TeraGram: A Structured Longitudinal Dataset of the Telegram Messenger

Anastasia Golovin[*1,2], Sebastian B. Mohr[*1,2], Arne I. Gottwald[3, 1, 4], Ulrik Hvid[5, 6], Srushhti Trivedi[7,8], Joao Pinheiro Neto[9], Andreas C. Schneider[1, 2], Viola Priesemann[1, 2]

[1]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany.
[2]Institute for the Dynamics of Complex Systems, University of Göttingen, Göttingen, Germany.
[3]Campus Institute for Dynamics of Biological Networks, University of Göttingen, Göttingen, Germany.
[4]Campus Institute Data Science, University of Göttingen, Göttingen, Germany.
[5]Biocomplexity, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
[6]PandemiX – Center for Interdisciplinary Study of Pandemic Signatures, Copenhagen, Denmark
[7]Institute of Medical Informatics, University Medical Center Göttingen, Germany
[8]Institute for Ethics and History of Medicine, University Medical Center Göttingen, Germany
[9]IDea_Lab, University of Graz, Graz, Austria
anastasia.golovin@ds.mpg.de, sebastian.mohr@ds.mpg.de, arne.gottwald@ds.mpg.de, ulrik.hvid@nbi.ku.dk, srushhti.trivedi@med.uni-goettingen.de, joaoxp@gmail.com, andreas.schneider@ds.mpg.de, viola.priesemann@ds.mpg.de

Abstract

Here we present a massive longitudinal dataset of public Telegram content, comprising over 5.9 billion messages dating from 2015 to 2025, collected from 712 thousand channels and groups, enriched with metadata on forwards, reactions, and polls. The dataset spans multiple languages including Russian and Farsi, representing countries where Telegram shows mainstream adoption, as well as Western languages where Telegram is used in specific sub-communities. The dataset has several advantages. First, when restricted by language, it provides a versatile example of an algorithm-free platform, contrary to many other social media platforms that are strongly influenced by opaque content-curation algorithms. Second, it enables comparative studies across different languages, communities, and user bases under identical platform affordances. The dataset thus offers a foundation for studying engagement patterns, network evolution, and community formation in the absence of algorithmic curation.

Code — https://github.com/Priesemann-Group/telegram_quality_control

Datasheet for datasets — https://github.com/Priesemann-Group/telegram_quality_control/DATASHEET.md

Dataset (preview) — https://zenodo.org/records/18262126

Introduction

Social media platforms play a central role in shaping public opinion (McGregor 2019; Okechukwu 2023). However, researchers seeking to understand this influence face a persistent challenge: most platforms use sophisticated and often proprietary algorithms to filter and prioritize content. When algorithms determine what users see, it is challenging to disentangle organic social dynamics from effects induced by proprietary algorithms.

Telegram stands out as a platform with minimal algorithmic intervention. Originally launched as an instant messaging app, it has evolved into a hybrid platform that combines private messaging with large-scale public broadcasting. Unlike most mainstream platforms, Telegram features a chronological feed and little to no algorithmic recommendations. Its emphasis on privacy and limited moderation has attracted activists and journalists who are seeking secure communication in authoritarian regimes (Su, Chan, and Paik 2022; Wijermars and Lokot 2022; Urman, Ho, and Katz 2021); it also made Telegram a hub for misinformation, extremist content, and illicit activities (Kiess 2025; Walther and McCoy 2021). This has led to pronounced differences in adoption and public sentiment towards the platform across countries: Telegram is widely adopted and trusted in many post-Soviet countries (Chernenko and Dutton 2025; Wijermars and Lokot 2022) and in Iran (Ghorbanzadeh and Saeednia 2018; Kargar and McManamen 2018) while viewed critically in the West for associations with illegal and extremist content (Kiess 2025; Walther and McCoy 2021).

To systematically study discourse on Telegram across these diverse contexts, we collected TeraGram, a large-scale, structured dataset of publicly available Telegram messages using a snowball crawling method (Goodman 1961). This approach systematically discovers and maps interconnected channels and communities, allowing to capture rich temporal and relational data on the platform.

