Disinformation Analysis on Telegram: A Metadata-Centered, Privacy-Aware Dataset

Jeong-Eun Choi
Fraunhofer SIT, ATHENE
Darmstadt, Germany
jeong-eun.choi@sit.fraunhofer.de

Karla Schäfer
Fraunhofer SIT, ATHENE
Darmstadt, Germany
karla.schaefer@sit.fraunhofer.de

York Yannikos
Fraunhofer SIT, ATHENE
Darmstadt, Germany
york.yannikos@sit.fraunhofer.de

Martin Steinebach
Fraunhofer SIT, ATHENE
Darmstadt, Germany
martin.steinebach@sit.fraunhofer.de

***Abstract*—Disinformation remains one of the major challenges in today's digital landscape. Although numerous detection approaches have been proposed, the vast volume and dynamic nature of social media data, combined with privacy and legal constraints, continue to hinder research and mitigation efforts. In this work, we present a dataset of 977 publicly accessible Telegram channels and groups relevant to disinformation in Germany, collected over more than one year, encompassing significant events such as the war in Ukraine and four state elections in Germany. To address privacy and legal concerns, we apply cleaning and pseudonymization processes while preserving structural and behavioral details. Our dataset enables metadata-based rather than content-based analysis, allowing efficient, low-resource exploration of dissemination patterns without requiring deep technical expertise or platform-specific knowledge. The dataset is designed to support broader network analysis of disinformation dynamics within Telegram's unique communication ecosystem while safeguarding privacy.**

***Index Terms*—Telegram, Disinformation, Dataset, Privacy**

I. Introduction

Recent studies highlight the significant role of Telegram in the spread of disinformation (false information spread with malicious intention), particularly in Germany. A study by the German Bertelsmann Foundation found that 19% of users reported encountering disinformation via messenger apps in general. During the COVID-19 pandemic, Telegram was specifically identified as a fertile ground for disinformation, with nearly one in four users reporting that they encountered disinformation on the platform either "very often" or "rather often" [1]. In a related report by the Alfred Landecker Foundation, Telegram is seen as the most important platform for conspiracy ideologies and right-wing extremism in Germany.[1]

Despite the evolving landscape of social media and the emergence of new platforms such as Bluesky, Signal, and Session, Telegram has become a significant platform in the context of disinformation and other illicit activities, including the sale of illegal goods. Its appeal lies in the combination of encryption, minimal moderation, and user anonymity, which attracts both, legitimate users and malicious actors. In addition to hosting disinformation, Telegram facilitates information exchange, for example through the dissemination of links to other platforms. This role is particularly significant given its comparatively smaller user base than other major platforms.

Nevertheless, despite Telegram's growing importance in the dissemination of information and disinformation, there remains a notable lack of publicly available datasets to support research on this platform, particularly in the context of disinformation.

We introduce our dataset as a foundation for metadata-based analysis of disinformation on Telegram. It includes a compact, but meaningful set of metadata metrics selected to capture dissemination patterns and behavioral signals. To protect privacy and reduce retraceability, we pseudonymized all channel/group names and IDs, and excluded full content. Instead, we offer high-level summaries such as automatically extracted keywords and sentiment, providing insight into the nature and tone of discourse without revealing privacy problematic content. The aim of this paper is to describe our dataset, present preliminary insights, and propose potential directions for future analysis.

Our contributions are threefold:

- Unlike existing Telegram datasets that focus solely on channels, our dataset is constructed using a careful selection of *channels and groups*, identified in collaboration with experts as relevant to disinformation dissemination in Germany.
- We apply *ethical safeguards*, including the pseudonymization of actors' (group/channel) identifiers and exclusion of content. This ensures stronger privacy protection while remaining useful for disinformation research.
- We provide *a curated set of raw and derived metrics* that balances analytical depth with data minimization to ensure usability at large scale.

This research work was supported by ATHENE in the project CRISIS as well as within the project DYNAMO.

[1] Link: Cemas

II. Background and Related Work

A. Disinformation & Social Media

Although disinformation on social media has been widely studied, findings often do not generalize across platforms and are fragmented across disciplines [2]. The creation, use, and spread of information are conditioned by the digital ecosystem, including platform design and how users perceive and use each platform, as well as cultural, geographical [2] and social context. As a result, comparative, platform-specific analyses are essential. Technical factors also matter as content creation is increasingly shaped by generative deep learning models, and content analysis relies on advanced methods such as multimodal analysis like speech-to-text, and image-to-text.

Another key aspect is that disinformation is not only about verifying whether a piece of information is true or authentic. It is also about how users employ that information and for what purpose, in other words, their intention. Many studies focus on verifying authenticity or facts, as this can provide clues to the possible intention of the author. For instance, if someone uses a picture as evidence but removes it from its original context, such behavior indicates a hidden motivation. Therefore, discussions on disinformation often concentrate on authenticity verification or fact-checking, leading to a strong focus on content-based analysis.

Given the scale of social media data, however, it is impractical to evaluate the credibility of each individual item solely through content-based approaches. We therefore suggest shifting attention to the holistic behavior of information exchange, focusing on actors [3] rather than isolated information items that are tied to a specific topic or language [2], [3]. This perspective can provide broader insights into the dissemination of disinformation and enable the development of more lightweight yet generalizable mitigation strategies.

