What we can learn from TikTok through its Research API

Francesco Corso
Politecnico di Milano & CENTAI
Milan, Italy
francesco.corso@polimi.it

Francesco Pierri
Politecnico di Milano
Milan, Italy
francesco.pierri@polimi.it

Gianmarco De Francisci Morales
CENTAI
Turin, Italy
gdfm@acm.org

ABSTRACT

TikTok is a social media platform that has gained immense popularity over the last few years, particularly among younger demographics, due to the viral trends and challenges shared worldwide. The recent release of a free Research API opens the door to collecting data on posted videos, associated comments, and user activities. Our study focuses on evaluating the reliability of the results returned by the Research API, by collecting and analyzing a random sample of TikTok videos posted in a span of 6 years. Our preliminary results are instrumental for future research that aims to study the platform, highlighting caveats on the geographical distribution of videos and on the global prevalence of viral and conspiratorial hashtags.

CCS CONCEPTS

• **Information systems** → *Information retrieval*; *Presentation of retrieval results*; **Social networks**.

KEYWORDS

online social networks, API, TikTok, conspiracy theories

ACM Reference Format:
Francesco Corso, Francesco Pierri, and Gianmarco De Francisci Morales. 2024. What we can learn from TikTok through its Research API. In *Proceedings of Diffusion of Harmful Content on Online Web Workshop (DHOW - WebSci '24).* ACM, New York, NY, USA, 5 pages. https://doi.org/3630744.3663611.

1 INTRODUCTION

TikTok is a social media platform for sharing short-form videos, known for its wide range of user-generated content, including lip-syncing, comedy, talent displays, and more. It has seen a steep increase in popularity, becoming one of the most prominent social media platforms on the Internet, with over 1 billion monthly active users and millions of videos posted every day around the world.[1]

TikTok has recently released a public Research API,[2] to which researchers can apply for access to gather data on videos, users, and comments via three respective primary endpoints. Such data availability initiative follows the examples of other platforms, such as Facebook[3] and Twitter,[4] which have been compelled to open to their data to researchers interested in studying the integrity of digital environments [2], especially under the pressure from the EU Digital Services Act.[5]

In the current post-API era [5], where most of the once-free APIs have been closed or converted to paid services, TikTok is still a relatively new frontier, even though the platform has been online for over five years. So far, the main approach to obtain data from TikTok has been to scrape and collect videos manually, as done, for instance, by Guinaudeau et al. [6] who studied political videos in the US, and showed differences in the activity of users of both TikTok and YouTube. Other works, such as the one by Medina Serrano et al. [11], also focused on US political discussions on TikTok, but employed a more content-based approach, and used wide-scale ML models on the video or the audio track of the videos they scraped. Similarly, other research described the usage of TikTok by organizations to communicate safety measures, best practices, and news during the pandemic [8, 13], and the impact of soft moderation labels employed by the platform for videos related to the COVID-19 pandemic [9]. Pera and Aiello [14] compared TikTok and YouTube, this time by using the TikTok Research API to look for climate change-related videos. Klug et al. [7] used a mixed-method approach to investigate the common assumptions of users about the TikTok recommendation algorithm. Lastly, a work similar to ours is that by McGrady et al. [10], who focused on gathering a random sample from social media (YouTube) and estimated the total number of videos present on the platform.

In this paper, we collect and analyze a random sample of TikTok videos posted in the 6 years spanning from January 2018 to December 2023 by means of TikTok Research API. We ask the following research question.

RQ1: What view of TikTok do we get through the lens of its Research API?

We provide a series of quantitative analyses on the data returned by the Research API and explore the potential implications for research relying on such a tool. By using repeated calls to the API, we build a worldwide random sample of over 500k videos (stratified per month) and analyze engagement metrics such as likes, shares, comments, and views. We highlight the temporal growth of the platform and show that the user base is dominated by Asian countries, with the USA as the only Western country in the top 10 in terms of shared videos. Lastly, we underline the effects of viral hashtags on driving engagement around videos that use the specific "For you" functionality of the platform, and offer an outlook on the prevalence of hashtags related to conspiracy theories.

