Exploring Brazilian TikTok and YouTube Shorts: A Public Dataset for Video Characterization

Tomas Lacerda[1], **Marcelo Sartori Locatelli**[1], **Igor Costa**[1], **Lorenzo Carneiro**[1], **Virgílio Almeida**[1], **Wagner Meira Jr.**[1],

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
tomaslm00@ufmg.br, locatellimarcelo@dcc.ufmg.br, igor.joaquim@dcc.ufmg.br,
lorenzocarneiro@dcc.ufmg.br, virgilio@dcc.ufmg.br, meira@dcc.ufmg.br

Abstract. *Short video platforms have garnered significant attention in recent years, with discussions ranging from concerns about inappropriate content and addiction to strategies for maximizing user engagement and screen time. Despite the large user base and growing relevance of these platforms, there is still a notable lack of comprehensive datasets focused on broad recommendation and moderation. This is especially true for TikTok, where API access is limited and collecting unbiased data is challenging. In this collection, we present a diverse and rich dataset from YouTube Shorts and TikTok's main feeds in Brazil, comprising over 35,000 videos. The dataset includes detailed engagement statistics, extensive video metadata, over 170,000 keyframes for visual analysis, and Safe Search API assessments for each keyframe. This rich resource fills a critical data gap, offering valuable tools for research on content categorization, user behavior analysis, and platform engagement strategies.*

1. Introduction

With the rapid rise in popularity of short-form media platforms in recent years, the world has witnessed a profound shift in how people consume and create content [Vázquez-Herrero et al. 2022][Anderson 2020]. This shift has brought about a range of both positive and negative consequences. On the positive side, these platforms have democratized access to marketing campaigns and offered new opportunities for small content creators, as well as a powerful means to popularize diverse forms of content [Valdovinos Kaye 2020][Guarda et al. 2021]. However, the negative impacts such as addiction, self-esteem challenges [Peng et al. 2022], harmful content dissemination, and the perpetuation of discrimination [Weimann and Masri 2023] cannot be overlooked.

As these platforms continue to evolve, the growing need to understand them more deeply has led us raise a dataset that can be used to explore the types of content that are recommended, and the demographics of their users. To facilitate this, we compile the *short_videos_dataset*, a dataset that contains more than 35,000 videos - 15,193 from TikTok and 10,925 from YouTube Shorts - collected during November 2024 to January of 2025. The data was gathered using a cold-start recommendation approach on both platforms, ensuring an unbiased sample of content trends.

Additionally we collected over 170,000 keyframes, which are individual frames that represent significant visual content within a video, 5 from each video filtering by the most significant frames, which may be a valuable asset for computer vision research and psychological studies on trending content, while also enabling the analysis of visual elements that contribute to virality and engagement. We further enrich the video data by processing each frame through Google Vision API's Safe Search analysis, serving as a baseline for detecting unsafe content within short video platforms.

2. Related Work

Even though short media platforms have been under scrutiny, researchers continue to struggle with collecting unbiased cold-start recommendation data, often resorting to content-specific approaches. For example, [Steel et al. 2023] developed datasets specifically focused on political content, targeting particular political events to study recommendation dynamics. In a similar vein, [Pinto et al. 2024] also constructed politically-oriented datasets, emphasizing event-centric data collection to analyze engagement patterns.

A more generic approach was proposed by [Shutsko 2020], where the authors characterized user engagement broadly across various content types. However, their dataset was limited by its smaller scale (due to is qualitative approach) and lack of public availability, constraining reproducibility and broader research applicability.

Our dataset complements these previous efforts by adopting a broader cold-start recommendation perspective to create a public resource with additional data, such as keyframes and SafeSearch filters, which have a plethora of potential uses.

3. Data Collection

The *short_videos_dataset* was created to assess the presence of potentially harmful content recommended by short media platforms. We aimed to capture unbiased cold-start recommendations by utilizing methods tailored to each platform: TikTok and YouTube Shorts. Data was consistently collected from Belo Horizonte, Brazil, ensuring an accurate reflection of the local recommendation experience.

3.1. TikTok

TikTok videos were collected using TikTokApi[1], the Unofficial TikTok Wrapper, by making requests to an application route that recommends 20 random videos from the For You page. The videos were stored in a JSONL file, maintaining a cold-start recommendation scenario by preventing TikTok from receiving feedback on previously recommended videos. The Origin Token used was from Belo Horizonte, Brazil to guarantee the consistency on the information gathered. To download the video for keyframe exctraction we used python's Pyktok[2], using *video_id* and *user_unique_id* to find the correct video.