To this respect, there is a lack of real data, that depicts a subset of disinformation landscape. There are some publicly available datasets (such as BuzzFeed Fake News Corpus [4]). However, these datasets primarily focus on detecting individual fake news articles or items, which reduces their usefulness for studying broader patterns of information dissemination and the role of different actors in the disinformation landscape.

DeFaktS [5] is a German-language dataset collected from X (formerly Twitter), containing 105,855 posts. It includes basic metadata such as likes, replies, and retweet counts. Like many other datasets, however, it struggles to trace connections between accounts because the collection process was post-driven rather than actor-driven. Therefore, although the dataset offers valuable annotations and labels, it provides limited insight into the dissemination patterns of actors and posts relevant to understanding disinformation.

Zhou and Zafrani [6] propose using network information to detect fake news by analyzing how news spreads across platforms. Their work focuses on identifying patterns in propagation, the users who spread the content, and the relationships among them, using data from X. We argue that such metadata-driven network analysis should also be applied to disinformation research on social media, not only to track the dissemination of disinformation but also to understand the dynamics of information exchange among actors within the disinformation landscape, i.e. promoting stronger contextual and longitudinal understanding [7].

B. Telegram

In 2024, Telegram adjusted its privacy policy to comply with law enforcement, and promised to disclose user phone numbers and IP addresses upon receipt of a valid court order[2]. However, in terms of disinformation, it is very unlikely that such a valid court order would be issued, as the dissemination of misleading content often does not always meet the legal threshold required for law enforcement intervention.

While Telegram offers an official API, its use in academic research has been limited. One possible reason is the comparatively smaller user base relative to mainstream platforms such as Facebook, Twitter (X), or YouTube. However, during the course of our research, we identified several additional factors that may contribute to this gap. These include the decentralized and semi-anonymous structure of Telegram, inconsistent or opaque definitions of platform metrics, and significant ethical and legal concerns around data collection, particularly when dealing with private or sensitive content.

Although the Telegram API provides access to posts and metadata, key metrics are often ambiguously defined or inconsistently implemented [8]. A possible solution is to simulate interactions manually to understand how the metrics are generated and what they actually represent. For example, the *forward count* field of a post does not indicate the total number of forwards, but rather the number of unique users who forwarded this specific post. Furthermore, original posts in groups (posts that were first posted in groups) always return a forward count of zero, whereas the same is not true for channels. In Telegram, channels are used for broadcasting, while in groups users can interact freely. Such small yet significant details create a barrier for initial exploration of Telegram, often due to a lack of either technical skills or contextual understanding.

Telegram's decentralized and semi-anonymous structure makes identifying relevant actors or communities challenging. Most existing datasets rely on seed lists and snowball sampling, tracing forwarded content back to its sources. For example, Baumgartner et al. [9] focused on right-wing extremism and cryptocurrency channels, while Gangopadhyay et al. [10] used TGStat[3] to collect the top 100 channels by various criteria and then apply snowball sampling. However, to the best of our knowledge, no existing dataset has systematically identified disinformation actors that are both active and relevant in Germany, and whose selection has been validated by journalism or fact-checking experts.

Accessing Telegram data often requires joining channels or groups, which raises ethical and legal concerns. Private

[2] Link: IBM

[3] Link: TGstat

or invitation-only groups are especially sensitive, as scraping content may violate privacy rights or platform terms. For datasets addressing sensitive topics like disinformation, public release can provoke backlash or legal challenges, especially if involved users feel targeted or exposed, even for public groups and channels. Therefore, we propose minimizing content data collection to preserve privacy while focusing on metadata-driven network analysis. This approach allows researchers to study actors and disinformation networks without compromising privacy.

III. Dataset

We present a dataset comprising 977 Telegram actors with data collected between 25 March 2022 and 30 June 2023. With this, the time span includes significant events such as the war in the Ukraine and four state elections in Germany (Niedersachsen, Nordrhein-Westfalen, Saarland and Schleswig-Holstein). Moreover, within the designated data collection period, the 9€ ticket was introduced in Germany, constituting a limited-time promotional offer on local public transportation. Concurrently, Donald Trump declared his intention to seek re-election for the presidency of the United States.

The initial 770 public Telegram actors were identified through 15 expert interviews with journalists, fact-checkers, and representatives from research and security authorities dealing with disinformation. The interviews, conducted between February and May 2022, served to map key actors involved in the spread of disinformation in Germany. These interviews were organized by our project partner, Hochschule der Medien Stuttgart, as part of the DYNAMO project[4] [11]. Further details pertaining to the interviews can be found in [12].

As some actors were found to be inactive during the period under review, an additional 207 actors were added using the snowball sampling method. Moreover, as previously noted, the forwards count of original posts (*forward_count_original*) is unavailable for all group posts and also missing in five channels, likely due to channel-specific privacy settings. More broadly, not all metadata fields are uniformly available across actors, as platform settings influence data visibility. Therefore, during our crawling process, channels or groups that did not allow crawling sufficient metadata were excluded from the final dataset. Finally, our dataset consisting of 676 channels and 301 groups was collected via the Telegram API[5].