[1] https://www.businessofapps.com/data/tik-tok-statistics/ accessed on 26/01/2024
[2] https://developers.tiktok.com/products/research-api accessed on 24/01/2024.

DHOW - WebSci '24, May 21–24, 2024, Stuttgart, DE

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/3630744.3663611.

[3] https://www.crowdtangle.com
[4] https://developer.twitter.com/en/use-cases/do-research
[5] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package accessed on 26/01/2024.

2 DATA COLLECTION

Our main constraint and objective for this research is to use exclusively the official TikTok Research API, which has severe restrictions on usage and availability. In particular, each API research organization has a quota of 1000 available requests per day. Since each request can return up to a maximum of 100 items, the resulting theoretical limit of the maximum number of elements available each day is 100 000. Given our aim of studying the geographical distribution of videos, we use a query that contains all the region codes described in the TikTok API documentation (note that Canada is not available by default in the API).[6] Additionally, we do not use any keywords for this research, in order to obtain a sample that is not conditioned by a specific topic. We use monthly queries since the maximum width of the time frame allowed for data collection by the API is 30 days. To meet the constraints imposed by the Terms of Service[7] we fix our collection quota to 1000 videos per month. Our sample thus aims to be stratified and have a uniform number of randomly sampled videos for each month in the period of our study. We send 10 requests, each of 100 videos, to the `/video/query` endpoint for each month from January 2018 to December 2023 (72 months). This yields a theoretical 72k items per day of extraction by using 720 of the daily 1000 queries available.

This collection process was run for 15 consecutive days from January 17th 2024 to January 31st 2024. If the maximum quota were reached each day, it would result in a dataset of over 1 million items, with 15k videos per month in the time frame of the study. Our methodology for data collection adheres to ethical standards as we do not try to deanonymize users. In addition, TikTok users have explicitly consented to the Terms of Service, which include the acknowledgment and approval of the transfer of personal data through the API.[8]

Finally, we make available the code we designed for the usage of the Research API. [9]

3 RESULTS

3.1 Evaluation of the API

Figure 1 shows the theoretical and real number of videos obtained from the data collection process described in Section 2. The API failed to meet the required quotas, delivering at most 65% (out of 72k) of the requested videos, a number which is in line with the data persistence of other social networks [4]. At the end of the data collection process, we obtained a total of 577 517 videos instead of the estimated 1+ million. The number of distinct users in the collection is almost the same as the number of videos, with only 0.34% repeated users in the sample. This result is expected given the extremely large number of users on the platform (over 1 billion).

The API documentation[6] explains that the requested quotas might not be met when videos marked as private or deleted appear in the response. These videos are not returned but are still counted by their internal system as part of the query result. This fraction of unavailable videos can thus give us an indication of the proportion

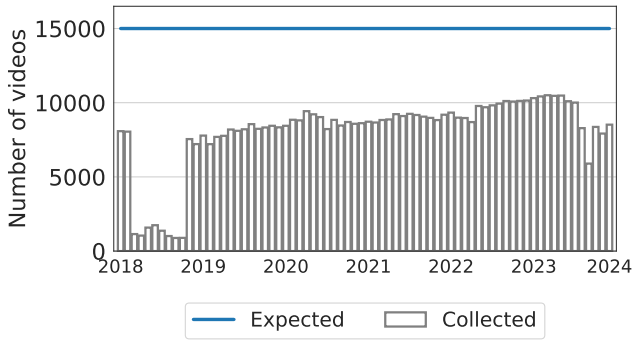

Figure 1: Time series of the data collection. The blue line represents the theoretical quota (maximum number of videos obtainable with the given number of API calls), while the histogram shows the obtained quota per month.