Several large-scale Telegram datasets have already been published. For example, the Pushshift Telegram dataset (Baumgartner et al. 2020) collected over 300 million messages starting from a seed of extremist and cryptocurrency channels, providing an early glimpse into Telegram's content landscape. Building on this, TGDataset (La Morgia, Mei, and Mongardini 2025) expanded coverage to 400 million messages across 120,000 channels, offering a more balanced view of the platform without focusing on one topic.

[*]These authors contributed equally.

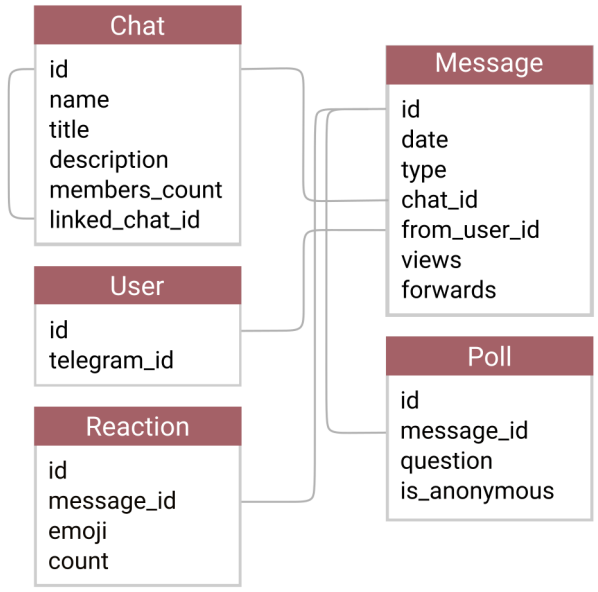

Figure 1: A simplified entity-relation diagram of the dataset. See SI, Fig. 8 for the non-simplified version.

Recent efforts have achieved even greater scales. The Telescope dataset features enriched metadata for 500 thousand public channels and message metadata for 71 thousand fully downloaded channels (Gangopadhyay et al. 2025).

In addition to those general-purpose datasets, several datasets focus on specific communities or geopolitical events. A dataset by Blas et al. on the 2024 US Presidential Election (Blas, Luceri, and Ferrara 2025) includes over one billion messages and uniquely incorporates private groups accessible by invite, with particular emphasis on English-speaking communities. Two datasets examine how the Russia-Ukraine war is covered on Telegram: the dataset by (Kireev et al. 2025) features examples of propaganda and moderation, while (Bawa et al. 2025) contrasts pro- and anti-Kremlin Telegram channels. Finally, the Schwurbelarchiv project (Angermaier et al. 2025) collected a smaller number of German channels, and used AI to transcribe multimedia content such as voice messages and videos. A summary of existing datasets is provided in SI, Tab. 2.

TeraGram provides the largest longitudinal record of public Telegram content to date (Tab. 1). Beyond its size, it improves over existing datasets in several aspects. First, while most existing datasets are distributed as JSON files that need to be parsed before analysis and might contain inconsistent schemas, TeraGram is structured in a relational format that enables computationally efficient analysis at scale (Fig. 1). The core relational tables Chats, Messages, Users, and Polls are populated with data directly available from the Telegram API, including message metadata (timestamp, view count, forward count), poll questions, and public chat information. Second, TeraGram aims to comprehensively capture Telegram-specific features that are absent or incomplete in prior work, including discussion groups, emoji reactions, polls, embedded URLs, hashtags, and reply-thread relationships. While some of these are available in prior work, no existing dataset provides all of them together. Third, rather than targeting a specific language or community, TeraGram spans a wide range of languages and public communities; we further augment the data by algorithmically inferring the primary language of each chat, enabling systematic cross-lingual analysis.

To preserve privacy, all user identifiers are pseudonymized, only public channel and group metadata is retained, and binary media blobs are excluded. The full message text is available to qualified researchers upon reasonable request.