Regarding the longitudinal aspect, we crawled data every day. Meaning that after a message is crawled, its information is not updated in the subsequent crawls. This approach is based on the fact that all messages available at the time of crawling can be collected unless they were deleted beforehand. To perfectly capture dissemination behavior one would need to track updates to each message daily, which was not feasible due to resource constraints. Instead, we chose to aggregate new posts over each day, providing a daily snapshot of dissemination patterns. Consequently, this dataset supports comparative analysis on a day-by-day basis, except for the very first crawl, which would have the latest 1000 posts. The first day of crawling varies for each channel/group, as they have been subsequently added at the beginning of the interviews or through snowball sampling.

Our dataset includes both, public channels (one-to-many) and groups (many-to-many), enabling analysis of different types of community interaction and information dissemination on Telegram which is often neglected in other existing Telegram datasets. For ensuring privacy, we pseudonymized channel and group IDs, as well as user IDs for each post. More details are described in Section III-A.

Compared to other existing datasets, our dataset is smaller than others (ex. 27,801 [9] and 71,048 [10]), but it spans a longer observation period (compared to snapshot [9] or 9-month [10]). While the latter [10] includes multilingual content, neither dataset was specifically designed for disinformation research.

A. Dataset Structure

The Telegram API provides two main types of data: information about actors and information about their individual posts. Accordingly, our dataset is organized into two main components, referred to as *collections*.

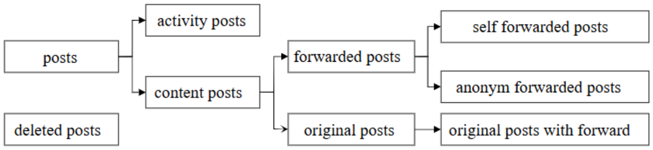


Fig. 1: Post Types

1) $Collection_{actors}$*:* The first collection contains aggregated metadata for all observed actors. Key metrics included in this collection are listed in Table I, which also provides a general statistic overview of our dataset for each metric. The different types of posts are visualized in Figure 1. In Table I, to represent the distribution of post types more meaningfully, we use percentages relative to POST_IDS_FOUND, rather than raw counts. In our dataset, however, we provide the raw values. The annotations in brackets indicate the data type of each metric in our dataset.

2) $Collection_{posts}$*:* Our second collection, consists of a separate file for each actor, containing metadata for every post posted. Each post is associated with a unique ID that is only unique within the context of its respective actor. See Table II for a detailed description of the metadata metrics and structure. In addition to metadata, we provide content features derived from the message itself, including text length, keywords extracted using YAKE [13], and sentiment analysis results using the hugging face model of tabularisai[6].

We chose YAKE for keyword extraction primarily due to its speed and effectiveness on short texts, which are typical

[4] Link: DYNAMO Project

[5] Link: Telegram

[6] Link:tabularisai

TABLE I: Overview of $Collection_{actors}$ - Although the dataset contains raw counts, we present percentages for some of the metrics in this table to enhance interpretability and comprehensibility.

Metric	Description	Statistic Overview
ID [INT]	Randomly generated unique ID for each actor	$total\_count = 977$
GROUP [BOOL]	True if actor is a group	$total\_group = 301$ ($total\_channel = 676$)
ACTIVE [BOOL]	True if active within 3 months (so, before 31 Mar 2023)	$total\_active = 767$ ($total\_deactive = 210$)
DAYS_ONLINE [INT]	Days online (based on earliest post posted)	$mean = 799.31$, $std = 352.43$, $min = 212$, $max = 2753$
MAX_PARTICIPANTS [INT]	Max participants reached	$mean = 14168.11$, $std = 31678.1$, $min = 1$, $max = 304680$
MIN_PARTICIPANTS [INT]	Min participants reached	$mean = 10685.13$, $std = 24893.62$, $min = 0$, $max = 245551$
MEAN_PARTICIPANTS [FLOAT]	Average participants	$mean = 12600.13$, $std = 28605.16$, $min = 1$, $max = 268090$
LATEST_PARTICIPANTS [INT]	Most recent participant count	$mean = 11713.87$, $std = 27072.81$, $min = 0$, $max = 245551$
POST_IDS_FOUND [INT]	Count of post IDs unique per actor	$mean = 8195.34$, $std = 31516.59$, $min = 2$, $max = 546171$
MISSING_POST_IDS [INT]	Missing post IDs detected from gaps	$mean = 3317.19$, $std = 42876.58$, $min = 0$, $max = 1318516$
REAL_AVG_POSTS_PER_DAY [FLOAT]	Days online (based on earliest post posted)	$mean = 799.31$, $std = 352.43$, $min = 212$, $max = 2753$
%CONTENT_POST [INT]	Posts with content	$mean = 95.07$, $std = 11.91$, $min = 1$, $max = 100$
%CONTENT_POST_W_TEXT [INT]	Content posts that has text data	$mean = 79.5$, $std = 17.82$, $min = 0$, $max = 100$
%ORIGINAL_POST [INT]	Posts posted originally by the current actor or within the current actor (channel/group)	$mean = 58.92$, $std = 28.56$, $min = 0$, $max = 100$
%FORWARDED_POSTS [INT]	Posts forwarded (either from other actors or reposting its own)	$mean = 36.08$, $std = 28.14$, $min = 0$, $max = 99$
%SELF_FORWARD [INT]	Original posts of the current actor that has been reposted	$mean = 0.56$, $std = 2.22$, $min = 0$, $max = 33$
MEAN_FORWARD_COUNT_ORIGINAL [FLOAT]	Forwarded counts of the original posts of the current actor	$mean = 116.70$, $std = 280.64$, $min = 8.70$, $max = 5792.31$
%ANONYMIZED_ORI_AUTHOR [INT]	Posts where the original author (user_id) is known	$mean = 79.51$, $std = 31.67$, $min = 0$, $max = 100$
%ANONYMIZED_ORI_CHANNEL [INT]	Posts where the original channel/group (channel or group_id) is known (i.e. where it was first posted); for original posts this is always known	$mean = 4.64$, $std = 7.53$, $min = 0$, $max = 76$
%ANONYMIZED_SPREADER [INT]	Posts where the original poster (user_id) is known; for original posts, the spreader and author are the same	$mean = 70.33$, $std = 45.24$, $min = 0$, $max = 100$

