Invasion@Ukraine: Providing and Describing a Twitter Streaming Dataset That Captures the Outbreak of War Between Russia and Ukraine in 2022

Janina Susanne Pohl[1], **Simon Markmann**[1], **Dennis Assenmacher**[2], **Christian Grimme**[1]

[1] Computational Social Science and Systems Analysis, University of Münster, Münster, Germany
[2] Computational Social Science, GESIS, Cologne, Germany
janina.pohl@uni-muenster.de, simon.markmann@alumnos.upm.es, dennis.assenmacher@gesis.org,
christian.grimme@uni-muenster.de

Abstract

Social media can be a mirror of human interaction, society, and historic disruptions. Their reach enables the global dissemination of information in the shortest possible time and, thus, the individual participation of people worldwide in global events in almost real-time. However, these platforms can be equally efficiently used in information warfare to manipulate human perception and opinion formation. Within this paper, we describe a dataset of raw tweets collected via the Twitter Streaming API in the context of the onset of the war, which Russia started in Ukraine on February 24, 2022. A distinctive feature of the dataset is that it covers the period from one week before to one week after Russia invasion of Ukraine. This paper details the acquisition process and provides first insights into the content of the data stream. In addition, the data has been annotated with availability tags, resulting from rehydration attempts at two points in time: directly after data acquisition and shortly before manuscript submission. This may provide information on Twitter moderation policies. Further, we provide a detailed list of other published dataset covering the same topic. On the content level, we can show that our dataset comprises several distinct topics related to the conflict and conspiracy narratives – topics that deserve more profound investigation. Therefore, the presented dataset is also made available to the community in an extended version with pseudonymized tweet content upon request.

Introduction

Social media have become a critical information space when considering world historical changes or disruptions (such as those caused by acts of war). In addition to the documentary value of the data[1], they are increasingly the subject of scholarly analysis to investigate and perhaps even explain social processes or opinion formation in the context of dedicated events.

In particular, the information space of social media has also emerged as a potential (concomitant) battleground for disinformation, hate speech, and fake news. Open accessibility in the sense of essentially free participation has a downside in addition to positive aspects of free expression: the technical infrastructure is susceptible to manipulation and can be misused for the automated duplication of content. As a result, any disinformation can be disproportionately disseminated and distort the image of opinion depicted by a platform (possibly by algorithmic means through recommenders). The media reception of this distorted opinion picture can lead to a strong multiplication of false or manipulative content. Although informational warfare is not a new phenomenon from the age of social media, its development has been greatly accelerated and globalized by using modern information and communication technologies (Stupples 2015; Prier 2017). Information warfare often involves disinformation campaigns, which can be carried out by very different means – from automation to (state) coordination of human actors (Metaxas and Mustafaraj 2012; Grimme et al. 2017; Keller et al. 2020).

The starting point for the analysis of such incidents, for the possible misuse of technical infrastructure, as well as for the detection of (disinformation) campaigns, is recording such activities and making them available to science. It is of great importance that relevant data sets on significant events are recorded and available for analysis by the scientific community. However, this idea of social media data sharing often contradicts reality (Bruns 2019; Assenmacher et al. 2021). Social media data is usually managed solely by platform operators and kept under lock and key through licensing agreements. In rare cases, platform operators make selected data fragments available to the scientific community. However, these are preprocessed and possibly truncated. Open documentation of the preprocessing and the original data is usually not available. Subsequent monitoring and querying of the actual activities on the platforms are hardly possible, as the platforms refuse to release problematic and blocked content, citing (often pretextual) data protection regulations or ethical considerations.

Therefore, it is even more critical to collect data directly during a crisis or war and make it available to the scientific community for scientific purposes. This paper describes a raw Twitter dataset collected via the Twitter Streaming API from one week before Russian forces began invading Ukraine on February 24, 2022, to one week after the invasion. In this paper, we comment on the value of the data, the method of data collection, and the ethical implications and

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] Even many international governmental bodies (e.g.https://www.nationalarchives.gov.uk/webarchive/ or https://www.archives.gov/) store social media data on official accounts for documentation purposes.

limitations of the dataset provided based on an initial technical report (Pohl et al. 2022b). Additionally, this paper highlights the dataset's potential on the content level by identifying various thematic areas of interest and placing them in the context of wartime action.

Related Datasets

Since the start of the Russian war on Ukraine, many social media datasets containing posts related to this conflict have been published. Here, we want to list and assess the content and value of other datasets in contrast to ours to provide the reader with an overview of available data sources. We searched for publicly available datasets by querying the digital libraries IEEE Xplore, Scopus, the Web of Science, the open access repositories arXiv and SSRN, as well as the (data) repositories Kaggle, Zenodo, and GitHub. In total, we found 18 publicly available datasets and four works in which the mode of data collection was explained, while the data itself is not available or behind a paywall. An overview of the datasets, including relevant meta information, can be found in Table 1.