With a total volume of 3.33 TB of data (1.43 TB excluding text) (Tab. 1 and SI, Tab 3), the dataset is distributed in a Parquet format that can be easily ingested into a relational database, accompanied by a representative CSV sample for accessibility. Both datasets are shared under the Open Data Commons Attribution (ODC-By) license that allows users to freely share, modify, and use the dataset as long as they attribute it. In this paper, we illustrate the utility of the dataset with preliminary analyses of language distribution, network degree, and topic clustering.

Overall, TeraGram provides a valuable resource for advancing research in social media analysis, online user behavior and computational social science — particularly within a platform characterized by minimal algorithmic interference.

Results

Dataset overview

Telegram supports two formats for public communication: channels and groups. *Channels* function as broadcasting services where only administrators can publish content, while subscribers receive and read these messages. *Groups*, in contrast, allow all members to post and participate in conversations. Furthermore, a channel can be linked to a dedicated group for its subscribers where members can discuss the channel's content and comment on individual posts. Such a group is termed a *discussion group*. Throughout this paper, we use the terminology from the Telegram API where the term *chats* serves as an umbrella term for both channels and groups.

Amounting to 3.33 terabyte of data, the TeraGram dataset presented in this study is one of the largest and most comprehensive Telegram datasets available to date (see Tab. 1 for summary statistics). The core dataset is organized into interrelated tables representing key Telegram entities, including messages, users, chats, polls, and emoji reactions (see Fig. 1 and SI, Fig. 8). This relational structure enables a detailed reconstruction and analysis of user interactions, group dynamics, and content dissemination patterns within the Telegram ecosystem.

Data was collected via the official Telegram user API using a snowball-crawling approach (see Methods for a detailed description). Starting from the 100 largest political channels by subscriber count (SI, Tab. 12), the crawler iteratively discovered new publicly accessible chats through forwarded messages. Discovered chats were prioritized for

Table	Count	Size
Messages	5.95B	1.0 TB
Message text (available on legitimate request)	5.51B	1.9 TB
Hashtags	498M	33 GB
URLs	655M	63 GB
Reactions	3.60B	319 GB
Discovered chats (umbrella term for channels and groups)	4.54M	1.1 GB
Downloaded chats (includes full message history)	712k	172 MB
Poll questions	21.29M	5.2 GB
Poll answers	79.44M	8.8 GB
Users (pseudonymized)	15.30M	10 GB
Channel members	3.57M	446 MB
Total		3.33 TB

Table 1: Summary of collected Telegram data. The values correspond to the approximate number of rows and disk size of the corresponding SQL table, including indexing.

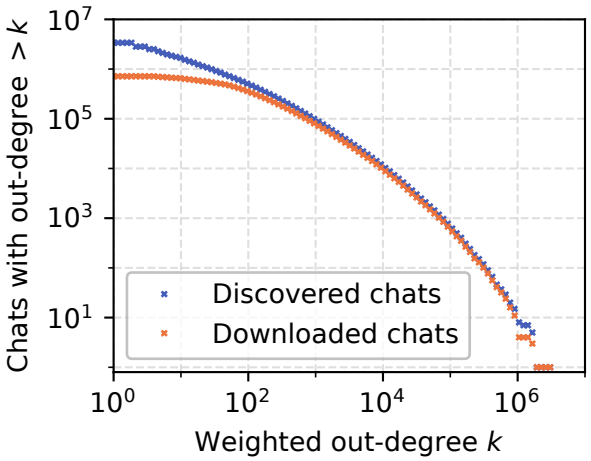

Figure 2: Complementary cumulative distribution function (CCDF) of the weighted out-degree of discovered and downloaded chats. The gap between the curves at low degrees is due to the crawler prioritizing chats for download proportionally to the degree. The out-degree is defined as the number of messages forwarded from a given chat to other downloaded chats, with edge weights proportional to the number of forwarded messages.

download based on their out-degree, i.e., the number of times they were forwarded in already downloaded chats. More forwards resulted in higher download priority. This approach ensures that the crawler prioritizes the highly influential chats from the tail of the out-degree distribution, which may present the hubs of the Telegram network (Fig. 2). Additionally, focusing on popular chats helps to protect user privacy, as users participating in smaller chats might not expect that their messages, though public, would be visible to a large audience.

