticaenperú

Table 1. List of Keywords and Hashtags Applied in the Query

demonstrate a proof-of-concept of this pipeline resulting in the collection of a rich, multimodal dataset of Spanish-based content originating from Peru, documenting a period of social unrest.

Namely, in November 2022, Peru's Congress attempted to impeach President Pedro Castillo on the charges of leading a criminal organization to profit off government contracts. In the face of imminent impeachment, Castillo attempted to dissolve Congress on December 7, 2022, but Congress removed Castillo instead. Former Vice President of Peru Dina Boluarte was subsequently sworn in as the new President of Peru. This sparked wide-spread protests, with more than 2,370 demonstrations across the country, including 66 fatalities, according to ACLED [2]. The discussion of the political crisis quickly moved to TikTok, which has 20 million users in Peru [6], providing a case study for the interactions between online activity and offline events.

Using the pipeline described in this paper, we collected 43,697 videos, along with 2,345 Whisper-generated audio transcripts [18] and 2,345 GPT-4 generated descriptions of the the content of the videos [1]. We are sharing the dataset publicly with the research community. We are also publishing our data collection pipeline codebase on Github.²

2 DATA COLLECTION

We continuously query the TikTok Research API³ based on a set of keywords and hashtags, as shown in Table 1, and published between November 20, 2022, and March 1, 2023. The quantity of videos published during the period are shown in Fig. 3. These dates cover videos from 2 weeks before the political crisis until several months after protests reached their zenith. The collection of videos' features includes their unique identifier, the video's publication time and country code, the author's username, the description of the video, the unique identifier of the music (if any) in the video, the number of *likes, comments, shares, and views* the video received at the time of collection, the list of unique identifiers of the effects (if any) applied in the video, the list of associated hashtags, the unique identifier of its associated playlist (if any), and the audio transcripts (if any).

After the data collection process, the actual video files are downloaded with the Pyktok Python module⁴. To download each video, the post's username and ID are extracted and used to create a link with the format https://www.tiktok.com/@username/video/id.

Between querying the videos from the TikTok API (50,631) and downloading them via Pyktok, 1,427 are privatized and unavailable to download; thus we recommend reducing the time between feature extraction and video download to maximize the content gathered. In addition, Pyktok fails to download posts that contain a slideshow of photos, which represents 1.5% of the collected videos. Pyktok will also occasionally fail to fully download videos due to internet connectivity issues, which can be resolved by querying the url again after approximately 100 seconds. Futhermore, Pyktok will successfully download but in a format invalid for analysis. As of now, this error is unexplained and appeared in approximately 9.2% of the data collected. Our dataset is reduced to 43,697 after the removal of the privatized videos, and unsuccessful downloads. Overall, these errors had little impact on our data collection.

¹https://github.com/gabbypinto/GET-Tok-Peru-data

 $^{^2} https://github.com/gabbypinto/GET-Tok-Peru\\$

³https://developers.tiktok.com/doc/about-research-api/

⁴https://github.com/dfreelon/pyktok

GET-Tok: A GenAl-Enriched Multimodal TikTok Dataset Documenting the 2022 Attempted Coup in Peru

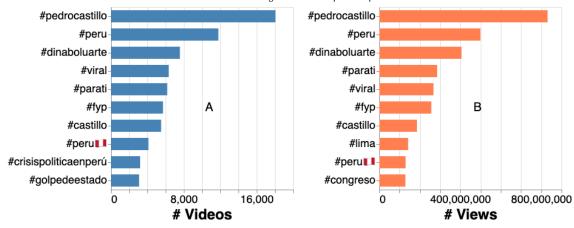


Fig. 2. Top 10 Most Frequent and Viewed Hashtags

Attribute	Sum
No. Users	17,181
No. Likes	63,539,322
No. Views	1,396,069,669
No. Comments	6,264,238
No. Shares	6.964.547

Table 2. Statistics on the Dataset

2.1 Detailed Statistics

Our dataset has 43,697 videos posted during critical political events during the observed period. Most notably, Pedro Castillo's attempted coup and removal from office were followed by Pro-Castillo protests and then protests against the disproportionate police response. The detailed description of the data collected, all of which have been published within Peru, is shown in Table 2 and Fig. 2.