in social media platforms like Telegram. Rather than using supervised extraction methods, we decide for YAKE, as YAKE depends on the statistical features, thus is more dependent on the real raw input rather than trained word embeddings or

TABLE II: Overview of $Collection_{posts}$

Metric	Description
ID [INT]	Post ID unique within the file
DATE_ACCESSED [DATE]	Date of crawling
DATE_POSTED [DATE]	Date when the post was posted
FORWARD_COUNT [INT]	Forward count of the post (unique number of users forwarding the post)
ORIGINAL_CONTENT_POST [BOOL]	True if original content post else forwarded content post (Empty i.e. Nan if activity post)
CONTENT_POST [BOOL]	True if content post else activity post
ORIGINAL_AUTHOR [INT]	Pseudonymized user ID of the original author
ORIGINAL_CHANNEL [INT]	Pseudonymized channel ID of the original channel
ORI_POST_ID [INT]	Original post ID from the original channel (exisists if it is a forwarded post)
TEXT_LENGTHS [INT]	Text length
KEYWORD_EXT [DICT [STR]]	Keywords extracted using KeyBERT, top three keywords
SENTIMENT [STR]	Sentiment classification using multilingual five class sentiment analysis model

semantic understanding capabilities of deep learning models.

Similarly, sentiment analysis with the model of tabularisai was included due to its computational efficiency and its suitability for analysing German-language content in social media contexts. While raw content undeniably offers significant analytical value, large-scale evaluation of ambiguous social media data poses scalability challenges. Our approach therefore emphasizes metadata-driven analysis complemented by lightweight, content-aware tools to enable efficient and context-sensitive exploration. For more qualitative analysis of the raw content, refer to the work [14] of our project partners.

We believe that, particularly for understanding broad behavioral trends and ecosystem-level dynamics, hybrid strategies are essential. For example, in our prior work [15], we applied regular expressions to detect posts containing scientific terminology. This allowed us to identify rather insignificant actors (ex. smaller number of participants) with limited reach or dissemination power, producing scientific content. Despite this, their posts were disseminated by larger actors which amplified their posts. This illustrates how combining metadata insights with targeted scalable content analysis can uncover hidden dissemination patterns and actor roles within platforms. Therefore, in order to address the tension between privacy and information, we decided to utilize lightweight information extraction methods to facilitate a general overview of the content.

IV. First Insights

The following analysis and results offer preliminary insights into the dataset. Concurrently, it presents possible ways of analyzing and extracting interesting data points of the dataset. The following results and analyses focus on identifying and examining significant actors.

A. About Metrics

A proper understanding of the metrics yields interesting insights. Below, we summarize some key findings from our data, alongside references to our previous analyses, especially relevant for dissemination of information or behavioral aspects of actors.

- MISSING_POST_IDS: Some actors exhibit a higher number of missing, i.e. deleted posts, while others have never deleted a post. This variation may indicate differences in content moderation practices. Comparing these subgroups can reveal which types of actors tend to remove content, hence, remove potentially disinformation.
- REAL_AVG_POSTS_PER_DAY: Certain actors post extremely frequently up to 4,046 posts per day. While high activity in groups can be explained by many participants posting, some channels (i.e. broadcasting) also exhibit very high average daily posts, possibly indicating management by multiple administrators.
- ORIGINAL_POST: Actors who mostly publish original posts can be regarded as content creators within the dataset, whereas those whose posts largely rely on reposting from others are better classified as spreaders.
- ANONYMIZATION: Telegram allows users to hide the identity of original authors, complicating user-level tracking. Consequently, analysis primarily relies on actor-level or post-level metrics. However, post dissemination paths cannot be fully reconstructed as only the original source is traceable. Therefore, large datasets and temporal ordering are necessary to approximate dissemination dynamics.
- SELF_FORWARD: Some actors frequently repost their own messages, which may indicate active efforts to amplify their content.
- FORWARD_COUNT_ORIGINAL: The forwarding count of an actor's original posts can serve as a proxy for their dissemination power within the network.
- USER_ID & ACTOR_ID: We identified 21,108 unique user_ids in the dataset. Although user identities are generally anonymized, overlaps in user_ids across different actors could offer valuable insights into user behavior and cross-actors dynamics. Additionally, we observed instances where posts originated from actors are not included in our dataset. These are marked with actor_id = 0, indicating the original source was outside our tracked ecosystem (our set of actors).