Overall, we discovered 4 382 659 channels and 160 617 groups, and fully downloaded 711 782 channels and 10 060 groups. Out of those, 4017 groups (40%) are discussion groups linked to channels. We also collected user profile data from users who posted in groups and from groups where the subscriber list was publicly visible. In those cases, we hashed usernames and first and last names prior to storage to ensure that this sensitive information is not retained at any point during the data collection process. When a user's phone number was visible, we retained only the country code and removed all other digits.

The data was collected between May 2025 and November 2025; however, since the Telegram API provides the whole history of a chat, the collected data spans from September 2015 to November 2025 — over a decade-long period (Fig. 3). Because chats were not updated after their initial download, chats encountered later in the crawling process contain more recent messages than those downloaded at the beginning of the collection period.

The dataset is shared under a two-tiered access model: rich metadata is openly available, while message text is accessible to researchers upon reasonable request. The metadata includes view counts and forward counts, which serve as measures of user engagement. When a message is forwarded from another chat, we link it to its original source message and originating channel when this information is available. Similarly, replies include references to the messages they respond to, creating explicit conversational threads. In addition, we also extracted two types of text markup entities: URLs and hashtags. These provide an efficient way to access structural text information without parsing the raw text. For messages containing binary data such as images, audio messages, and videos, we store only the metadata and any accompanying text captions.

We also collected data on special features of Telegram such as polls and emoji reactions. Poll data include the question text, answer options, and aggregate vote counts when available. Reactions are stored as aggregated emoji counts per message without user-level information. Message reactions can be used as a measure of user engagement or as an efficient way to evaluate sentiment of a given message.

Quality control

Given the scale and diversity of our Telegram corpus, rigorous quality control is essential to ensure data integrity. As outlined below, our quality control workflow combines automated integrity inspection with manual spot-checks on a stratified random sample of messages. We also publish a datasheet in accordance with (Gebru et al. 2021) that provides comprehensive documentation of dataset provenance. Together, those steps verify data integrity and quality before any downstream analysis.

Data integrity Following the approach recommended by (Elazar et al. 2024), we execute three automated quality control checks: duplicate detection, length distribution analysis, and n-gram inspection. The goal of those checks is to catch scraping artifacts, duplicates indicating spam or bot

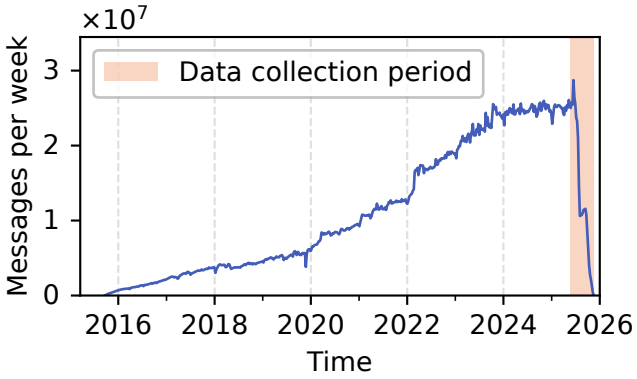

Figure 3: Number of messages posted per week in the dataset. The time series spans September 2015 to November 2025. Apparent variations during the collection period are influenced by the crawling procedure: chats were downloaded only once, so those discovered later in the crawl contribute more recent messages than chats collected earlier.

