CARMA 2024

6th Int. Conf. on Advanced Research Methods and Analytics
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/CARMA2024.2024.17765

The Invasion of Ukraine Viewed through Large-Scale Analysis of TikTok

Benjamin Steel [1], Sara Parker [2], Derek Ruths [1]

[1]Department of Computer Science, McGill University, Canada, [2]Media Ecosystem Observatory, McGill University, Canada.

How to cite: Steel, B.; Parker, S.; Ruths, D. 2024. The Invasion of Ukraine Viewed through Large-Scale Analysis of TikTok. In: 6th International Conference on Advanced Research Methods and Analytics (CARMA 2024). Valencia, 26-28 June 2024. https://doi.org/10.4995/CARMA2024.2024.17765

Abstract

The vast majority of TikTok analysis done to date has used small, manually-collected datasets. This leaves open the questions of how to do large-scale analysis of phenomena on TikTok and what can be learned about unfolding current events. In this paper, we seek to better understand how such events present themselves on TikTok by conducting a first examination of large-scale user and content dynamics around the invasion of Ukraine in 2022. As this is among the first studies to conduct large-scale (i.e., involving millions of data points) data collection and analysis of TikTok data, our contributions also include insights into best (and less-than-great) practices for such critical research tasks. Furthermore, we have open-sourced our Ukraine-invasion dataset and provide a software library that can be used to collect TikTok data.

Keywords: *tiktok; social media; Ukraine.*

1. Introduction

TikTok has rapidly become socially, politically, and economically important. The platform now boasts over one billion active users, and over a quarter of people below age 25 in the United States consider TikTok their primary news source (Matsa, 2022, Stokel-Walker, 2022). As has been witnessed on other social platforms, at this scale of adoption, trends on TikTok can have society-scale effects. There is urgency, then, in answering key questions: how can large-scale analysis be done on such a platform to study society-scale dynamics? What can be gleaned from such analysis about human behavior surrounding current events?

These questions are largely unaddressed. Existing research on TikTok has almost exclusively focused on small, curated datasets. This is due, in part, to the challenges posed by the design and novelty of the platform as compared to Twitter, Reddit, and other microblog-centric platforms. We considered these questions by undertaking the first large-scale study of TikTok content dynamics around the evolving Russian invasion of Ukraine - ultimately building and

working with a dataset of 9,500 (video) posts, 4.4 million comments, and 2.6 million users. This study forced us to engage with a number of methodological challenges that have long been settled on other platforms like Twitter - specifically, how to collect a representative dataset. As a result, part of our work provides an early guide for researchers seeking to use TikTok as a medium or object of study. With our dataset in hand, we looked at how language usage changed during the first months of the invasion, the dynamics of invasion-related topics, and capability of existing bot detection methods on the platform.

We make four contributions. First, we establish that TikTok manifests deeply contentious aspects of geopolitical events. Second, our findings make clear the need for new bot detection systems and paradigms for establishing dataset representativeness for TikTok. Third, this study provides a (albeit imperfect) template for TikTok data collection. And fourth, this study provides a dataset and data collection library: github.com/networkdynamics/pytok for future studies using TikTok.

2. Background

Prior studies of TikTok relating to the invasion of Ukraine have tended to involve small, manually collected datasets, typically involving between 50 and 500 posts, often without consideration of comments or metadata for those videos (ElHawary, 2023; Primig et al., 2023; Badola, 2023). Of existing TikTok studies, the largest we know of is (Medina Serrano et al., 2020) which studied partisan behavior using 8000 posts. To the best of our knowledge, this is the largest dataset of content on TikTok relating to the invasion of Ukraine by orders of magnitude. Our intent here has been to leverage this scale in order to assess phenomena that would be difficult or impossible to reliably measure with smaller datasets.