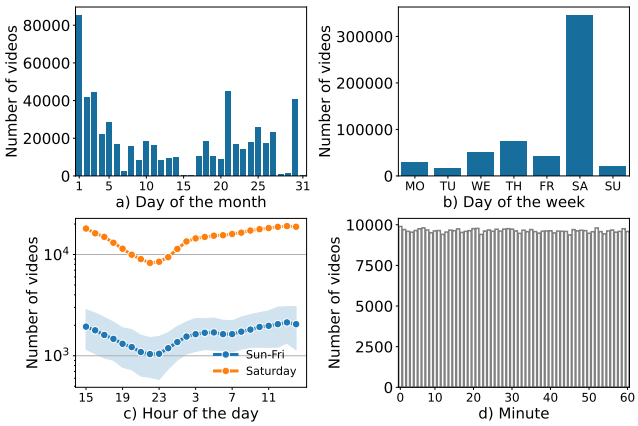

Figure 2: Number of videos posted (a) for each day of the month, (b) for each day of the week, (c) for each hour of the day (UTC), and (d) for each minute of the hour. The API shows a bias at the daily level, but not at the minute level.

of unavailable content on TikTok, month by month. There is a growing trend in the number of returned videos, which finds its peak in March 2023, probably because older videos are more likely to be deleted by users. There is also a large gap in the period from March 2018 to December 2018 where the returned data points barely surpassed 500 items per month. A similar but smaller drop appears around July 2023. We do not know the causes behind the missing videos in 2018, but we can speculate that it is possibly due to an error in the internal systems of the API. For this reason, we exclude the year 2018 from the following analyses.

3.2 Temporal Bias

Here, we investigate temporal patterns in the videos present in our sample. First, we analyze the frequency of appearance of all the different days of the month. In Figure 2(a), we show the cumulative number of videos posted for each day of the month, indicating that our data is not a perfect random sample, since the distribution is

[6] https://developers.tiktok.com/doc/research-api-specs-query-videos
[7] https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en accessed on 24/01/2024
[8] https://www.tiktok.com/legal/page/eea/privacy-policy/en accessed on 24/01
[9] https://github.com/orsoFra/python_tiktok-research-api

not uniform across all days of the month. For instance, we observe an unusual amount of videos posted on the first day of the month along with some missing days (e.g., the 15th). Figure 2(b) shows the number of videos posted for each day of the week. Saturday is the day when the majority of the videos are posted (>55%) followed by Thursday and Wednesday. Both of these observed facts are probably due to a malfunctioning of the Research API internal mechanisms, as we do not have reasonable evidence showing that these two phenomena are generated by user behaviour on the platform. Figure 2(c) shows the number of videos posted for every hour (UTC zone). We plot two time series to show the difference in volumes for Saturday compared to the rest of the days of the week. We find similar behavior to what was shown on other social media platforms like Twitter [15], where there is a peak of posted videos in the early afternoon. Despite the fact that videos posted on Saturday are three times larger than on other days of the week, the distribution pattern of the posting time remains very similar. Finally, Figure 2(d) shows the distribution of the minutes of the posting time. We find a different behavior compared to what has been evidenced by Pfeffer et al. [15] on Twitter (now X), where 15% of the data they collected was generated in the first minute of the hour in which they were posted. The distribution on TikTok is instead an almost uniform across all minutes. Pfeffer et al. suggested their result was due to bots' activity and programmed tweets. TikTok also allows scheduling video releases in advance, but only to Creators and Business accounts,[10] which are a small minority of the user-base of the platform. Thus this functionality seemingly does not influence the pattern of the posting times.

3.3 Distribution of interactions

Let us now focus on the interaction indicators available in our sample. Figure 3 shows the complementary cumulative distribution function (CCDF) of the four features available on the API: the number of views, likes, shares, and comments for each video. The plot shows a scaling behavior typical of social networks [1]. There has been a progressive increase in the median values for views and likes over the years, as these indicators follow the platform's growth. The order of the subplots is increasing in 'strength' of interaction [11], with the lowest defined as the visualization of the video, the second with a like, the third with a share, and the fourth with a comment. Indeed, the latter two have maximum values of two orders of magnitude lower than views and likes.