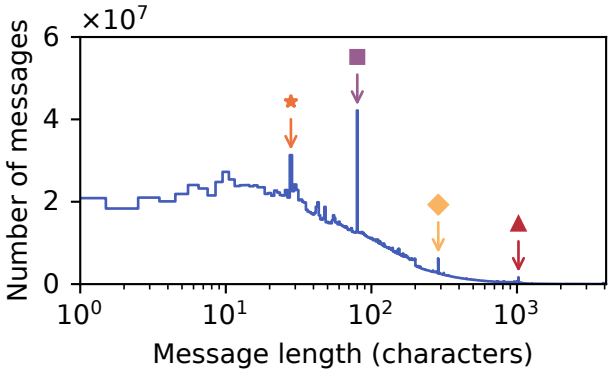

Figure 4: Distribution of the message length shows several distinct peaks caused by systematic patterns. Star: YouTube links; square: message redacted by Telegram for copyright or Terms of Service violation; diamond: an advertisement message mass-posted in one Arabic channel; triangle: caption length limit.

activity, and anomalies in message length distributions that can be due to automated truncation.

Duplicate detection analysis reveals that 53.7% of messages are unique. Duplications can be explained by several factors: very short messages (e.g., single-word responses or greetings) that naturally occur frequently; messages containing only URLs without additional text; or advertisement posts that are frequently reposted without changes in content. Duplicates were not removed from the dataset.

Length distribution analysis shows several strong peaks at specific message lengths (Fig. 4). We investigated the origin of the four most prominent peaks. The peak at 28 characters (orange star) corresponds to YouTube links that follow the format "https://youtu.be/abcdefghijk". The peak at 80 characters (purple square) is due to messages redacted by Telegram for copyright or Terms of Service violations. In such cases, the original message text is replaced by either "This message couldn't be displayed on your device due to copyright infringement." or "This channel can't be displayed because it violated Telegram's Terms of Service." Both are 80 characters long. The peak at 288 characters (yellow diamond) was caused by an advertisement for other Telegram channels that was mass-posted in one Arabic channel. Finally, the peak at 1024 characters (red triangle) corresponds to the caption length limit, i.e., the message length limit for messages containing multimedia content.

Finally, we performed n-gram analysis to detect artifacts in text such as unusual punctuation, spam, and near-duplicate messages. For this, we extract the top 10 most popular unigrams, 3-grams, and 10-grams from all English-speaking chats (SI, Tables 4-6). Following the approach in (Elazar et al. 2024), we did not filter stop words or clean the text beyond converting it to lowercase, as the goal is to detect artifacts rather than describe message content. The most popular n-grams show a mix of typical English phrases, especially in unigrams and 3-grams, and inorganic strings that become more common in 10-grams. Those inorganic strings are due to boilerplate text that some channels append to every message, text separators, or redaction messages from Telegram discussed earlier. We also find an additional redaction message: "This channel can't be displayed because it violated local laws." This text is displayed whenever a message was forwarded from a chat that violated local laws to a normal chat. Such chats are called restricted in the Telegram API and are marked in the dataset as such in the field `is_restricted`.

Bot activity A key quality concern is the prevalence of bot-generated messages, as automated posts can skew analyses of human behavior and information spread. Registered Telegram bots, easily identifiable via the API, constitute just 0.4 % of users and contribute 0.14 % of messages in our dataset. More insidious "troll farm" or LLM-driven accounts that mimic human activity are harder to detect and beyond this paper's scope (see, e.g., studies on coordinated inauthentic behavior (Cinelli et al. 2022)). Nevertheless, we do not expect such accounts to be ubiquitous in the dataset. First, our data span back to 2015, predating the recent LLM boom. Second, the dataset primarily consists of channels where posting is restricted to administrators. Consequently, the most significant risk of bot contamination stems from channel administrators using bots to manage their channel. We leave the investigation of this scenario for dedicated future work.

Analysis of the Telegram dataset

To characterize the thematic and informational range of the dataset, we conduct analyses focusing on language use, external references, and topical structure.