3. Dataset

The preparation of our dataset involved three steps: (1) building software for collecting data from TikTok, (2) designing and executing a methodology for collecting a broad, topically-coherent set of posts, and (3) filtering the data obtained. Work still needs to be done to improve TikTok data collection - we offer our process as a step along that path. We have released the data and code required to rebuild the dataset here: github.com/networkdynamics/ukraine-tiktok.

3.1. Collection Software Library

We required a method that both searches and downloads TikTok content. While TikTok does have a public API, it is not yet widely accessible. We investigated TikTok scraping libraries but found that none fit our needs: some libraries are browser automation-based, which results in slow collection times, while others are entirely requests-based, which is vulnerable to TikTok

backend API changes. We therefore developed an open-source library github.com/networkdynamics/pytok based off of github.com/davidteather/TikTok-Api. Our approach strikes a balance between prior approaches: browser automation for initial access, HTTP requests for fast access, and fallbacks to browser automation again if this fails. Our library has already been used by multiple other researchers, showing its demand.

3.2. Collection Methodology

Data access methods are limited on TikTok compared to Twitter, Reddit, or other platforms, with the main pathways being algorithmic feed, user-based search, and keyword-based search. Due to the black-box nature of the TikTok recommendation algorithm, we opted for a keyword-based search method. Using our library, we collected videos using a combination of hashtag and general search functionality. Hashtag searches are limited to the 1000 most viewed videos with that hashtag, so to expand the video set, we also used the general search functionality, which generally returns more videos for a search term. This approach is still non-ideal for data collection: the fuzzy nature of the black-box search functionality can return content unrelated to the search term; additionally videos appeared to be ordered in no available parameter-based order. Despite these limitations, we considered this the best of available options. We used a *seed-and-snowball* approach to expand the list of search terms: we did an initial informal search of TikTok, to find the most common hashtags used in popular videos related to the war, and we used this set as the seed set of hashtags (the first ten in the list below). We collected videos tagged with these terms, and examined the ranked co-occurring hashtags in the collected video descriptions to expand our search terms. The final hashtag set (with translations, and the corresponding language) was as follows: *standwithukraine*, *russia*, *nato*, *putin*, *moscow*, *zelenskyy*, *stopwar*, *stopthewar*, *ukrainewar*, *ww3*, володимирзеленський (Volodymyr Zelenskyy, ukr) , славаукраїні (Glory to Ukraine, ukr), путінхуйло (Fuck Putin, ukr), россия (Russia, ukr), війнавукраїні (War in Ukraine, ukr), зеленський (Zelenskyy, ukr), нівійні (No war, ukr), війна (War, ukr), нетвойне (No war, rus), зеленский (Zelenskyy, rus), путинхуйло (Fuck Putin, rus), *#denazification, #specialmilitaryoperation, #africansinukraine, #putinspeech, #whatshappeninginukraine*.

We ran this collection process in July 2022 and April 2023. The behavior of historical search in TikTok is currently uncharacterized - raising concerns over completeness. However, as we see in the count of content over time in Fig. 1, the number of results returned across our collection period was stable, implying that query timing does not greatly affect the sample. We also collected the comments for each of these videos (limiting to 1000 comments per video to ensure reasonable collection duration), to provide additional text data for analysis.

Due to the noise of the search functionalities, and hashtag abuse, there were some irrelevant videos in the raw dataset. We therefore removed this content from the data to produce the final

dataset, ensuring it contains only videos related to the invasion of Ukraine. We define "related to the invasion" as any video containing one of the following items:

Depictions or discussions of combat; Support or protest for either side's war efforts, including propagandistic content; Any mention of Putin or Zelenskyy during the invasion; Critical political and military leaders engaging with the invasion; Videos about direct social or economic outcomes of the war; Speculation about the war; Videos about the militaries of countries involved in the war posted during the invasion (as implied propaganda).