Most published papers presenting datasets cover posts from the micro-blogging platform Twitter. Haq et al. (2022) collected original tweets, retweets, and quotes via the Twitter Streaming API from two days before the start of the war until November 2022. First, they used general hashtags related to the Ukraine war (e.g., #russia, # ukraine, etc.) but later added hashtags related to specific battlegrounds to their query. Shevtsov et al. (2022) started their data collection via the Streaming API for a vast list of multilingual hashtags on February 24 until September and "backfilled" to February 22 via the Search API. Similarly, Chen and Ferrara (2022) employed backfilling, streaming data from February 22 onward and supplemented it to January 1 by filtering for hashtags in multiple languages. In an associated work, Pierri et al. (2022) added labels to the dataset according to the tweet's availability via the API in November 2022. Smart et al. (2022) only used the Twitter Streaming API to gather original tweets, retweets, and quotes related to variants of the hashtags #IStandWithRussia and #IStandWithUkraine from February 23 to March 8. In contrast, Caprolu, Sadighian, and Di Pietro (2022) only used the historic Search API to collect original English tweets from January to March 2022, using a general list of hashtags related to the conflict as a filter. However, the authors only described their method and query but did not publish tweet IDs. Gosh (2022)'s dataset is also not freely available but behind a paywall of IEEE's dataport. The authors collected 55K unique tweets related to popular hashtags (popular at this point in time) in 57 different languages. Also, they provided metadata like the number of likes, retweets, etc., which enables a user network analysis. Likewise, (Soares et al. 2022) published their data on the Canadian data-sharing platform Borealis along with network analysis to identify opposed user groups. Their Twitter data is related to the claim that the Russian forces conducted a chemical attack in Mariupol. Park et al. (2022) collected tweets issued by over forty Russian news outlets for one week after the start of the war, searched for often-used hashtags, and used these hashtags for further data collection from Twitter via the historic Search API, thus covering the Russian perspective on Twitter. Similarly, Toraman et al. (2022) covered the Turkish perspective on the conflict, focusing on online manipulation. They used fact-checking platforms to identify mis- and disinformation events and then used the Twitter Search API to collect related tweets. Additionally, they labeled their dataset w.r.t. whether the tweet contains facts or false information. Thapa et al. (2022) collected tweets and images related to tweets, generating a dataset of 5.6 K image/text pairs. Additionally, they labeled their data according to whether the tweets contained hate speech or not.

Next to the datasets presented in an accompanying paper, many datasets were only uploaded on data-sharing platforms together with documentation. Münch and Kessling (2022) published their Twitter dataset related to the hashtags #ukraine and #bucha on OSF. They used the Streaming API from February 27, backfilling data from February 1 onward via the Search API. Preda (2022) published his Twitter dataset collected via the Search API related to the hashtag #slavaUkraini on Kaggle, covering Ukrainian and English tweets issued in March 2022. Purtova (2022) published her dataset on Kaggle, including English tweets related to Russian and Ukrainian troops, borders, and support. The data covers the time frame from January 1 to March 16. In contrast, the dataset by Mukherjee (2022) and BwandoWando (2023b) are not tagged with the list of used hashtags. Mukherjee (2022) covers tweets from December 31, 2021, to March 6, 2022, while BwandoWando (2023b) updates his dataset queried every 15 minutes from the Search API regularly since the day before the war started. Note that the Kaggle dataset contains complete tweet objects, while the other publicly available datasets only contain tweet IDs, adhering to Twitter's Terms of Service (ToS).

Only some datasets also exist for other platforms. Pierri et al. (2022) collected Facebook data using CrowdTangle from January 1 related to the same multilingual hashtags they used in their previous work (Chen and Ferrara 2022). However, they neither publish the complete dataset nor the IDs of Facebook posts. Zhu et al. (2022) use Pushshift to gather related data on the war from Reddit (comments from various Ukraine- and military-related subreddits) from the day of the beginning of the war until June 2022. Hanley, Kumar, and Durumeric (2023) also use the Reddit API and Pushhift to gather data from the subreddit *Russia* and other political subreddits to understand the spread of Russian state media narratives. Another dataset by BwandoWando (2023a) also covers seven months of Reddit data from the subreddit *Ukraine* after the beginning of the war. Chen et al. (2022) and Fung and Ji (2023) collected data from the Chinese social media platform Weibo using a publicly available Python scraper (version 1.0.6). Both started their data collected a few days before the beginning of the war, covering the Chinese perspective on the conflict. Finally, next to their Twitter dataset covering the Russian perspective, Park et al. (2022) also sample data from the Russian Platform VKontakte related to Russian media outlets. This dataset covers over 20 million posts starting from the beginning of 2021 to mid-2022.

Paper	Platform	Collection Time	API	No. Posts	Upload	Languages	Avbl.	Rh.
Haq et al. (2022)	Twitter	22/02/21 - 22/11/06	Stream	> 30 M	GitHub	Multiple	Free	No
Shevtsov et al. (2022)	Twitter	22/02/22 - 22/09/21	Both	> 30 M	GitHub	Multiple	Free	No
Chen&Ferrara (2022)	Twitter	22/02/22 - 22/10/01	Both	> 30 M	GitHub	Multiple	Free	No
Smart et al. (2022)	Twitter	22/02/23 - 22/03/08	Stream	5.2 M	figshare	Eng	Free	No
Caprolu et al. (2022)	Twitter	22/01/27 - 22/03/23	Search	5.5 M	–	Eng	No	No
Gosh (2022)	Twitter	22/02/22 - 22/04/18	Search	55 K	Dataport	Multiple	Paywall	–
Soares et al. (2022)	Twitter	22/04/06 - 22/04/13	Search	246 K	Borealis	Eng	Free	No
Park et al. (2022)	Twitter	22/02/24 - 22/05/14	Search	18.5 M	GitHub	Multiple	Free	No
Toraman et al. (2022)	Twitter	2020 - 2022	Search	10.3 K	GitHub	Eng, Tur	Free	No
Thapa et al. (2022)	Twitter	22/02/22 - 22/03/28	Search	5,7 K	GitHub	Eng	Free	No
Münch&Kessling (2022)	Twitter	22/02/01 - today	Both	> 30 M	OSF	Multiple	Free	No
Preda (2022)	Twitter	22/03/01 - 22/03/24	Search	30 K	Kaggle	Eng, Ukr	Free	No
Purtova (2022)	Twitter	22/01/01 - 22/03/06	Search	325 K	Kaggle	Eng	Free	No
BwandoWando (2023b)	Twitter	21/08/19 - today	Search	> 30 M	Kaggle	Multiple	Free	No
Mukherjee (2022)	Twitter	21/12/31 - 22/03/04	Search	225 K	Kaggle	Eng	Free	No
Pierri et al. (2022)	Facebook	22/01/01 - 22/04/24	CrowdTangle	19.5 M	–	Multiple	No	No
Zhu et al. (2022)	Reddit	22/02/24 - 22/06/13	Pushshift	8.3 M	GitHub	Eng	Free	No
Hanley et al. (2023)	Reddit	22/01/01 - 22/03/15	Pushshift	5.5 M	–	Eng	No	No
BwandoWando (2023a)	Reddit	22/02/28 - 22/09/06	API	1.82 M	Kaggle	Eng	Free	No
Chen et al. (2022)	Weibo	22/02/19 - 22/03/05	Scrapy	100 K	GitHub	Chn	Free	No
Fung&Ji (2023)	Weibo	22/02/21 - today	weibo-scraper	3.5 M	GitHub	Chn	Free	No
Park et al. (2022)	VK	21/01/01 - 22/05/15	VK Open	21 M	GitHub	Rus	Free	No
This Paper	Twitter	22/02/17 - 22/03/03	Stream	8.7 M	GESIS	Eng	Free	Yes