A small proportion of the videos collected from the collected posts included TikTok-generated transcripts, which has the potential to contain rich-linguistical data. The percentage of videos with its transcripts included, with respect to the publication date, range between 1.4 to 15.3%. Due to the low proportion of transcripts provided by the TikTok API, we applied OpenAI's Whisper to generate the transcripts for a sample of 2,345 videos. The videos were chosen based on the videos' publication dates. The specific set of dates were chosen to provide a preliminary analysis of the comparison between the two types of transcripts surrounding critical dates within the period.

3 AI-POWERED DATA AUGMENTATION

3.1 Generating Whisper Transcripts

The transcripts are generated through a two-step process. Initially, the audio is extracted from the video content using FFmpeg, a widely recognized software tool for processing multimedia files.⁵ Following the extraction of audio, we utilize OpenAI's Whisper [18], to produce the transcripts.⁶ Specifically, we opt for the 'large-v3' model of Whisper due to its higher accuracy and semantic richness in transcription when compared to the 'base' model. To illustrate the effectiveness of our method, we compare transcripts generated by Whisper with those provided by TikTok in

⁵https://ffmpeg.org/

⁶https://github.com/openai/whisper

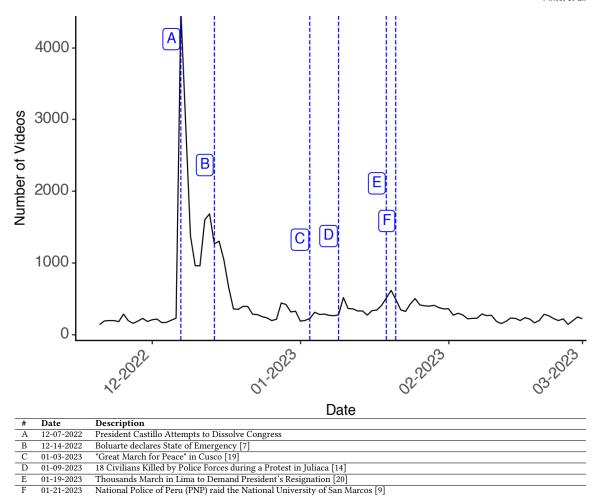


Fig. 3. Timeline of events and volume of TikTok posts.

Fig. 5, highlighting discrepancies in capitalization (yellow), punctuation (red), and spelling (blue). Notably, an instance of misspelling in the TikTok transcript was correctly spelled in the Whisper-generated transcript, showcasing the better performance of the latter. This analysis included posts from specific dates: November 20, 2022; November 23, 2022; December 14, 2022; January 18, 2023; and January 22, 2023. Our limited collection is primarily due to the audio extraction process and Whisper's limited throughput of approximately 19 videos per hour. The comparison of transcript proportions between TikTok and Whisper is depicted in Fig. 4.

We evaluate the similarity of 182 transcripts through several methods. The transcripts are compared based on their word count, resulting in a Pearson correlation r of 0.9816 and R^2 of 0.9635. Thus, indicating that the word count between the TikTok and Whisper transcripts were nearly equivalent. We apply the Jaccard similarity index based on the set of words presented in each transcript, resulting in a mean score of 0.76. To further assess the similarity between the transcripts, we employ BETO embeddings [8]—a BERT-based model tailored for Spanish language data. The average cosine similarity was 0.927 indicating a high degree of similarity between the two sets of transcripts.

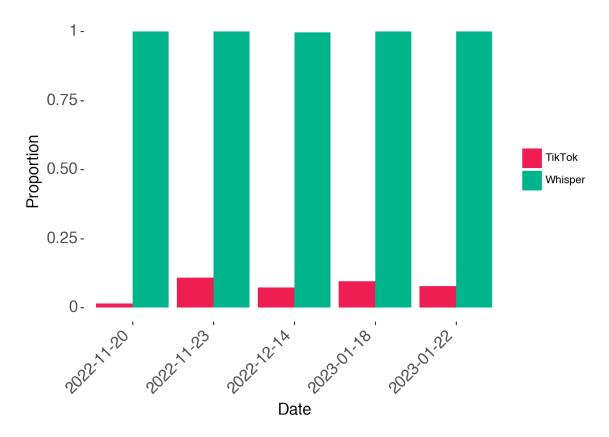


Fig. 4. Share of video transcripts extracted by TikTok and Whisper.

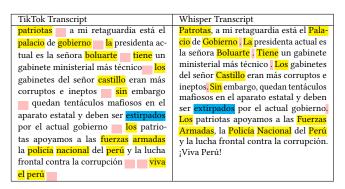


Fig. 5. Transcripts from TikTok and Whisper with yellow for capitalization, red for punctuation, blue for spelling discrepancies.