With this definition, we took a sample of 300 videos from the dataset, opened each video in TikTok, and manually labelled them as related or not. We found that of the videos that were still available, 63% were related to the war. 29% were no longer available. That such a high percentage of the content collected was not available 6 months after collection shows how ephemeral content is on TikTok. We fine-tuned a RoBERTa large language model (LLM) (Liu et al., 2019) using our 300 labelled videos descriptions to classify the videos as being related to the invasion or not, then used it to filter the dataset to only war related videos. For this filtered set, we found that of those still available, 93% were related to the war. Post-filtering, the full dataset contains text and metadata for approximately 9.5 thousand videos related to the invasion of Ukraine with 4.4 million comments, from 2.6 million users. Of users who posted a video, the mean number of videos posted was 1.7, with a max of 152. Of users who posted a comment, the mean number of comments was 1.7, with a max of 832. Given that we limited our max number of comments per video to 1000, the mean number of comments per video was 766. Fig. 1 shows the number of videos and comments in our dataset over time.

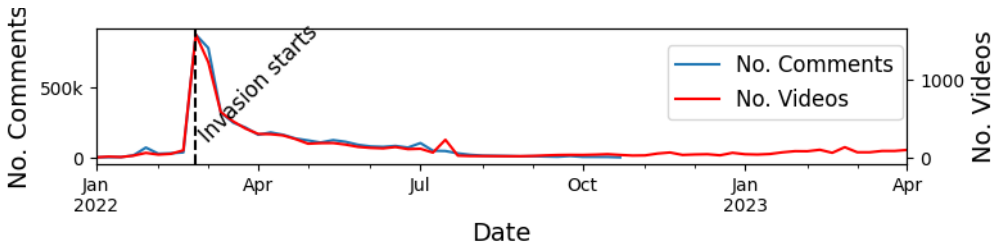

Figure 1: Number of videos and comments in our dataset over time.

4. Experiments

4.1. Words and Languages

We first took a macro view of how language evolved on the platform. Specifically, we wanted to measure the use of languages and words from the beginning of the invasion and onwards, to understand if there was any one changepoint at the start of the invasion, or if there were more complex, ongoing changes as communities and the platform itself responded to the unfolding

crisis. To do this temporal analysis, we used simple keyword, hashtag, and language searches in the comments we have from the videos, with language data provided by TikTok data. Where common keywords have multiple spellings, we have searched for all of these terms, and summed the counts to find the final search count. Note that these results are not meant to provide a comprehensive understanding of language patterns, but rather to show what possible new effects can be uncovered with large-scale analysis of TikTok data.

At this level of social interaction, we saw some behavioural changes over time. Notably, we saw evidence of mass movement to the Ukrainian language instead of Russian over time by users who at some point used Ukrainian, Fig. 2. We also saw a change from majority English text to Russian text over the course of the invasion in Fig. 3, indicating sustained attention from Russian speakers, or a decrease in attention from English speakers, or both. It seems likely that the medium of video allows greater mutual interaction between different language populations, allowing language dynamics that may not be seen on a text based platform.