B. About Longitudinal Analysis

As mentioned earlier, the first crawl collected the latest 1,000 posts, while subsequent crawls added only new posts per day. For longitudinal day-by-day analysis, it is therefore important to focus on posts from the second crawl onwards. At the same time, the historical data from the initial crawl makes it possible to identify actors whose original posts received more forwards in the daily crawls than in the first crawl, which we refer to as GROWING_ACTORS. This suggests that their activity and dissemination strength increased during the observed period, since daily crawls would normally show lower forward counts because posts have been online for less than 24 hours. It is notable that six actors experienced a surge in forwards to their original posts, with a percentage increase exceeding 100% (see Table III).

TABLE III: Only for Active Actors (Total of 767): Daily Forwards Counts of Original Posts vs. Initial Crawl

Feature	Count
# of actors with New Original Posts	503
MEAN % Difference in Forward Counts of Original Posts	-38.58 (↓)
MIN % Difference in Forward Counts of Original Posts	-100.0 (↓)
MAX % Difference in Forward Counts of Original Posts	3392.25 (↑)
STD % Difference in Forward Counts of Original Posts	157.91
# of actors with Increasing Forward Counts of Original Posts (MEAN % Difference in Forward Counts of Original Posts is positive ↑)[GROWING_ACTORS]	22

As comparison, we identified the top 22 actors with the highest forward counts of original posts, which we refer to as STRONG_ACTORS. We found out that the mean forward counts for the GROWING_ACTORS was 52.83 while for the STRONG_ACTORS it was 496.29. There was only one actor that was in both groups, indicating that this channel was or became a strong actor in general that has increased in its dissemination power during the period of observation. This actor, with pseudonymized ID 6998685360, has an average forward of original posts (daily) of 728.0 with the maximum increase in forward counts compared to the initial crawl. However, this actor has posted a single original post (being an audio file) during the period of observation, thus suggesting it to be an outlier, while analyzing the audio file would still be of importance, later discussed in Section V.

TABLE IV: STRONG_ACTORS vs. GROWING_ACTORS

Feature	Count
Latest Participants vs. Min Participants	
Average % Difference in Participants for STRONG_ACTORS	17.48 (↑)
Average % Difference in Participants for GROWING_ACTORS	656.40 (↑)
% Original Posts (of All Content Posts)	
Average % for STRONG_ACTORS	82.68
Average % for GROWING_ACTORS	67.00
Average Content Post per Day	
Average % for STRONG_ACTORS	6.48
Average % for GROWING_ACTORS	5.35
Average Days Online	
Average % for STRONG_ACTORS	844.36
Average % for GROWING_ACTORS	1209.09

Comparing the number of participants between the two groups STRONG_ACTORS and GROWING_ACTORS, see Table IV, shows that STRONG_ACTORS has higher number of latest participants while GROWING_ACTORS has a higher increase in participants. Moreover, the STRONG_ACTORS tend to have a higher percentage of original posts and also a higher number of content posts posted per day. Surprisingly, the mean days online were much lower for STRONG_ACTORS (see Table VII).

TABLE V: Qualitative Observation for Actors with Neutral or Positive Tones

Psedo_id	Top Sentiments	Topics
STRONG_ACTORS		
3936952829	Very Positive	William Toel
8081868486	Neutral	Politics (USA, Germany), Covid-19
GROWING_ACTORS		
1103442536	Very Positive	Elysion, Russia, Chlidren
1603243930	Neutral	Scheduling for Demonstration
2516624129	Neutral	German Politics

Using the sentiment labels, we were able to observe that both actors predominantly disseminate content posts with ‘Very Negative’ sentiment. Among STRONG_ACTORS, the actor with the ID 3936952829 had most posts that were ‘Very Positive’ (56.96%) and 8081868486 had most posts that were ‘Neutral’ (42.22%). Among GROWING_ACTORS 1103442536 had most posts that were ‘Very Positive’ (49.90%) and 1603243930 and 2516624129 had most posts that were ‘Neutral’ (77.73% and 41.11% respectively). The qualitative observation using the keywords are presented in Table V.

C. *Influential Actors*

Finally, we counted all posts that were included in the whole dataset, where the posts originated from the actors of GROWING_ACTORS or STRONG_ACTORS. See Table VI for the results. For STRONG_ACTORS the mean of the total posts found was 99,764.36.

Furthermore, we define the *influencial_actors* as an actor whose original post is most frequently found in other actors (groups and channels). We found that from the STRONG_ACTORS, 5 actors belonged to the top 50 *influencial_actors*.

TABLE VI: Count of Messages Originated from different Actors (within Active Actors)

Actor Type	Mean	Std	Min	Max
STRONG	99,764.36	108,037.29	338	377,312
GROWING	4,685.91	12,526.08	0	49,511
ALL	67,755.88	345,983.42	0	5,998,356
INFLUENTIAL	1,406,988.82	1,499,859.26	334,750	5,998,356

Furthermore, we analyzed the top 22 most INFLUENCIAL_ACTORS in comparison with STRONG_ACTORS and GROWING_ACTORS (see Table VI and Table VII). One of the most interesting finding is that while all actors of STRONG_ACTORS and GROWING_ACTORS were channels, 18 of the 22 INFLUENCIAL_ACTORS were groups. As described, in groups it is not possible to extract the forwards count of the original posts directly from the original author. Therefore, in order to observe the forwards of the original posts of groups, the only possibility is to find the posts (ex. using post IDs and actor ID) within the selected dataset.