3.2 Generating Video Descriptions with GPT-4

For the purpose of conducting a comprehensive multimodal analysis, we use GPT-4 to generate descriptions of the activities depicted, as well as any text displayed, within these videos. We specifically use the gpt-4-vision-preview

model,⁷ one of the best models available at the time of writing this paper [1]. While OpenAI has not specified the exact dataset used for training their models, it is conceivable that the training data may include information pertinent to the political situation in Peru. This could explain any difference in performance for this pipeline applied to the present dataset compared to any other datasets analyzed in the future.

The description generation sub-pipeline is twofold. We first process the videos using the OpenCV Python package, ⁸ a tool that enabled us to capture 1 out of every 300 frames from the videos. These frames are then fed into GPT-4 along with the prompt: "These are frames from a video. Please generate an objective description of what is happening in the video and, in a separate paragraph, provide any text that appears in the video." We were able to generate descriptions for 2,345 videos.

3.3 GPT-4 Multimodal-Based Stance Detection

The described pipeline showcases a powerful application in classification tasks, particularly using GPT-4 to assess the stance conveyed by various inputs. This approach is applied to numerous posts, illustrated in Fig. 1a, aiming to categorize each post based on its sentiment towards Pedro Castillo's presidency—whether in support, opposition, or neutral.

The integration of audio and video frame extraction, alongside generating detailed descriptions of the video content, enables us to achieve significantly more accurate classifications than those based on the typically brief video captions. By employing the comprehensive methodology outlined in Fig. 1a, we can create detailed textual representations of the content. An additional layer of analysis is introduced by including detailed descriptions of the video content generated by GPT-4 as input.

The classification prompt is as follows: "Based on the transcript, description of what is occurring and shown in the video, and a frame from the video, classify the input as 'support' if the input shows support for Pedro Castillo as President of Peru, or as 'no support' if the input does not show support for Pedro Castillo as President of Peru, or as 'neutral' if input is neutral."

A few examples of the inputs, along with their corresponding "video_description," provided by the API, are shown in Fig. 6. The examples suggest good agreement with expectations; in future work, we will use human annotation to quantify this agreement.

4 CONCLUSIONS

In summary, this study presents a novel approach to collecting and analyzing social media content, focusing on TikTok videos from the political crisis in Peru. Utilizing generative AI, including the Whisper and GPT-4 models, we crafted a comprehensive pipeline that enhances the collected data with detailed audio, video, and textual analysis. The result is a valuable multimodal dataset of 43,697 videos, equipped with enhanced transcripts and video descriptions. This dataset bridges a significant research gap by focusing on non-English content and regions typically overlooked in social media studies. Our methodology showcases the potential of AI in extracting nuanced insights from social media narratives, contributing to a deeper understanding of the relationship between online discourse and real-world events. By sharing our dataset and codebase, we encourage further exploration into the impacts of digital platforms on societal and political dynamics, highlighting the critical role of AI in advancing social science research.

⁷https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

⁸https://pypi.org/project/opencv-python/



(A) Vamos Pedro
castillo el pueblo
está contigo #parati #viralvideo
#tiktok #peruanadas #love
#pedrocastillopresidente2021
Translation: Let's go Pedro Castillo the
people are with you Translation (screen):
Let's go Pedro Castillo



(B) #antaurohumala #huanta #pedrocastillo #cierredelcongreso #golpedeestado #insurgencia #andahuaylas #peru #peruanos Translation (screen): God protect our homeland



(C) castillo orejas de burro Translation: Castillo donkey ears. Translation (screen): Castillo Castillo doesn't know how to govern Donkey ears doesn't know how to listen.

Fig. 6. Videos and their captions separated by the detected stance: (A) 'Pro-Castillo' (B) 'Neutral' (C) 'Anti-Castillo'

Limitations. The dataset sheds light on Peru's political crisis discussions online but has limitations. It doesn't fully capture the sentiment on former President Castillo's attempted coup, considering Peru's 68.68% social media penetration and approximately 20 million TikTok users out of a 34.5 million population. This study's methodology introduces GPT-4 models for data augmentation, which, despite not being open-source or free, offers a new approach. However, Whisper encounters issues with background noise, omits emojis from the source videos and AI-generated video descriptions present ethical dilemmas due to potential biases.