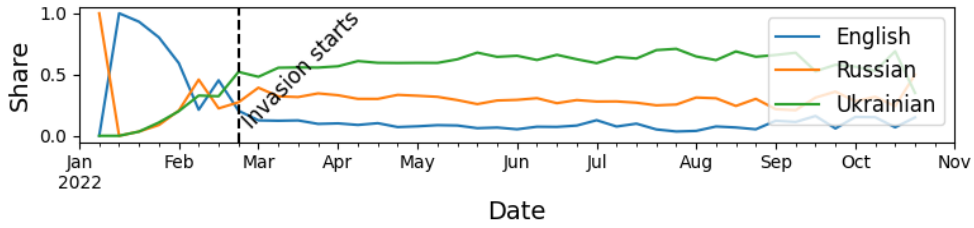

Figure 2: Language use for users who at some point use the Ukrainian language

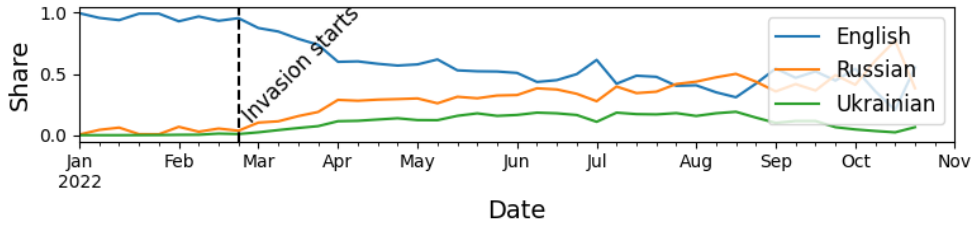

Figure 3: Language use over time

We also examined country and leader name mentions, finding that attention on Putin remained consistent throughout the war, but attention on Zelenskyy stays low (Fig. 4). Conversely, we saw sustained attention on Ukraine, but quickly diminishing attention on Russia in Fig. 5. This presents a curious juxtaposition among TikTok users: maintaining a focus on Ukraine and events within it while paying attention to Putin (rather than Zelenskyy).

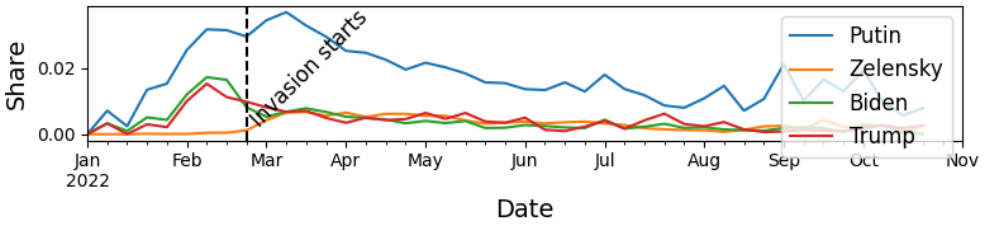

Figure 4: Leader mentions over time.

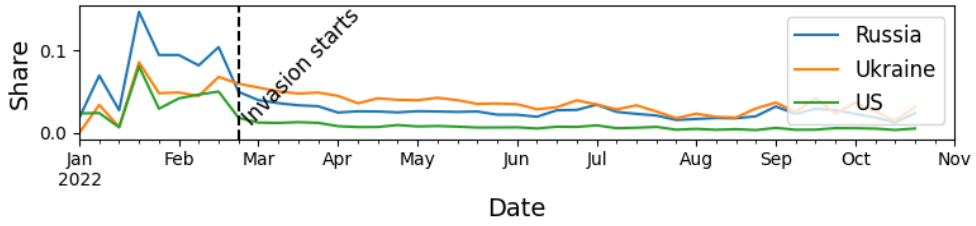

Figure 5: Country mentions over time.

4.2. Topics

We used topic modelling to examine video descriptions, as they provide us a view into the major themes on the platform. We used the BERTopic library (Grootendorst, 2022) and a multilingual Twitter fine-tuned LLM (DeLucia, 2022). In Fig. 6 we can see a diverse range of topics reflecting various political perspectives across the platform. Alongside popular topics aligned with general public media discourse that we would expect to see (NATO and Biden's relationship with the war, the position of Poland in the war), we also see more TikTok specific discussion, such as Eurovision's part in European solidarity and fears of an invasion of Alaska, indicative of nuanced discourse unique to the platform.

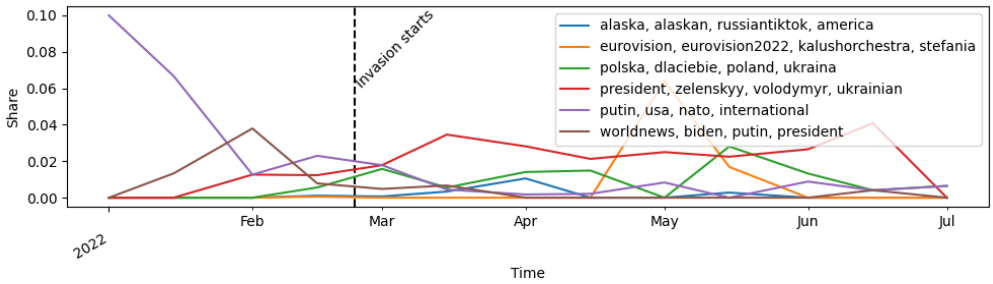

Figure 6: Selection of topics over time.