We identify the primary language of each downloaded chat using the `fast_langdetect` library (LlmKira 2025; Joulin et al. 2016b,a). For each chat, we use the first 100 messages since creation, all concatenated into one long string, where we remove URLs and whitespace. Chats with

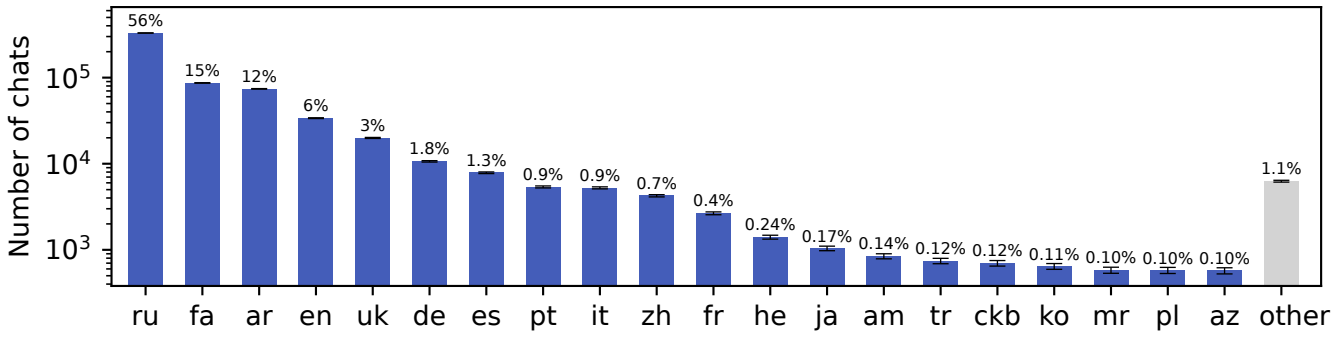

Figure 5: Languages of fully downloaded chats. Chat language is classified based on the text of the first 100 messages. Language codes follow the ISO standard. Error bars give the 95% CI interval.

a language-classification confidence score below 0.8 are excluded from the analysis. Error bars on language frequency counts are computed using Wilson score confidence intervals, treating each language count as a binomial proportion of the total corpus.

The dataset covers a wide range of languages, with top-5 languages accounting for approx. 91.7 % of all chats (Fig. 5). We find a high fraction of Russian (56 %) and Farsi (15 %) chats, which reflects Telegram's mainstream adoption in Russia and other post-Soviet countries and Iran. English chats make 6 % of downloaded chats, which still amounts to 33 829 chats.

To evaluate the quality of our language classification, we analyzed whether forwarded messages typically originate and arrive within a chat of the same language (e.g., Russian to Russian). This is the case for 83.3% of forwards where the source chat was successfully downloaded. In almost half the cases where source and destination were classified differently, either the source (25%) or destination (19%) was English. The former fraction likely represents non-English chats forwarding English material, while the latter, perhaps, represents chats that write in English even though they belong to a non-English subnetwork. As a further control, we manually inspected sets of 100 random messages from Russian, English, Arabic and German chats and found a low number of misclassifications (5%, 9%, 1% and 3%, respectively), consistent with the reported 95% accuracy of `fast-langdetect` (LlmKira 2025; Joulin et al. 2016b,a).

To evaluate the prevalence of misinformation in the data, we assessed the reliability of news domains shared in messages using the aggregated "wisdom of experts" domain rating by (Lin et al. 2023). Since the domain rating primarily covers the US media landscape and has limited coverage of Russian and Iranian news sources, we only performed this analysis on English-language chats. For every chat, we extracted all URL entities from the corresponding table. URL rehydration was not necessary, as link shorteners are rarely used on Telegram; for example, Bitly links make up only 1.2% of URLs. We applied a blacklist to remove domains like search engines and social media platforms that appear in the Lin et al. dataset but do not refer to news sources

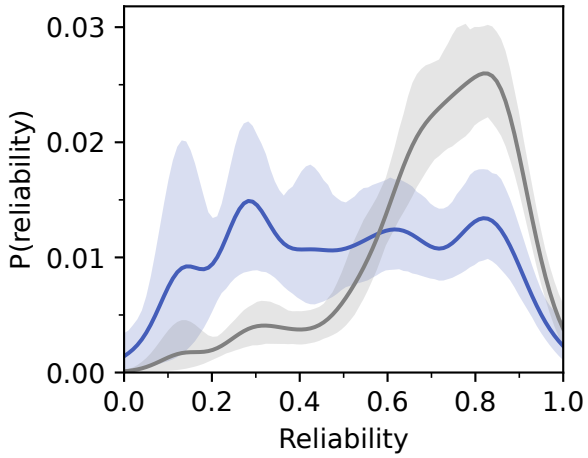

Figure 6: High prevalence of unreliable URLs in English-speaking Telegram chats compared to a mainstream platform like Twitter. The bands represent the 95% CI interval obtained by clustered bootstrapping on domains.