Ethical Considerations. In compliance with the TikTok Research API Terms of Service, we regularly update and prune our dataset, ensuring its alignment with current API offerings and adherence to TikTok's guidelines. Shifting focus, our study employs GPT-4 and Whisper to generate textual and visual content analyses. OpenAI's nondisclosure of the models' training data raises ethical concerns about potential biases. Acknowledging this, we aim to prioritize bias detection and mitigation in our future work, enhancing the research's ethical foundation.

Acknoledgements. This work was supported in part by DARPA (contract no. HR001121C0169).

REFERENCES

- [1] J. Achiam et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.
- [2] ACLED. ACLED Data, 2023. URL https://acleddata.com/. Accessed on: 12-01-2023.
- [3] N. Alonso López, P. Sidorenko Bautista, F. Giacomelli, et al. Beyond challenges and viral dance moves: Tiktok as a vehicle for disinformation and fact-checking in spain, portugal, brazil, and the usa. *Anàlisi*, 2021.
- [4] C. H. Basch, Z. Meleo-Erwin, J. Fera, C. Jaime, and C. E. Basch. A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on tiktok. *Human vaccines & immunotherapeutics*, 17(8):2373–2377, 2021.
- [5] C. H. Basch, B. Yalamanchili, and J. Fera. # climate change on tiktok: a content analysis of videos. Journal of community health, pages 1–5, 2022.
- \cite{Matter} T. Bianchi. Social media usage in peru statistics & facts, 2024.
- [7] R. Cano. Peru's new government declares police state amid protests, 2022. URL https://apnews.com/article/politics-latin-america-peru-pedrocastillo-dda20b05a4269d40d89b080fd1aab7e6.
- [8] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. Spanish pre-trained bert model and evaluation data. In PML4DC at ICLR 2020, 2020.
- [9] D. Collyns. Police violently raid lima university and shut machu picchu amid peru unrest, 2023. URL https://www.theguardian.com/world/2023/jan/22/peru-police-make-violent-raid-on-limas-san-marcos-university.

- [10] C. A. Davis, G. L. Ciampaglia, L. M. Aiello, K. Chung, M. D. Conover, E. Ferrara, A. Flammini, G. C. Fox, X. Gao, B. Gonçalves, et al. Osome: the iuni observatory on social media. *PeerJ Computer Science*, 2:e87, 2016.
- [11] A. Espinoza and C. A. Piña-García. Propaganda and manipulation in mexico: A programmed, coordinated and manipulative "pink" campaign. Journalism and Media, 4(2):578–598, 2023.
- [12] Y. Li, M. Guan, P. Hammond, and L. E. Berrey. Communicating covid-19 information on tiktok: a content analysis of tiktok videos from official accounts featured in the covid-19 information hub. Health education research, 36(3):261–271, 2021.
- [13] J. Lima, M. Santana, A. Correa, and K. Brito. The use and impact of tiktok in the 2022 brazilian presidential election. In *Proceedings of the 24th Annual International Conference on Digital Government Research*, pages 144–152, 2023.
- [14] B. McDonald and A. Tieffenthäler. How peru used lethal force to crack down on anti-government protests, 3 2023. URL https://www.nytimes.com/ 2023/03/16/world/americas/peru-protests-police.html.
- [15] J. C. Medina Serrano, O. Papakyriakopoulos, and S. Hegelich. Dancing to the partisan beat: A first analysis of political communication on tiktok. In Proceedings of the 12th ACM Conference on Web Science, pages 257–266, 2020.
- [16] A. Moir. The use of tiktok for political campaigning in canada: The case of jagmeet singh. Social Media+ Society, 9(1):20563051231157604, 2023.
- [17] V. Purushothaman, T. McMann, M. Nali, Z. Li, R. Cuomo, and T. K. Mackey. Content analysis of nicotine poisoning (nic sick) videos on tiktok: retrospective observational infodemiology study. *Journal of medical internet research*, 2022.
- [18] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning, pages 28492–28518. PMLR, 2023.
- [19] Reuters. Conservative peruvian groups march for peace ahead of fresh protests, 2023. URL https://www.usnews.com/news/world/articles/2023-01-03/conservative-peruvian-groups-march-for-peace-ahead-of-fresh-protests.
- [20] S. Schmidt. Protesters flood peru's capital to demand president step down, 2023. URL https://www.washingtonpost.com/world/2023/01/19/lima-peru-fire-protests/.
- [21] L. Shang, Z. Kou, Y. Zhang, and D. Wang. A multimodal misinformation detector for covid-19 short videos on tiktok. In 2021 IEEE international conference on big data (big data), pages 899–908. IEEE, 2021.