3.4 Region prevalence

Figure 4 shows the top 10 countries by number of videos in our sample. The first is India, with over 12% of videos, followed by Indonesia and then the US, which is also the only Western country in the top-10 list. We further investigate this aspect by plotting the yearly prevalence of the top 10 countries over the span of the dataset Figure 5.

The most evident feature is the prevalence of videos from India and Southeast Asia in general. From 2019 to mid-2020, India was the most prominent country in our dataset, with over 40% of the total videos sampled from early 2019. The rapidly descending trend

[10]https://www.tiktok.com/business/en-US/blog/introducing-video-scheduler-now-you-can-plan-tiktoks-in-advance accessed on 20/02/2024

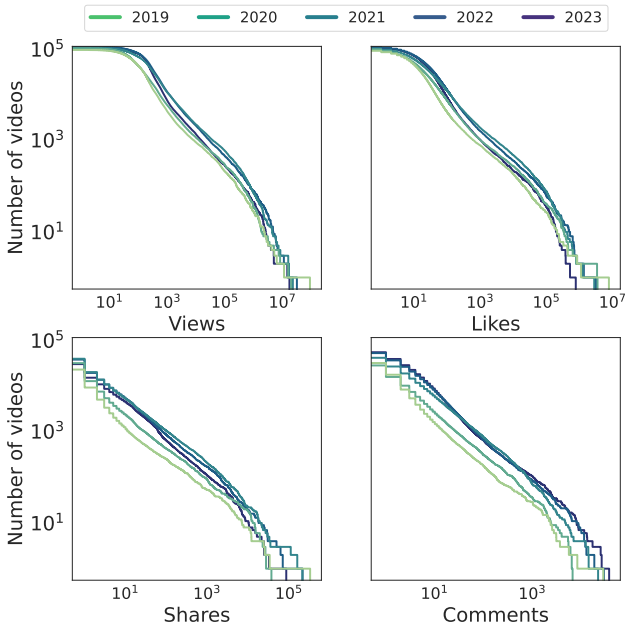

Figure 3: CCDFs of the four main interactions on TikTok: number of views, of likes, of shares, and of comments for videos per year. All the features have a heavy-tailed distribution. The yearly platform growth is evident in the shift to the right of each feature. Axes are on a logarithmic scale.

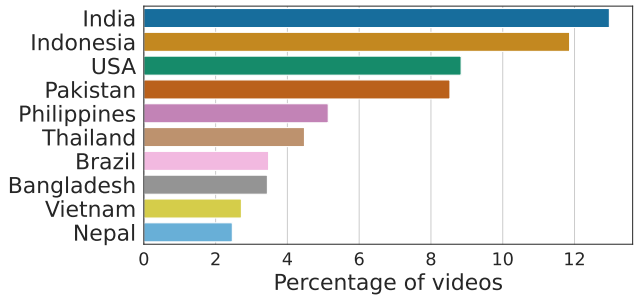

Figure 4: Top 10 regions by prevalence in the dataset with relative percentage of prevalence in the sample. India is still the largest one historically, despite the ban in 2020.

is due to a nationwide censorship policy applied in June 2020 which affected TikTok and other Chinese applications [16]. Note how the top 10 countries represent just over 60% of the total videos sampled from the API, thus indicating again a heavy-tailed distribution.

3.5 Effects of viral hashtags

Since their creation on Twitter in 2007, hashtags have morphed into fundamental and pervasive elements of social media culture [3]. Originally designed for categorization and conversation facilitation, hashtags now play a fundamental role in content discovery and trendsetting across various platforms. TikTok is no different in this aspect. Users make use of hashtags to define and aid the discovery of the content they post. Some of these hashtags are especially

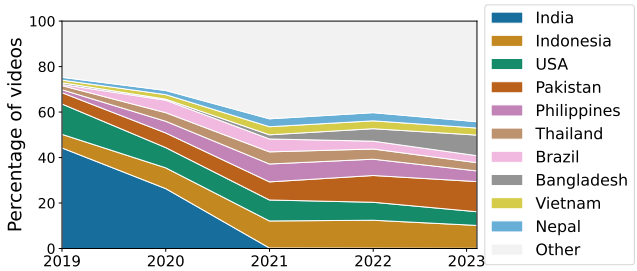

Figure 5: Yearly prevalence of the top 10 regions in our sample. The light-grey area represents all the other regions collected. Most countries in the top 10 are in Asia.