Table 1: Overview of the published datasets and related metadata on the Russian Invasion of Ukraine for various platforms. Next to information about the collection time and mode as well as basic statistics, the table also shows whether the dataset is freely available (avbl.) and information whether the posts are still online or not is given (Rh.).

Although such a high number of datasets were already published, the dataset we provide in this work offers a new perspective on the ongoing conflict as documented on Twitter. To the best of our knowledge, it offers a unique insight into some of the unfiltered Twitter communication related to the invasion of Ukraine by Russian troops. First, it covers a unique timespan collected via the Twitter Stream API, as seen in Table 1. We started one week before and ended one week after the beginning of the war with our data collection. Other papers (e.g., Haq et al. (2022) or Shevtsov et al. (2022) backfill their data from before the beginning of the war with the Search API). By using the Streaming API during this timespan, content later flagged as problematic (and thus, subsequently made unavailable) is also included in our dataset. Further, our data collection only includes original tweets, in contrast to, e.g., Smart et al. (2022). Additionally, our use of the hashtags #ukraine, #russia, and #conflict is broad enough to cover various trending topics. Since our query frequently reached the stream cap of the API during our data collection process, we are confident that our dataset contains the most extensive collection of unique tweets compared to the other datasets. In contrast, Chen and Ferrara (2022) uses a long, multilingual list of keywords to query the Stream API. Presumably, the broader data stream may hit the 1% cap of the API earlier than with our small set of keywords. Lastly, we also label our dataset according to the tweet's availability directly after the start of the conflict in March 2022 and months later, in January 2023, after Elon Musk's purchase of Twitter. Thus, we also provide valuable information on Twitter's moderation policies then and now.

Overall, the provided dataset augments the already available datasets by providing a new perspective that comprises

1. two weeks of original English tweets before and after the beginning of the war,
2. including later-on suspended content and meta-information,
3. covering many trending topics due to the use of rather general hashtags for querying the Stream API,
4. and labeled according to their availability via the Search API at two distinct points in time.

Data Description

We gathered 8.7 million original tweets between February 17 and March 3, 2022, produced by 2.3 million individual user accounts in total. To enrich the data with additional information on how many tweets were deleted since they were posted, we rehydrated the dataset twice, i.e., we checked their availability via the Twitter Search API. To help fellow researchers working with this dataset, we additionally conduct a superficial content analysis to explore possible research topics.

Rehydration

After we finished our data collection process, we checked each tweet's status via the Twitter Search API for the first time in March 2022 and repeated this process in early January 2023 to see how it changed. The results can be seen in Figure 1 and Table 2, which displays the total number of tweets posted daily and how many got deleted at the two

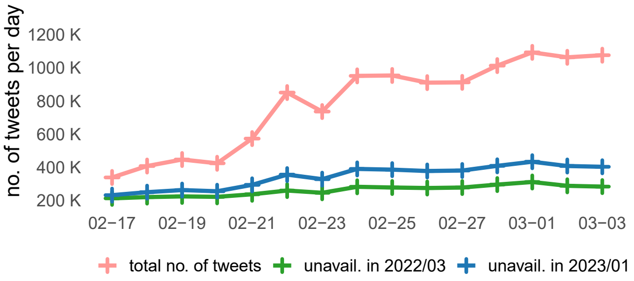

Figure 1: Number of collected tweets per day over the collection period. Tweets are separated according to their availability after the two rehydration processes.

Date	Unavail. 22/03	Unavail. 23/01	Online
22/02/17	9.2 %	22.8 %	0.3 %
22/02/18	9.7 %	23.8 %	0.3 %
22/02/19	9.8 %	25.3 %	0.3 %
22/02/20	9.8 %	24.8 %	0.3 %
22/02/21	10.0 %	25.1 %	0.3 %
22/02/22	9.2 %	23.9 %	0.4 %
22/02/23	8.6 %	23.8 %	0.3 %
22/02/24	11.0 %	25.3 %	0.5 %
22/02/25	10.4 %	24.7 %	0.4 %
22/02/26	10.6 %	25.0 %	0.3 %
22/02/27	11.0 %	25.3 %	0.3 %
22/02/28	11.7 %	25.6 %	0.3 %
22/03/01	12.4 %	26.9 %	0.3 %
22/03/02	10.2 %	24.0 %	0.3 %
22/03/03	9.5 %	23.1 %	0.2 %

Table 2: Percentage of unavailable tweets at the two rehydration time points. The last column displays how many tweets unavailable in the first inquiry were available in the second.

points under consideration. The numbers more than doubled, from roughly 10 % to 24 %. Compared to other Twitter studies investigating suspended content, this fraction is significantly more extensive and seems to deviate from the norm (Chowdhury et al. 2020; Majó-Vázquez et al. 2021). Nevertheless, the reasons for content/account removal can be manifold (e.g., violations of the ToS or self-determined deletion). Obtaining this knowledge in advance is impossible, as the Search API does not disclose this information.