The INFLUENCIAL_ACTORS do not have larger number of participants, higher percentage of original posts or are longer online, compared to the STRONG_ACTORS. However, we can observe that they have much higher number of mean posts

TABLE VII: [INFLUENCIAL_ACTORS vs. STRONG_ACTORS vs. GROWING_ACTORS within active actors] vs. INTERACTIVE_ACTORS

Actor Type	#groups / #channels	Mean # latest participants	Mean % original posts	Mean Posts per Day	Mean Days Online
INFLUENCIAL_ACTORS	18 / 4	23,894.32	77.16	315.53	523.5
STRONG_ACTORS	0 / 22	103,762.22	82.68	6.48	844.36
GROWING_ACTORS	0 / 22	9,535.91	67.00	5.35	1,209.09
INTERACTIVE_ACTORS	17 / 5	3,217.54	32.21	178.60	492.5

per day (Table VII). This is very likely due to the fact that groups are designed for a many-to-many interaction so that all participants are involved in producing the content. This is probably also the reason why users in these groups are more engaged to spread the posts.

D. Actors Interaction

TABLE VIII: [INFLUENTIAL_ACTORS vs. STRONG_ACTORS vs. GROWING_ACTORS within active actors] vs. INTERACTIVE_ACTORS: Mean Count of Influenced Actors

INFLUENTIAL	STRONG	GROWING	Overall	INTERACTIVE
37	25	6	88	239

With our dataset, it is also possible to derive interaction behaviors between actors, by observing how many original posts of an actor has been posted in other actors. In other words, while the previous section defined *influence* in terms of posts spread, we observe in terms of actors (referred as INTERACTIVE_ACTORS). Similar to the previous analysis, we compare STRONG_ACTORS, GROWING_ACTORS, INFLUENTIAL_ACTORS in terms of how many actors were influenced by the original posts of these actors, see Table VIII.

Among the top 22 of INTERACTIVE_ACTORS, 4 of the INFLUENCIAL_ACTORS are included in the list of INTERACTIVE_ACTORS as well. We observed that in terms of count of messages originated from actors, INTERACTIVE_ACTORS had more of their original posts (341,421.36, see Table VI for comparison) posted within our dataset, while having the lowest mean days online (492.5), lower mean latest participants (3,217.54), lower mean % of original posts (32.21) but relative higher mean posts per day (178.60). We can presume that within this set of actors, there are actors that are actively trying to interact and spread their original posts. This could also be an indication of actors aiming to increase their influence. Similar to INFLUENCIAL_ACTORS, the INTERACTIVE_ACTORS are consisting of more groups than channels (17 groups and 5 channels), see Table VII.

E. Post Dissemination

As an example for identifying post dissemination, we observed the actor with ID 6667842874 from STRONG_ACTORS which had the highest amount of original posts (total of 319,264) found within the dataset. We first identified the top 3 posts (post IDs: 40332, 67217, 63031) that were most frequently found (66, 40, 35 - respectively) within the dataset and tried to reconstruct the spread of messages in terms of time posted. The results are visualized in Fig. 2.

As demonstrated in Fig. 2, the majority of posts are forwarded during the initial days following their publication. The red lines in the figure indicate the earliest time at which the post appears in our dataset.

The post with ID 40332 was not included among the posts crawled for the original actor (ID 6667842474), so the red line represents the earliest occurrence found in other actors. Interestingly, even if a post is deleted in the original channel or group, its content can still be retrieved if it has been forwarded to other actors. This post was forwarded for six days. The content of this post is promoting the forwarding and spreading of a channel, and has a very positive sentiment label.

The post with ID 67217 was included in the original actor's posts, so the red line represents the time of the original post. This post was forwarded for three days. It is about nuclear power, and has a very negative sentiment label.

The post with ID 63031 was forwarded for a period of seven days. It mentions vaccines and has a very positive sentiment label.