(SI, Tab. 7). Of the remaining URLs, we randomly sampled 1% to reduce computational costs, which resulted in 861 thousand URLs, and fit a kernel density estimator with a Gaussian kernel to estimate the probability distribution. To compute the CI intervals, we perform clustered bootstrapping with domains as clusters; in other words, instead of resampling individual URLs, we sample entire domains with replacement 1000 times.

Overall, we observe a high prevalence of URLs with a reliability score below 0.6 in English-speaking Telegram chats (Fig. 6). For comparison, we perform the same analysis on a Twitter dataset of all tweets within a 24-hour period on September 21, 2022 (Pfeffer et al. 2023) (Fig. 6, gray line). The resulting distributions differ distinctly: URLs on Twitter are concentrated within the 0.6 to 1.0 reliability range. In contrast, the distribution for English-speaking Telegram varies widely, with 60% of URLs lying below the 0.6 threshold.

To explore the main themes discussed in chats, we applied

BERTopic (Grootendorst 2022), an unsupervised topic modeling pipeline, to messages from the four most prevalent languages in our dataset: Russian, Farsi, Arabic, and English. For each language, we filtered chats with a language classification score above 0.8 and extracted all messages from those chats. We cleaned the messages by removing URLs and stripping consecutive whitespace, then filtered out messages shorter than 50 characters and system messages identified in quality control analysis. We then randomly sampled 1 million messages per language and performed topic modeling on each set, with each run being executed on an NVIDIA A100 40GB GPU. To produce readable labels for each topic, we use BERTopic's ChatGPT integration (`gpt-4o-mini` model), where we provide it with topic keywords identified by BERTopic and a sample of messages and ask it to return a short label in English. The preprocessing code and topic modeling pipeline with all hyperparameters are available in our GitHub repository.

All four languages feature discussions about politics, current world events, as well as non-political topics such as sports and music (Fig. 7, SI, Fig. 10, and SI Tables 8-11 for the top-40 topics in tabular format). However, English top-40 topics also include entries like "racial discourse", "antisemitic narratives", "climate change hoax" and "jfk assassination conspiracy" that suggest far-right and conspiracy content. In contrast, Russian and Farsi topics cover a more diverse set of day-to-day topics like books, fashion, art, and music, whereas Arabic topics are predominantly religious.

Usage guide

The dataset is designed for flexible use across many fields, including computational social science, digital humanities, and machine learning applications. It is provided as a relational database, with tables representing messages, channels, users, polls, and reactions. This structure supports efficient queries on content, metadata, and user-channel interactions.

The dataset adheres to FAIR principles (Wilkinson et al. 2016) to ensure researchers can seamlessly download and integrate the data into existing workflows:

- **Findable:** Both the full dataset and subsampled CSV tables are indexed with DOIs on established data-sharing platforms.
- **Accessible:** Public portions of the dataset are available for download without authorization.
- **Interoperable:** The full dataset is provided as Parquet files, which is an open-source table format supported by many data processing systems. The subsampled dataset is provided in CSV format for readability and easy inspection.
- **Reusable:** To facilitate reuse, we provide schema documentation, access instructions, example SQL queries, and preprocessing scripts in our GitHub repository. Dataset provenance and limitations are documented in an accompanying datasheet.

Ethical use is strongly encouraged. Researchers are advised to respect the original context of communication and avoid deanonymization efforts.