To investigate possible changes in Twitter's moderation policy, we computed which amount of unavailable tweets from the first rehydration (March 2022) was available again at the second rehydration (January 2023), so-called "zombie" tweets. The reappearance of tweets can happen if, e.g., a user is suspended for some time and can publish tweets again or in case of reactivation due to changes in moderation policy. As can be seen in Table 2, on each date, 0.2 % to 0.5 % of the tweets were available again in January 2023 after they were not available in March 2022. At least in this context, no relaxation of Twitter's moderation policies can be observed.

Content Analysis

In order to provide readers with a comprehensive overview of the potentially problematic topics contained within our dataset, we analyzed the tweets' content. Given the large number of tweets in our dataset, we randomly sampled one million tweets from the entire time. To visualize the various topics present in the data, we utilized a 2-dimensional corpus projection using Uniform Manifold Approximation (McInnes, Healy, and Melville 2018) on Sentence-Bert (S-BERT) embeddings (Reimers and Gurevych 2019). We used a threshold of 0.8 for the cosine similarity to assess the distance between vectors.

Our approach produced many clusters, which we subsequently filtered by eliminating clusters containing fewer than 250 elements. This resulted in 76 clusters, as depicted in Figure 2. As this number of clusters remained relatively high, we selectively colored only those related to the call for support for Ukraine. This shows the manifold topics in our dataset that other researchers can investigate further. It is worth mentioning that Figure 2 also reveals a peculiarity: a smaller cluster attached to the boundaries of a larger cluster, colored in red. This cluster comprises tweets about a cryptocurrency scam that occurred at the war's outset. On February 26, 2022, the official Twitter account of the Ukrainian government announced that it would accept donations in Bitcoin, Ethereum, and Tether and included wallet IDs for each currency[2]. This announcement prompted several accounts to post identical tweets but with alternate wallet IDs substituted for the original ones. We investigate this campaign later in this chapter.

Another potentially polarizing topic we integrated into our investigation is a Russian narrative for starting the invasion: the biolab narrative (i.e., the telling that the US is running laboratories for researching bioweapons in Ukraine). Although this topic is not present in the most relevant topics detected by our clustering approach, it was extensively covered in the media and used in official statements of the Russian Ministry of Defense. Thus, we will explore the occurrence of this conspiracy theory in our data as another research topic, which may be investigated using our dataset.

We refrain from conducting a more in-depth analysis of other topics in the dataset here. However, the result of our analysis in Figure 2 shows the various topics and research opportunities our dataset provides for other researchers.

Cryptocurrency Scams As can be seen in Figure 2 in the red clusters, a cryptocurrency scam was issued during the first weeks of the war. Here, scammers exchanged the Ukrainian wallet IDs with their own IDs to receive the money from the donations. We were interested in assessing the campaign's success and finding out whether people donated money to the wrong wallets, so we checked a sample of the wallet IDs via blockchain[3] that were posted most often and had not been deleted yet. We could not find any transactions on these wallets after February 26, 2022, meaning the campaign stopped shortly after the outbreak of the war.

[2] https://twitter.com/Ukraine/status/1497594592438497282

[3] https://www.blockchain.com/

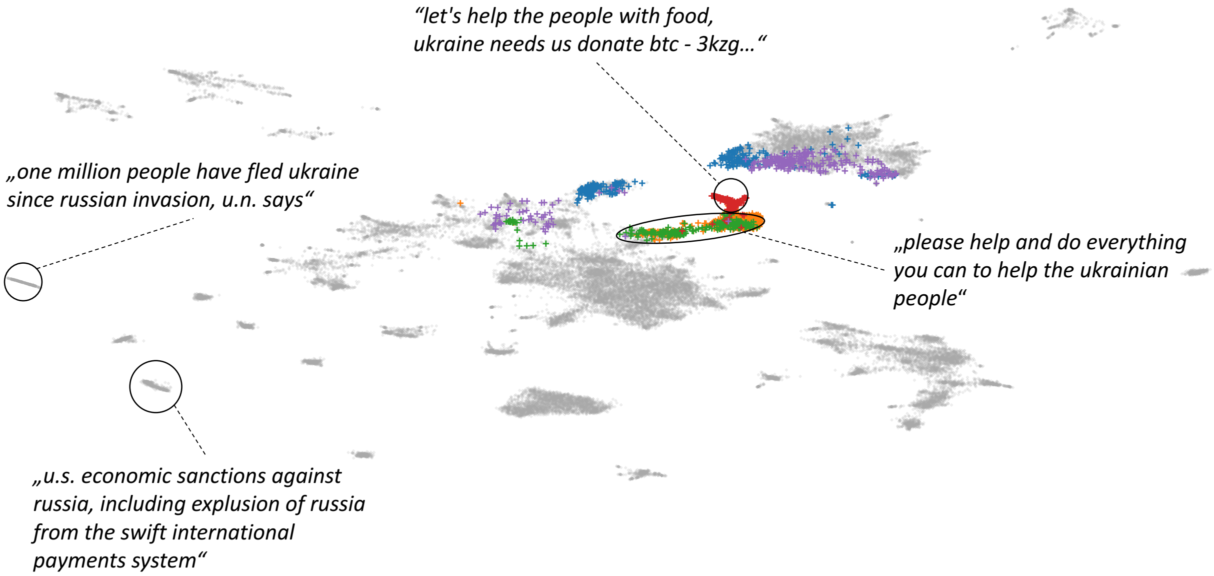

Figure 2: A 2-dimensional corpus projection using Uniform Manifold Approximation (McInnes, Healy, and Melville 2018) on S-BERT sentence embeddings (Reimers and Gurevych 2019). The colored observations highlight tweets discussing potential ways of helping and supporting Ukraine. The red cluster represents tweets associated with the Cryptocurrency scam described in Section . The Figure further illustrates our dataset's extensive scope of additional topics.