V. Limitations

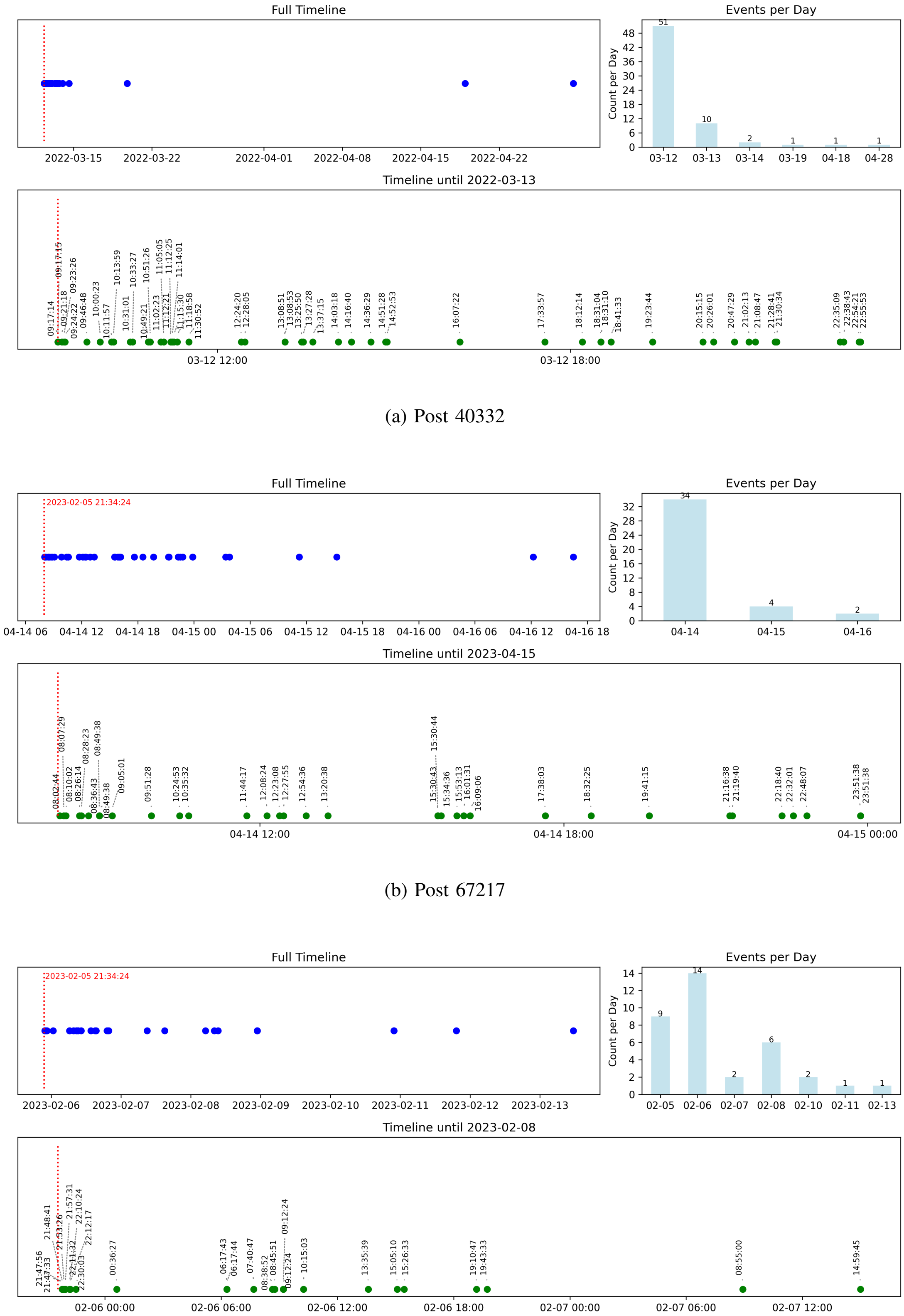
A potential limitation of our dataset is the absence of the unprocessed content of the posts. This decision was deliberate in order to safeguard privacy and to provide a dataset that supports initial exploration of Telegram across disciplines without compromising ethical standards. Instead, we provide automatically generated keywords and sentiment labels. These were selected on the basis of scalability, but their performance was not the focus of a detailed evaluation. In future work, we plan to examine content-based methods such as topic modeling to extend the range of derived content-based labels.

Our dataset also does not include labels that indicate whether a post constitutes fake news or disinformation. This was not feasible given resource constraints and, more importantly, due to the inherent difficulty of classifying the intention behind each post. For this reason, we focused on identifying and analyzing influential actors in the German disinformation landscape through interviews with experts.

To gain a more comprehensive understanding of the landscape, a comparable dataset of non-disinformation actors in Germany would be highly valuable. We therefore suggest complementing our dataset with other German-language Telegram datasets for comparative analysis (e.g. [10]).

Our raw dataset also includes multimodal elements, such as uploaded images and audio files. We plan to explore methods for generating captions or transcriptions to transform these into analyzable text data, while carefully considering privacy requirements. Additionally, our dataset captures the use of

Fig. 2: Post dissemination for actor ID 6667842874



(a) Post 40332

(b) Post 67217

(c) Post 63031

URLs, which creates opportunities for inter-platform analysis. We intend to further explore this dimension and provide a dedicated dataset as part of future work.

VI. Discussion & Future Work

Our dataset provides valuable insights into the disinformation landscape in Germany and, more broadly, into the use of Telegram. It enables the exploration of behaviors and interactions between actors as well as the dissemination of posts. Although the dataset consists of daily snapshots collected over more than a year from over 900 channels and groups, it remains relatively small in scale. Both the raw and derived metrics are well-documented, offering guidance for further research and interpretation.

A distinctive characteristic of the dataset is its daily crawling approach, which captures more fine-grained behavioral patterns than datasets limited to a single snapshot. Beyond the study of disinformation, the dataset can also serve to investigate Telegram usage in general. For instance, it highlights behavioral differences between channels and groups that influence dissemination dynamics. A temporal analysis is also possible with this dataset. Furthermore, it can be employed for visualizing Telegram networks.