3.2. Youtube Shorts

Collecting data from YouTube Shorts posed a challenge since there was no app-based recommendation route available through an API. We implemented a web scraper that systematically extracted video URLs from the recommended section of YouTube Shorts, simulating user scrolling behavior. Consistency was maintained by using headers set to Belo Horizonte, Brazil, and controlling the scrolling duration for each video, minimizing the influence of external factors on the recommendation algorithm. After collecting

[1] https://github.com/davidteather/TikTok-Api

[2] https://github.com/dfreelon/pyktok

URLs, we utilized the Official YouTube API[3] to retrieve detailed video metadata. We used yt_dlp[4] library to download the video for keyframe exctration.

3.3. Post-processing

After downloaded, each video went through a keyframe detection algorithm[5] that identified the five most visually distinct frames within a 5,000-frame window, ensuring maximum diversity of visual information using a peak estimate difference algorithm. Subsequently, each identified keyframe was analyzed using the Google Vision API's SafeSearch [6] feature. API responses were recorded under the columns `keyframe1` to `keyframe5`, providing content classification across five categories: *Medical*, *Adult*, *Racy*, *Spoof*, and *Violence*. Each category is rated on a scale from 0 (`UNLIKELY`) to 4 (`VERY LIKELY`), enabling a systematic assessment of potentially NSFW content across both platforms.

3.4. Analysis

The final dataset is comprised of 15,193 TikTok videos and 10,925 YouTube Shorts videos, complete with metadata, keyframe analysis results, and their respective extracted keyframes. This comprehensive resource enables robust analysis of content trends, visual computation techniques, and platform-specific recommendation behaviors.

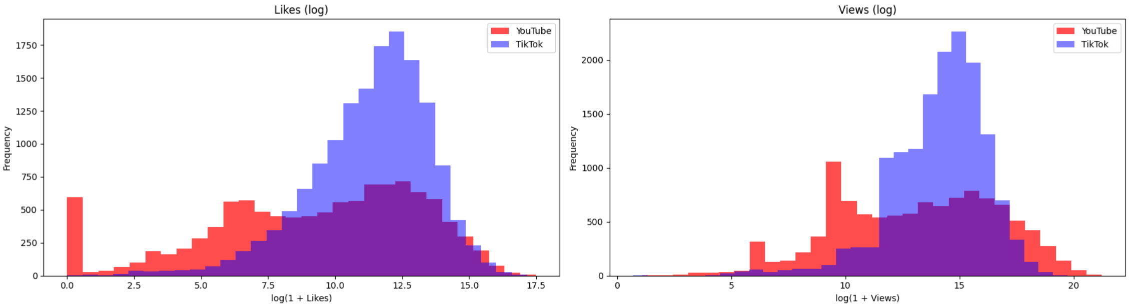

Figure 1. Distribution of likes and views (log) on both platforms

In Figure 1, The left histogram shows the distribution of *likes* and the right shows *views*, both using a $\log(1 + x)$ transformation to reduce skewness. TikTok videos (blue) tend to receive more likes and views, while YouTube content (red) shows a wider and more varied distribution, including a significant portion with low engagement. The overlapping regions highlight shared engagement ranges. We can also notice that YouTube has a much higher 0-like video recommendations, which may indicate that TikTok does not recommend videos that have yet to be liked, making it less beginner friendly.

Table 1, for example, reveals that YouTube Shorts has significantly higher average view counts than TikTok, while TikTok exhibits a slightly higher and more consistent like-to-view ratio. This can suggest that although YouTube reaches a broader audience, TikTok users may engage more actively with the content they watch, which may indicate a better and more personalized algorithm from TikTok's side.

[3] https://developers.google.com/youtube/v3

[4] https://github.com/yt-dlp/yt-dlp

[5] https://github.com/joelibaceta/video-keyframe-detector

[6] https://cloud.google.com/vision/docs/detecting-safe-search

Table 1. Engagement metrics across platforms, with a 95% confidence interval.

Platform	Views Avg (±CI)	Likes Avg (±CI)	Likes/Views Avg (±CI)
YouTube	19.44M ± 1.29M*	449.5K ± 27.6K	0.0749 ± 0.0339
TikTok	5.55M ± 0.21M*	411.7K ± 16.3K	0.0827 ± 0.0011

Table 2. Keyframe metrics across platforms, with a 95% confidence interval.

Platform	Adult	Medical	Spoofed	Violence	Racy
YouTube	0.347 ± 0.005	0.727 ± 0.005	1.436 ± 0.012	0.711 ± 0.004	1.116 ± 0.007
TikTok	0.443 ± 0.005	0.614 ± 0.004	1.154 ± 0.009	0.631 ± 0.004	1.145 ± 0.006

Table 2 highlights subtle but relevant platform differences: TikTok presents higher average scores for Adult and Racy content, suggesting a more permissive visual landscape. Conversely, YouTube shows a higher Spoofed score, possibly indicating a greater prevalence of AI-generated or manipulated media on the platform.

4. Applications

The *short_videos_dataset* can be used to gain deeper insights into how short-video platforms recommend content, exploring the factors that contribute to virality and optimizing short-video performance.