Table 1: Top 10 hashtags by frequency of use in our sample and their virality (manual assessment).

Hashtag	Viral?	Hashtag	Viral?
fyp	Yes	fypシ	Yes
foryou	Yes	viral	Yes
duet	No	tiktok	No
capcut	No	parati	Yes
foryoupage	Yes	trending	Yes

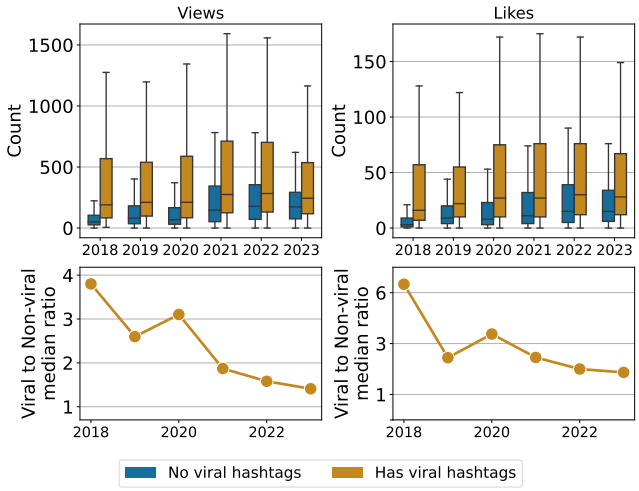

Figure 6: (Top) Yearly distributions of views and likes of videos according to whether they use 'viral' hashtags. (Bottom) Yearly ratio of the medians of views and likes for videos that use 'viral' hashtags vs. those that do not.

employed because they allegedly boost the visibility of the content, by exposing it to the TikTok recommendation algorithms.

Table 1 shows the top 10 most frequent hashtags used in our sample, with a manual classification of the intent of virality of the hashtag. This classification is based on the perceived purpose of the hashtags to recall a particular functionality of the social platform: the 'For You Page', where the average user of TikTok spends over

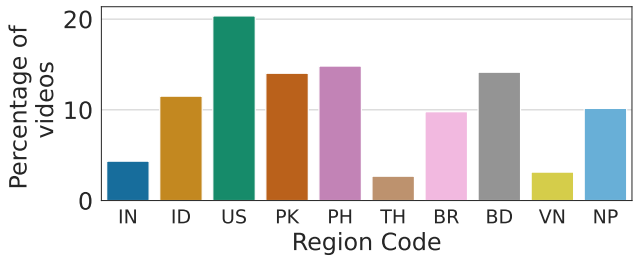

Figure 7: Percentage of videos that use viral hashtags in the top 10 countries by prevalence.

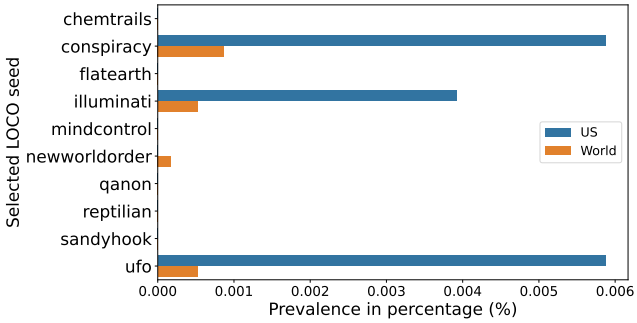

Figure 8: Percentage of videos in the world and in the US that use conspiracy-related hashtags.