The dataset also offers a solid foundation for identifying indicators of significant actors on Telegram. This is relevant not only for the study of disinformation but also for social media analysis, as it allows the detection of influential actors who may not be visible when relying solely on simple raw metrics, such as the INFLUENTIAL_ACTORS we defined in this work.

Future work will focus on generating datasets from other social media platforms with an emphasis on data minimization. Another key objective is to develop generalizable and scalable indicators for identifying significant actors using this dataset, which can subsequently be evaluated on other Telegram datasets.

VII. Conclusion

This paper presents a comprehensive dataset for analysing the dissemination of posts and behavioural patterns of actors on Telegram. The actors were carefully selected by experts to identify key actors within the German disinformation environment of Telegram. Due to privacy and legal constraints, we did not release the raw data containing retraceable IDs and full post content. However, through a thorough analysis of the extracted metadata, we have restructured the data to support metadata-driven research approaches that balance utility with ethical and privacy considerations.

Our work addresses key challenges arising from Telegram's decentralized and semi-anonymous nature, offering a valuable resource for understanding disinformation spread in this important but underexplored social media platform. Beyond introducing the dataset, our paper provides first insights and demonstrates how it can be used to support further analyses, offering both concrete examples and guidance for future research.

VIII. FAIR Data Statement

Our dataset adheres to the FAIR principles to promote transparency, accessibility, and reusability.

- Findable: The dataset is assigned a persistent identifier (DOI: 10.5281/zenodo.16994657) and is registered in Zenodo (Link), to ensure easy discovery by researchers and automated systems.
- Accessible: Data files and documentation are openly available via Git:Dynamo. Access to sensitive data components is regulated through pseudonymization to comply with ethical and legal standards. It also serves as a point of contact for inquiries.
- Interoperable: The dataset is provided in the widely used, machine-readable format CSV, facilitating integration with diverse analysis tools and platforms.
- Reusable: Comprehensive documentation, including data collection methods and usage guidelines, is provided. The dataset is shared under a clear license (Creative Commons Attribution 4.0 International) to enable lawful reuse and citation.

References

[1] L. Bernhard, L. Schulz, C. Berger, and K. Unzicker, "Verunsicherte Öffentlichkeit: Superwahljahr 2024: Sorgen in Deutschland und den USA wegen Desinformationen," p. 67 p., 2024. Artwork Size: 67 p. Publisher: Bertelsmann Stiftung.

[2] E. Broda and J. Strömbäck, "Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review," *Annals of the International Communication Association*, vol. 48, pp. 139–166, Apr. 2024.

[3] D. Plikynas, I. Rizgelienė, and G. Korvel, "Systematic Review of Fake News, Propaganda, and Disinformation: Examining Authors, Content, and Social Impact Through Machine Learning," *IEEE Access*, vol. 13, pp. 17583–17629, 2025.

[4] F. Pierri and S. Ceri, "False News On Social Media: A Data-Driven Survey," *ACM SIGMOD Record*, vol. 48, pp. 18–27, Dec. 2019.

[5] S. Ashraf, I. Bezzaoui, I. Andone, A. Markowetz, J. Fegert, and L. Flek, "Defakts: A german dataset for fine-grained disinformation detection through social media framing," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 4580–4591, 2024.

[6] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *SIGKDD Explor. Newsl.*, vol. 21, p. 48–60, Nov. 2019.

[7] E. Aïmeur, S. Amri, and G. Brassard, "Fake news, disinformation and misinformation in social media: a review," *Social Network Analysis and Mining*, vol. 13, p. 30, Feb. 2023.

[8] K. Schäfer and J.-E. Choi, "Transparency in messengers: A metadata analysis based on the example of telegram," in *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, pp. 1–3, 2023.

[9] J. Baumgartner, S. Zannettou, M. Squire, and J. Blackburn, "The pushshift telegram dataset," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 840–847, May 2020.

[10] S. Gangopadhyay, D. Dessí, D. Dimitrov, and S. Dietze, "Telescope a longitudinal dataset for investigating online discourse and information interaction on telegram," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 19, pp. 2423–2433, Jun. 2025.

[11] T. Setz, I. Vogel, M. Steinebach, K. Bader, n. krämer, G. Hornung, Y. Yannikos, L. Rinsdorf, J. Kluck, and C. Jansen, *Desinformationen und Messengerdienste: Herausforderung und Lösungsansätze*, pp. 317–370. 01 2022.

[12] L. Rinsdorf, K. Bader, and C. Jansen, "Telegram als Plattform für staatsskeptische Akteur:innen," in *Politischer Journalismus* (C. Nuernbergk, N. F. Schumacher, J. Haßler, and J. Schützeneder, eds.), pp. 131–146, Nomos Verlagsgesellschaft mbH & Co. KG, 2024.

[13] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, “Yake! keyword extraction from single documents using multiple local features,” *Information Sciences*, vol. 509, pp. 257–289, 2020.

[14] K. Bader, K. F. Müller, and L. Rinsdorf, “Wanderer between the worlds: Telegram use from the users’ perspective,” *Publizistik*, vol. 70, pp. 133–155, May 2025.

[15] J.-E. Choi, K. Schäfer, and Y. Yannikos, “Scientific Appearance in Telegram,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, pp. 2091–2096, May 2024.