4.3. Bots

Bot detection is a crucial tool in large-scale social media analysis. We used a free and open source Twitter bot classifier focusing on generalizability (Ram, 2021), operationalizing the classifier features with the closest TikTok feature. We did not anticipate that this model would work perfectly on TikTok data, but we wanted to understand how poor the performance would be. In short, the performance was very bad. 99.2% were scored as likely to be a bot by the bot detection system, which, looking at the data, is unlikely. Examining features that were most attended to by the classifier, we find that verified status, following count, and account age are the top contributors to these classifications. It's clear that the predictive relationships between being a bot or not are very different between TikTok and Twitter.

5. Discussion

TikTok manifests intriguing social dynamics around current events. Where language is concerned, we found clear shifts in population use of languages that track and are explainable within the context of the invasion of Ukraine. That geopolitical events would manifest themselves in changes in language use is a striking example of the unexpected cultural impacts we can uncover through large-scale social media analysis. The large-scale dynamics in topical engagement shows how TikTok reflects discourse around the unfolding invasion. Moreover, the dynamics surrounding country and leader attention underscores the phenomena that can emerge and be studied within platforms with a short video-sharing mechanism.

TikTok requires revisiting fundamental aspects of large-scale analysis. Our work highlighted the need for substantial additional work in (at least) two areas. First, we as a research community lack the frameworks and baseline statistics for assessing and designing representative datasets from TikTok. Second, tools for identifying bot-generated content are completely lacking. Until we develop methods to address them, these issues will hamper large-scale studies of TikTok. In spite of the limitations highlighted, we consider the data preparation process, the data collection library, and the dataset to be valuable resources for the community as we collectively improve our capacity to conduct large-scale studies on TikTok. As evidenced by our findings, this effort is warranted: there is a great deal of social, political, and economic value to be gleaned from society-scale studies of the TikTok platform.

References

Badola, P. (2023). Russia and Ukraine: A Content Analysis of "The World's First TikTok War".

DeLucia, A., Wu, S., Mueller, A., Aguirre, C., Resnik, P., & Dredze, M. (2022, December). Bernice: A multilingual pre-trained encoder for Twitter. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 6191-6205).

ElHawary, D. M. M. (2023). TikTok Battlefield: Comparative Analysis of English and Arabic Language Representations of The 2022 Russian Ukrainian Conflict On TikTok (Doctoral dissertation, The American University in Cairo (Egypt)).

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Matsa, K. E. (2022). More Americans are getting news on TikTok, bucking the trend on other social media sites. Pew Research Center, 21.

Medina Serrano, J. C., Papakyriakopoulos, O., & Hegelich, S. (2020, July). Dancing to the partisan beat: A first analysis of political communication on TikTok. In Proceedings of the 12th ACM Conference on Web Science (pp. 257-266).

Primig, F., Szabó, H. D., & Lacasa, P. (2023). Remixing war: An analysis of the reimagination of the Russian–Ukraine war on TikTok. Frontiers in Political Science, 5, 1085149.

Ram, R., Kong, Q., & Rizoiu, M. A. (2021, March). Birdspotter: A tool for analyzing and labeling twitter users. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (pp. 918-921).

Stokel-Walker, C. (2022). TikTok wants longer videos-whether you like it or not. Wired. com. URL: https://www. wired. com/story/tiktok-wants-longer-videos-like-not/[accessed 2022-07-10].