60% of their time [17]. The use of this hashtag indicates the will of the author of the video to 'invite' the algorithm to show their content on the 'For You Page', thus potentially widening their audience. Figure 6 (top) tests this effect by comparing the distribution of views and likes for videos that use at least one 'viral' hashtag in their description (approximately 15% of the total) to the rest of the videos which do not use these hashtags. Videos that use 'viral' hashtags have significantly more views and likes compared to the ones that do not use them (Two-sided Mann-Whitney, $p < 0.001$). This behavior is present throughout all the years considered in our study, but if we observe the ratio between the medians of the two distributions (Figure 6, bottom) we see that the trend tends to decrease in the more recent years. This result suggests a possible adjustment of the recommendation algorithm to give less weight to the presence of these hashtags. Figure 7 shows the top 10 regions by prevalence with the relative percentage of videos that use 'viral' hashtags. It is noteworthy that even non-English speaking countries make use of English hashtags. This result possibly indicates the intent of the authors to reach an international audience. However, researchers should take care of potential biases when searching the API with specific English hashtags.

3.6 Prevalence of Conspiracy Theories

We are also interested in giving a preliminary outlook on the presence of hashtags that are related to mainstream conspiracy theories. We employed the LOCO dataset by Miani et al. [12] to obtain a list of the most prevalent conspiracy seeds (keywords), focusing on

the top 20 seeds (which describe approximately 45% of the articles on LOCO). We then filtered manually this list, to remove those keywords that were too generic (e.g., *5g*, *coronavirus*, *climatechange*, *barackobama*) and keep only the ones that are widely recognized as conspiracy theories. This resulted in a list of nine seeds of conspiracy theories that we employed as hashtags to search into our dataset plus the keyword 'conspiracy' as an additional check. We show the results of this search in Figure 8.

Only three of the nine seeds are present in the dataset, with very low percentages, with the hashtag 'conspiracy' being slightly more prevalent. The prevalence is higher if we focus on the 'US' region, since the hashtags themselves are in the English language. We also compute the 99% confidence intervals of these percentages via Clopper-Pearson method, which resulted in the order of 10^{-5}, thus not visible in the plot. Considering that the dataset we collected gathers videos from all around the world and that the number of videos posted on TikTok amounts to tens of billions ,[11] this analysis estimates that the number of conspiracy videos could amount to hundreds of thousands on a worldwide scale.

4 CONCLUSIONS

We collected a monthly-based stratified random sample of videos posted on TikTok by using the official Research API. We then provided a series of relevant insights and statistics about the performance of the API service and the data we obtained, especially oriented to the researchers who are planning to use this service. Our results describe a significant inability of the API to meet the quotas of requested videos, with a possible internal problem of data quality since querying videos from 2018 provided much fewer results compared to other periods. We then showed the growth of likes, views, comments, and shares over time, while also providing informative statistics about the global demography of the social media platform. Researchers should particularly pay attention to the latter, since the majority of videos on the platform originate from Asian countries, and authors in those countries also employ English-language hashtags. Finally, we showed that the videos that use typical engagement-oriented 'viral' hashtags have statistically more views and likes compared to the rest of the sample.

As with any empirical work, our research is subject to limitations. First of all, our sample is not uniformly random through the 6 years we set for the collection since it is stratified by month. The probability of a video being sampled from all the videos posted on TikTok in six years is much lower than the probability of the same video being sampled in the month of its creation. Moreover, the TikTok Research API service is a black-box system and we cannot explore the inner mechanisms that provide us with these supposedly-random results. This issue compounds with the lack of transparency for what concerns the removed content on this platform, which is currently inaccessible to researchers.

The presence of an official research API by TikTok has opened several research possibilities. It is now possible to study discourse quality on TikTok and examine several problems that afflict other platforms such as disinformation and coordinated inauthentic behavior. For instance, looking at their prevalence in these samples can give us an estimate of their presence on the whole social media. Still, the service offered is far from being ideal, due to the limited number of available requests and the convoluted documentation.