Furthermore, we investigated why some tweets with the wrong wallets got deleted, and others did not. One user posted the fake donation tweet 132 times, yet none of these tweets were deleted. This is an example of how not even spamming the tweet resulted in deletion. We even found one tweet from another user, which was still available on January 11, 2023. From the data exploration, we could not find any rule (applied by Twitter) determining whether a fake donation tweet got deleted or not.

Biolab Narrative Another conspiracy narrative that has flourished since the start of the conflict in late February is that Russia attacked Ukraine because the Ukrainian government allowed the US to establish laboratories to develop biological weapons (Maschmeyer 2021; Chakravarty 2022). Especially in the QAnon community in the US, this conspiracy has found some followers[4]. We evaluated to what extent this process can be traced in our Twitter dataset by searching for variations of the keywords *biolab* and *bioweapon*.

The results are presented in Table 3: the date after the number of tweets containing the keywords is given next to the percentage of these tweets being unavailable each time we rehydrated the dataset. Until February 24, 2022, the average amount of tweets per day in which one of the tags appeared was around 20. On February 24, 2022, this number suddenly rose to 698 and the following day to 1888. All of a sudden, the topic gained popularity right after the first Russian attacks. Until March 03, 2022, the number of daily tweets decreased slightly.

Date	No. Tweets	Unavail. 22/03	Unavail. 23/01
22/02/17	22	5 %	36 %
22/02/18	10	10 %	40 %
22/02/19	14	0 %	71 %
22/02/20	24	0 %	25 %
22/02/21	12	0 %	8 %
22/02/22	26	4 %	92 %
22/02/23	29	10 %	52 %
22/02/24	698	14 %	45 %
22/02/25	1888	10 %	33 %
22/02/26	1562	11 %	34 %
22/02/27	1686	9 %	31 %
22/02/28	1863	9 %	28 %
22/03/01	1823	11 %	35 %
22/03/02	1327	10 %	37 %
22/03/03	1470	8 %	34 %

Table 3: Amount of (unavailable) tweets related to the biolabs conspiracy theory in our dataset over time.

Considering our rehydration analysis in March 2022, a clear pattern can be observed: before the start of the war, Twitter's content moderators or the users themselves deleted less than ten percent of the tweets related to the biolabs conspiracy theory. Then, until the last day considered in our dataset, over ten percent of the tweets were deleted when the conspiracy theory got reignited by the start of the war. Nearly ten months later, our second rehydration analysis reveals that the absolute number of unavailable tweets either increased or stayed equal. These results highlight the highly dynamic nature of Twitter datasets (Zubiaga 2018), as

[4] https://www.cnn.com/2022/03/09/media/biolab-ukraine-russia-qanon-false-conspiracy-theory/index.html

well as the importance of our dataset, which preserves these deleted tweets in a pseudonymized form and, with them, a relevant part of the conspiracy pattern published around the considered event.

Dataset Creation

We collected the dataset live from the Twitter Stream API. From February 17 to March 03, 2022, we received 1 % of tweets related to the hashtags #ukraine, #russia, and #conflict. We decided to use these extensive terms to be open to any direction the conflict might take and to collect as many sub-topics as possible since we could not know which hashtags would be trending in advance. As the conflict evolved, we did not narrow our search terms to minimize our dataset's content shift and bias, as well as to be open to any upcoming trends.

After a general inspection of our dataset, we identified tweets unrelated to the Russian-Ukrainian conflict. Thus, we decided to post-process our dataset by filtering the tweets for posts that contain the words *ukraine*, *ukrainian*, *russia*, *russian*, *putin*, *zelensky* or *kremlin*. After the post-processing, our dataset consists of 8.7 million tweets produced by 2.3 million unique user accounts.

Additionally, we rehydrated the tweets to receive the tweets' availability information to enrich our dataset. We did so the first time in the middle of March, i.e., after our data collection phase, to check which tweets were immediately deleted by either Twitter or their original author. In January 2023, we repeated the analysis for several reasons: first, we wanted to check whether more tweets were unavailable nine months after our initial research. Second, we tried to assess the effect of Elon Musk's purchase of Twitter. His announcement[5] to make Twitter's content moderation policies more moderate might affect their Tweet deletion behavior. Finally, we pseudonymized the author names of the now unavailable tweets and any user mentioned in the tweet.

Usage Notes

Given the mode of data collection, the context, the results of our initial analysis, and the labeling of the dataset, we propose potential areas of research that can be investigated, as well as limitations that must be kept in mind while using our dataset.

Research Areas of Interest

Content On the content level, our data basically allow the analysis of communication activities of users (personal opinion), state actors (presentation of their perspective, propaganda), and journalists (reporting). At the same time, the data are also interesting for researching disinformation, propaganda, campaign narratives, or automation.

For example, the results of our initial analysis in Section show evidence of scams and conspiracy theories. The fake cryptocurrency wallet IDs included in genuinely-looking tweets provide an opportunity to investigate fraud over time and across platforms. Even before the war, the biolab conspiracy theory was discussed by researchers and media (Maschmeyer 2021; Chakravarty 2022). Here, our dataset provides research opportunities to study how the campaign supported this theory evolved after it was reignited by the start of the war.