Beyond recommendation analysis, this dataset holds significant value for visual computation research. It can be leveraged to cluster keyframes, categorize videos based on visual similarities, and study content distribution trends. Additionally, it serves as a valuable resource for harmful content detection and moderation, allowing researchers to combine Google Vision Safe Search data with individual keyframe analysis to assess the prevalence of inappropriate or harmful material across platforms.

Tables 3 and 4 provide a detailed overview of the dataset's contents. Additionally, the folders containing the extracted keyframes are named using generated identifiers that do not correspond to the original video IDs. This was done for ethical purposes, in order to prevent the re-identification of users.

Table 3. YouTube dataset description

Column	**Description**
folder_id	folder_id in which keyframes are contained
description	Video description
category_id	Video category id
defaultLanguage	Default Video Language
duration	Video duration
caption	If subtitles are available
view_count	Number of views
like_count	Number of likes
dislike_count	Number of dislikes
comment_count	Number of commentaries
keyframe1–5	Google Vision Safe Search Analysis
day	Collected at (DD-MM-YYYY)

Table 4. TikTok dataset description

Column	**Description**
folder_id	folder_id in which keyframes are contained
collectCount	The ammount of times a user has saved the video
commentCount	Number of commentaries
diggCount	Number of likes
playCount	Number of views
shareCount	Number of shares
author_followerCount	Author follower count
author_heartCount	Total of likes on the authors profile account
author_videoCount	Total of published videos
desc	Video description
keyframe1–5	Google Vision Safe Search Analysis
day	Collected at (DD-MM-YYYY)

5. Ethical Considerations

While this dataset enables valuable insights into Brazilian short video social media, it also involves the analysis of personal data, which TikTok users may not have fully consented to. Although all data in this dataset was made publicly available by its authors, it is possible that users did not anticipate such analytical uses when publishing their content. In light of this, we have dehydrated the dataset to ensure that no individual can be easily identified. While elements like keyframes are hard to fully dehydrate, their use should remain within fair use, aimed at deepening understanding of Brazil's short video phenomenon, which might otherwise remain unexplored. To ensure ethical use, keyframes won't be publicly available but can be requested via a dedicated link in the dataset description.

6. Usage Restriction

This dataset is provided strictly for academic research purposes. Any use for commercial, promotional, or non-academic objectives is prohibited. The dataset should be handled responsibly and ethically, ensuring that all analyses and publications derived from it respect the privacy and dignity of the individuals whose content is included.

7. Final Considerations

Research on the impacts of social media platforms continues to increase, reaffirming the importance of finding reliable and unbiased data. In this context, the present work strives to contribute to a better understanding of the impacts of these types of platforms within the Brazilian context.

Some of the dataset's main limitations stem from ethical considerations, particularly because TikTok's API is officially available only in the United States. Given this constraint, other studies have also resorted to similar methods[Steel et al. 2023], reinforcing the need to remove user-based information to preserve individual privacy.

The dataset can be accessed here[7].

8. Acknowledgments

We are especially grateful to Lucas Gervásio, Raul Ferreira, Gabriela Lacerda, and Gabriel Pains for generously sharing their computational resources, which were essential for processing video data. This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

Anderson, K. E. (2020). Getting acquainted with social networks and apps: it is time to talk about tiktok. *Library hi tech news*, 37(4):7–12.

Guarda, T., Augusto, M. F., Victor, J. A., Mazón, L. M., Lopes, I., and Oliveira, P. (2021). The impact of tiktok on digital marketing. In *Marketing and Smart Technologies: Proceedings of ICMarkTech 2020*, pages 35–44. Springer.

Peng, C., Lee, J.-Y., and Liu, S. (2022). Psychological phenomenon analysis of short video users' anxiety, addiction and subjective well-being. *International Journal of Contents*, 18(1):27–39.

Pinto, G., Burghardt, K., Lerman, K., and Ferrara, E. (2024). Get-tok: A genai-enriched multimodal tiktok dataset documenting the 2022 attempted coup in peru. *arXiv preprint arXiv:2402.05882*.

Shutsko, A. (2020). User-generated short video content in social media. a case study of tiktok. In *Social Computing and Social Media. Participation, User Experience, Consumer Experience, and Applications of Social Computing: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*, pages 108–125. Springer.

[7] https://zenodo.org/records/15446544

Steel, B., Parker, S., and Ruths, D. (2023). The invasion of ukraine viewed through tiktok: A dataset. *arXiv preprint arXiv:2301.08305*.

Valdovinos Kaye, D. B. (2020). Make this go viral: Building musical careers through accidental virality on tiktok. *Flow*, 27(1).

Vázquez-Herrero, J., Negreira-Rey, M.-C., and López-García, X. (2022). Let's dance the news! how the news media are adapting to the logic of tiktok. *Journalism*, 23(8):1717–1735.

Weimann, G. and Masri, N. (2023). Research note: Spreading hate on tiktok. *Studies in conflict & terrorism*, 46(5):752–765.