[11] https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2023-4/ accessed on 10/03/2024

REFERENCES

[1] Lada A Adamic, Rajan M Lukose, Amit R Puniyani, and Bernardo A Huberman. 2001. Search in power-law networks. *Physical review E* 64, 4 (2001), 046135. (Cited on 3)

[2] Dominik Bär, Francesco Pierri, Gianmarco De Francisci Morales, and Stefan Feuerriegel. 2023. Auditing targeted political advertising on social media during the 2021 German election. *arXiv preprint arXiv:2310.10001* (2023). (Cited on 1)

[3] Andreas Bernard. 2019. *Theory of the Hashtag*. John Wiley & Sons. (Cited on 3)

[4] Tugrulcan Elmas. 2023. The impact of data persistence bias on social media studies. In *Proceedings of the 15th ACM web science conference 2023*. 196–207. (Cited on 2)

[5] Deen Freelon. 2018. Computational research in the post-API age. *Political Communication* 35, 4 (2018), 665–668. (Cited on 1)

[6] Benjamin Guinaudeau, Kevin Munger, and Fabio Votta. 2022. Fifteen Seconds of Fame: TikTok and the Supply Side of Social Video. *Computational Communication Research* 4, 2 (Oct. 2022), 463–485. https://doi.org/10.5117/CCR2022.2.004.GUIN (Cited on 1)

[7] Daniel Klug, Yiluo Qin, Morgan Evans, and Geoff Kaufman. 2021. Trick and please. A mixed-method study on user assumptions about the TikTok algorithm. In *Proceedings of the 13th ACM Web Science Conference 2021*. 84–92. (Cited on 1)

[8] Yachao Li, Mengfei Guan, Paige Hammond, and Lane E Berrey. 2021. 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 (June 2021), 261–271. https://doi.org/10.1093/her/cyab010 (Cited on 1)

[9] Chen Ling, Krishna P Gummadi, and Savvas Zannettou. 2023. " Learn the Facts About COVID-19": Analyzing the Use of Warning Labels on TikTok Videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 554–565. (Cited on 1)

[10] Ryan McGrady, Kevin Zheng, Rebecca Curran, Jason Baumgartner, and Ethan Zuckerman. 2023. Dialing for Videos: A Random Sample of YouTube. *Journal of Quantitative Description: Digital Media* 3 (Dec. 2023). https://doi.org/10.51685/jqd.2023.022 (Cited on 1)

[11] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Dancing to the Partisan Beat: A First Analysis of Political Communication on TikTok. *12th ACM Conference on Web Science* (July 2020), 257–266. https://doi.org/10.1145/3394231.3397916 Conference Name: WebSci '20: 12th ACM Conference on Web Science ISBN: 9781450379892 Place: Southampton United Kingdom Publisher: ACM. (Cited on 1, 3)

[12] Alessandro Miani, Thomas Hills, and Adrian Bangerter. [n. d.]. LOCO: The 88-million-word language of conspiracy corpus. *Behavior research methods* ([n. d.]), 1–24. (Cited on 4)

[13] Adam M Ostrovsky and Joshua R Chen. 2020. TikTok and its role in COVID-19 information propagation. *Journal of adolescent health* 67, 5 (2020), 730. (Cited on 1)

[14] Arianna Pera and Luca Maria Aiello. 2023. Shifting Climates: Climate Change Communication from YouTube to TikTok. *arXiv preprint arXiv:2312.04974* (2023). (Cited on 1)

[15] Juergen Pfeffer, Daniel Matter, Kokil Jaidka, Onur Varol, Afra Mashhadi, Jana Lasser, Dennis Assenmacher, Siqi Wu, Diyi Yang, Cornelia Brantner, Daniel M. Romero, Jahna Otterbacher, Carsten Schwemmer, Kenneth Joseph, David Garcia, and Fred Morstatter. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. http://arxiv.org/abs/2301.11429 arXiv:2301.11429 [cs]. (Cited on 3)

[16] Lin Song and Avishek Ray. 2023. "How can a small app piss off an entire country?": India's TikTok ban in the light of everyday techno-nationalism. *Inter-Asia Cultural Studies* 24, 3 (2023), 382–396. (Cited on 3)

[17] Chris Stokel-Walker. 2020. TikTok's global surge. *New Scientist* 245 (2020), 31. https://api.semanticscholar.org/CorpusID:216240836 (Cited on 4)