Campaign Analysis Some labeled clusters in Figure 2 show campaigns containing tweets calling for help for Ukraine. These original campaigns, presumably launched by the Ukrainian government, target international organizations, politicians, and organizations seeking support in defense against the Russian invasion. Identifying and proofing Russian-backed campaigns remains an open research task for the future. Thus, our dataset is highly suitable for exploring and studying not only a large number of varying topics related to the war but also state-backed social media campaigns via possibly coordinated automated multiplication of content.

Twitter's Moderation Policies Since we labeled our dataset with the tweet's availability in March 2022 and January 2023, it can be used to study the results of Twitter's content moderation efforts. Our analysis of the cryptocurrency scams revealed, at first glance, no discernible pattern of which accounts were suspended or which tweets were deleted related to this scam. Furthermore, future work could also encompass studying the potential influence of Elon Musk's announcement to relax the moderation policies after he bought Twitter in October 2022.

Benchmarks Our analysis already revealed campaigns, scams, and conspiracy theories. Thus, our dataset is very well suited to be used as a benchmark dataset for algorithmic approaches that aim to detect such activities. Researchers could produce labels according to our findings or their analysis and train machine learning algorithms to identify this (problematic) content. Additionally, following the approach of (Pohl et al. 2022a), the identified campaigns can be used as blueprints to construct artificial campaign artifacts to challenge existing detection approaches in an adversarial manner.

Limitations

Use of Stream API Note that we reached the cap of the 1 % stream rate of the Twitter Stream API during our data collection multiple times. Although our dataset accurately represents trending topics during that time, we do not claim complete accuracy. Additionally, conclusions drawn from this data are only based on the (unknown) sampling strategy of Twitter which eventually provides the captured data stream. Thus, we suggest cross-checking and comparing the opinion climate represented in our dataset with the ones included in the related datasets (for reference, see Table 1) to get a complete picture of the actual mindset on Twitter during that time.

Original Tweets Further, we only collected original tweets and discarded duplicates like retweets and quotes. Although this makes the opinions represented in our dataset more diverse, no conclusions can be drawn on the existence

[5] https://edition.cnn.com/2022/04/26/tech/twitter-free-speech-elon-musk/index.html

of retweet networks or cascades, for example (Kessling and Grimme 2020). The dataset suits tasks focusing on the content level rather than analyzing the network perspective.

Content Bias Another decision we made during the data collection was to focus on English tweets and to exclude posts in other languages like Ukrainian or Russian. Although other platforms like Telegram or VKontakte are used more by Russians, at least some pro-Russian tweets (directed to the Russian population or the Russian-speaking Ukrainian population) are more likely to be tweeted in Russian. Thus, the dataset is probably biased toward an English-speaking international view of people active on a micro-blogging platform like Twitter.

Selected Keywords As stated in Section , we initially used a rather general list of hashtags to collect the data and later filtered using a narrower set of more important keywords. While this kept our collection process open to upcoming trends, it also allowed for a potential bias. On the one hand, there might still be some tweets in the dataset unrelated to the ongoing conflict, while on the other hand, we might have filtered some tweets that were related to the conflict but did not contain any of the most important keywords.

Evolution of Dataset Other limitations stem from data sharing limitations in Twitter's ToS. Although it is allowed to share tweet IDs, researchers cannot share the original posts, metadata, or user information. Other researchers can recreate (rehydrate) the entire dataset based on the IDs, provided no tweets have been deleted. However, this is highly unlikely as these datasets evolve through deleted tweets or blocked users (Zubiaga 2018). We refer the reader to Assenmacher et al. (2021) for more limitation details of sharing data collected from social media platforms.

Dataset Availability

The topics discussed in this dataset are potentially sensitive and may contain personal data. Because of this and Twitter's ToS, we do not release the entire dataset, including all metadata. Nevertheless, it is essential to support open science and to ensure that researchers (almost) anywhere in the world can access the content to conduct all kinds of research on global events.

To deal with these two conflicting goals, we publish two different datasets. First, we provide a publicly-available dataset adhering to Twitter's ToS containing the tweet's unique identifiers and the result of our rehydration efforts. This data includes all the ids of the initial 8.7 million tweets we collected. Every researcher with a valid Twitter research or developer account can retrieve the tweet objects of all tweets that are still available, i.e., that were neither deleted by Twitter nor by their original authors. This dataset can be found in a GESIS data archive. Here, due to the archive regulations, the data can be stored long-term so that it is available for researchers for a long time in the future:

https://doi.org/10.7802/2555

Second, we publish a restricted-access dataset containing the unavailable tweets without meta-data. Everyone can apply for access to this dataset, which will be granted if (and only if) they meet certain conditions. Most importantly, applicants must be affiliated with a (private or public) research agency and only use the dataset for noncommercial purposes. Further, we pseudonymize the tweets such that all user mentions are replaced with hash digests, and no direct connection can be made to any specific user. Access to the dataset can be requested here:

https://doi.org/10.5281/zenodo.7804846

When writing this paper, severe changes were made to the Twitter API, whose extent and magnitude need to be clarified. While we cannot predict the future implications of the changes to the Twitter API, we remain committed to ethical and responsible research practices. Nonetheless, the actual value of this dataset lies in its potential to facilitate cross-disciplinary research and collaboration. Thus, we invite any researchers to collaborate with us on this topic and use this dataset in a joint project. We hope our dataset will encourage others to build upon our work, foster a more open and collaborative research community, and contribute to the broader scholarly discourse.

Conclusion

This paper has presented a dataset collected before and during the invasion of Ukraine by the Russian army in early 2022. A unique feature of the dataset is that the data recording started one week before February 24, 2022 (the onset date of the invasion) and continued for a week afterward. The collected data originates from the Twitter Streaming API and thus includes content removed by users or Twitter afterward. In addition, the availability of all included tweets was checked and annotated at two points in time (early March 2022 and early January 2023). This information is helpful regarding potentially inappropriate content and Twitter's treatment of data in crises and afterward.

In addition to a comprehensive and up-to-date contextualization of the dataset within the ecosystem of topic-related and already available datasets, the paper also includes an analysis of the dataset presented here. This demonstrates that the dataset contains at least three topics of interest for research (crypto-scam, biolabs narrative, and a counter-campaign by Ukraine seeking international support). The observed artifacts also contain various traces of (apparently) semi-automated action. This demonstrates the potential and essential contribution of the dataset to the community and future research.

Ethical Impact

From a pure research perspective, this dataset contains valuable information on online communication patterns in a contemporary historical context. However, there has always been a conflict between full access to online content produced by individuals with individual interests in data and privacy protection on the one hand and the opportunity to get insights into the underlying mechanics of (dis-)information, campaigning, and manipulation content on social media platforms on the other hand. Consequently, we are trying to find an intermediate solution that respects the privacy of individuals while enabling researchers to investigate the dynamics and content of information warfare.

We follow the path of other research endeavors that share Twitter data in the context of abusive language detection (which is a similar sensitive topic) (Founta et al. 2018). While we publicly share the complete tweet ID list, we also provide access to the removed content upon explicit request. We only allow academics to access this content and ensure that they agree to our conditions, among other things, not redistributing the dataset and using it for scientific purposes only. We also exclude researchers connected to the immediate war parties, i.e., Russians, Ukrainians, or related organizations, for the time being. The accessible content is reduced to tweet text and stripped from user metadata. In addition, we replaced all user names mentioned in the tweets with pseudonyms and removed the associated tweet IDs to prevent the association of tweet content and specific users.

Further, users of the data set should consider the following ethical aspects:

- Researchers using this dataset and deriving conclusions from it are advised to reflect on the implications of their findings in the context of ongoing (dis-) information operations or campaigns. False or speculative conclusions may contribute to mis- or disinformation in the ongoing conflict and influence the public reception of the involved war parties.

- Researchers using the dataset on the content level should consider the implications of their analysis results on individuals or organizations implicitly mentioned in the textual data. Hyperlinks and indirect mentions of persons/organizations that we have not been able to anonymize in an automated manner may cause direct or indirect harm to the data subjects. Therefore, publishing hidden personal information and/or references to institutions should be treated very carefully.

- In the context of using single data artifacts (like using single tweets in showcases or as representative examples), researchers should consider that the individual "right to be forgotten" (GDPR) has to be respected. As there is no linkage between tweet IDs and textual content in our data set, data users should verify the open availability of specific content or rely on aggregated content presentation.

Although some ethical conflicts exist between possible harmful consequences due to the publication of tweet texts and the provision of research data, we have decided to publish the data. We consider these datasets (as all of the other datasets mentioned above) essential and of public interest in the context of the historical event of the Russian invasion of Ukraine and the subsequent war. To mitigate any potential adverse effects of the publication, we have taken various measures to prevent data de-anonymization. At the same time, we ensure that the sensitive text data cannot be accessed uncontrollably but are still available to researchers.

Acknowledgments

The authors acknowledge support from the BMBF-funded project HybriD (funding ref. 16KIS1531K) and the European Research Center for Information Systems (ERCIS).

References

Assenmacher, D.; Weber, D.; Preuss, M.; Calero, V. A.; Bradshaw, A.; Ross, B.; Cresci, S.; Trautmann, H.; Neumann, F.; and Grimme, C. 2021. Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review*, online first.

Bruns, A. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11): 1544–1566.

BwandoWando. 2023a. Reddit r/Ukraine Dataset (53K thrds, 1.76M cmnts). https://www.kaggle.com/datasets/bwandowando/ukrainesubredditthreadsandcomments. Accessed: 2023-04-17.

BwandoWando. 2023b. Ukraine Conflict Twitter Dataset. https://www.kaggle.com/dsv/4809572. Accessed: 2023-04-17.

Caprolu, M.; Sadighian, A.; and Di Pietro, R. 2022. Characterizing the 2022 Russo-Ukrainian Conflict Through the Lenses of Aspect-Based Sentiment Analysis: Dataset, Methodology, and Preliminary Findings. *ArXiv*, arXiv:2208.04903.

Chakravarty, P. 2022. The war in Ukraine and its economic fallout. https://policycommons.net/artifacts/2288078/the-war-in-ukraine-and-its-economic-fallout/3048212/. Accessed: 2023-04-17.

Chen, B.; Wang, X.; Zhang, W.; Chen, T.; Sun, C.; Wang, Z.; and Wang, F.-Y. 2022. Public Opinion Dynamics in Cyberspace on Russia–Ukraine War: A Case Analysis With Chinese Weibo. *IEEE Transactions on Computational Social Systems*, 9(3): 948–958.

Chen, E.; and Ferrara, E. 2022. Tweets in Time of Conflict: A Public Dataset Tracking the Twitter Discourse on the War Between Ukraine and Russia. *ArXiv*, abs2203.07488.

Chowdhury, F. A.; Allen, L.; Yousuf, M.; and Mueen, A. 2020. *On Twitter Purge: A Retrospective Analysis of Suspended Users*, 371–378. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370240.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Fung, Y. R.; and Ji, H. 2023. A Weibo Dataset for the 2022 Russo-Ukrainian Crisis. *ArXiv*, arXiv:2203.05967.

Gosh, S. 2022. UKRUWAR22: A Collection of Ukraine-Russia War Related Tweets. https://ieee-dataport.org/documents/ukruwar22-collection-ukraine-russia-war-related-tweets#files. Accessed: 2023-04-17.

Grimme, C.; Preuss, M.; Adam, L.; and Trautmann, H. 2017. Social Bots: Human-Like by Means of Human Control? *Big Data*, 5(4): 279–293.

Hanley, H. W. A.; Kumar, D.; and Durumeric, Z. 2023. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. In *Proceedings of the 17th International Conference on Web and Social Media (Publication Status: Accepted)*. AAAI.

Haq, E.; Tyson, G.; Lee, L.-H.; Braud, T.; and Hui, P. 2022. Twitter Dataset for 2022 Russo-Ukrainian Crisis. *ArXiv*, abs/2203.02955.

Keller, F. B.; Schoch, D.; Stier, S.; and Yang, J. 2020. Political Astroturfing on Twitter: How to Coordinate a Disinformation Campaign. *Political Communication*, 37(2): 256–280.

Kessling, P.; and Grimme, C. 2020. Analysis of Account Engagement in Onsetting Twitter Message Cascades. In Grimme, C.; Preuss, M.; Takes, F. W.; and Waldherr, A., eds., *Disinformation in Open Online Media*, 115–126. Cham: Springer International Publishing.

Majó-Vázquez, S.; Congosto, M.; Nicholls, T.; and Nielsen, R. K. 2021. The Role of Suspended Accounts in Political Discussion on Social Media: Analysis of the 2017 French, UK and German Elections. *Social Media + Society*, 7(3): 20563051211027202.

Maschmeyer, L. 2021. Digital Disinformation: Evidence from Ukraine. *CSS Analyses in Security Policy*, 278.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, arXiv:1802.03426.

Metaxas, P. T.; and Mustafaraj, E. 2012. Social Media and the Elections. *Science*, 338(6106): 472–473.

Mukherjee, S. 2022. Ukraine War Tweets. https://www.kaggle.com/datasets/shyambhu/ukraine-war-tweets. Accessed: 2023-04-17.

Münch, F. V.; and Kessling, P. 2022. ukraine_twitter_data. https://doi.org/10.17605/OSF.IO/RTQXN.

Park, C. Y.; Mendelsohn, J.; Field, A.; and Tsvetkov, Y. 2022. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. *ArXiv*, arXiv:2205.12382.

Pierri, F.; Luceri, L.; Jindal, N.; and Ferrara, E. 2022. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. *ArXiv*, arXiv:2212.00419.

Pohl, J.; Assenmacher, D.; Seiler, M.; Trautmann, H.; and Grimme, C. 2022a. Artificial Social Media Campaign Creation for Benchmarking and Challenging Detection Approaches. In *Workshop Proceedings of the 16th International AAAI Conference on Web and Social Media*, ICWSM-16. Atlanta, GA, USA: AAAI.

Pohl, J.; Seiler, M. V.; Assenmacher, D.; and Grimme, C. 2022b. A Twitter Streaming Dataset collected before and after the Onset of the War between Russia and Ukraine in 2022. *Social Science Research Network (SSRN)*.

Preda, G. 2022. Slava Ukraini Tweets. https://www.kaggle.com/datasets/gpreda/slava-ukraini-tweets. Accessed: 2023-04-17.

Prier, J. 2017. Commanding the Trend: Social Media as Information Warfare. *Strategic Studies Quarterly*, 11(4): 50–85.

Purtova, D. 2022. Russia-Ukraine war - Tweets Dataset (65 days). https://www.kaggle.com/datasets/foklacu/ukraine-war-tweets-dataset-65-days?resource=download. Accessed: 2023-04-17.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shevtsov, A.; Tzagkarakis, C.; Antonakaki, D.; Pratikakis, P.; and Ioannidis, S. 2022. Twitter Dataset on the Russo-Ukrainian War. *ArXiv*, arXiv:2204.08530.

Smart, B.; Watt, J.; Benedetti, S.; Mitchell, L.; and Roughan, M. 2022. #IStandWithPutin Versus #IStandWithUkraine: The Interaction of Bots and Humans in Discussion of the Russia/Ukraine War. In Hopfgartner, F.; Jaidka, K.; Mayr, P.; Jose, J.; and Breitsohl, J., eds., *Social Informatics*, 34–53. Cham: Springer International Publishing.

Soares, F. B.; Saiphoo, A.; Gruzd, A.; and Mai, P. 2022. Navigating the fog of war during the Russia's invasion of Ukraine: An exploratory network analysis of tweets about an alleged chemical attack in Mariupol. https://socialmedialab.ca/2022/05/10/navigating-the-fog-of-war-during-the-russia-invasion-of-ukraine/. Accessed: 2023-04-17.

Stupples, D. 2015. The next war will be an information war, and we're not ready for it. https://theconversation.com/the-next-war-will-be-an-information-war-and-were-not-ready-for-it-51218. Accessed: 2023-04-17.

Thapa, S.; Shah, A.; Jafri, F.; Naseem, U.; and Razzak, I. 2022. A Multi-Modal Dataset for Hate Speech Detection on Social Media: Case-study of Russia-Ukraine Conflict. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, 1–6. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Toraman, C.; Ozcelik, O.; Şahinuç, F.; and Can, F. 2022. Not Good Times for Lies: Misinformation Detection on the Russia-Ukraine War, COVID-19, and Refugees. *ArXiv*, arXiv:2212.00419.

Zhu, Y.; Haq, E.-u.; Lee, L.-H.; Tyson, G.; and Hui, P. 2022. A Reddit Dataset for the Russo-Ukrainian Conflict in 2022. *ArXiv*, arXiv:2206.05107.

Zubiaga, A. 2018. A Longitudinal Assessment of the Persistence of Twitter Datasets. *Journal of the Association for Information Science and Technology*, 69(8): 974–984.