Discussion

In this work, we present TeraGram, a large-scale structured dataset of the publicly available Telegram ecosystem. Our dataset is, to the best of our knowledge, the largest to date; it also contains rich metadata on various features of Telegram such as polls, emoji reactions, URLs, hashtags, discussion groups, forwards and replies. Message texts are available upon request. The structured format provides a convenient and efficient way for researchers to analyze various aspects of the platform.

The dataset spans a wide range of languages, with billions of messages in Russian, Farsi, and Arabic, and hundreds of million in English and other European languages. The high proportion of Russian and Farsi content reflects Telegram's popularity in Russia and other post-Soviet countries and Iran.

Telegram usage varies across languages. While Russian and Farsi chats feature more mainstream topics, English chats share URLs of remarkably low reliability: 59% fall below the commonly accepted 0.6 threshold for classification as reliable sources. This contrasts with mainstream English social media, where a comparable study found only about 15% of sources to be questionable (Di Martino et al. 2025). Direct comparison for Russian and Farsi is not currently possible, as existing reliability rankings lack sufficient coverage of Russian and Iranian news sources.

Several important limitations must be considered when working with this data. First, since the crawler prioritizes queued chats based on the number of messages that were forwarded from them, it is biased towards popular chats. We also cannot reach chats that do not belong to the same connected component as our seed chats. Additionally, we did not crawl private conversations and groups. Researchers should therefore exercise caution when extrapolating findings to broader user behavior on Telegram, especially in contexts where private or low-visibility conversations play a critical role.

Second, Telegram itself has undergone significant evolution during the nearly decade-long period that our data covers. Early on, the platform operated with minimal algorithmic intervention, but in 2024, features such as recommendations of related channels, sponsored messages, and enhanced content-discovery tools have been introduced (Telegram 2024). This evolution implies that interaction dynamics likely differ between 2017 and 2025. Moreover, natural shifts in language and community norms over time introduce additional variance. Consequently, any analysis pooling data across this period must account for this temporal heterogeneity.

Overall, the presented dataset allows a wide range of future investigations into online communication and information dynamics. For instance, recent work already leveraged the data to study the spread of information via external URLs (Ventzke 2025), fine-tune LLMs (Brockers, Ehrlich, and Priesemann 2025), or improve misinformation detection algorithms (Keßler et al. 2026).

We anticipate that the dataset will support diverse downstream applications, including network modeling, bot detec-

Figure 7: Topics in Russian and English chats identified using BERTopic. While both English and Russian datasets contain topics like sports and current events, the English dataset includes a subset of far-right topics (e.g., "antisemitic narratives," "climate change hoax"). In contrast, Russian topics predominantly reflect mainstream diverse interests, including books, fashion, art, and music. See Tables 8-11 in the SI for the top-40 topics of the four languages in a tabular format.

tion, and community formation across multilingual and longitudinal corpora.

Conclusion

In this paper, we collected a longitudinal dataset of public Telegram chats spanning nearly a decade worth of data. The dataset thus provides a unique opportunity to study organic information diffusion across diverse community structures on a platform with minimal moderation and algorithmic interference.

Methods

Data collection

Telegram provides no official data access for researchers, necessitating a custom collection pipeline. We built an asynchronous crawler using Pyrogram (Pyrogram 2025), a Python interface to the MTProto Telegram API, which interacts with the platform through standard client authentication.

We employed a snowball crawling strategy (Goodman 1961) that exploits Telegram's message-forwarding feature to discover interconnected chats. Starting from a curated seed set of 100 public channels (see SI, Tab. 12), the crawler recursively identified new chats by resolving the origins of forwarded messages. This approach enables scalable, structurally informed data collection (La Morgia, Mei, and Mongardini 2025; Baumgartner et al. 2020).

To maximize coverage of influential chats, we assigned a download priority to queued chats equal to their observed out-degree, i.e., the number of times the crawler encountered forwards from the chat in already downloaded chats. The crawler always downloaded the chat with the highest priority of all discovered and not yet downloaded chats, ensuring central network hubs were captured first (SI, Algorithm 1). This priority queue approach optimizes the discovery of large, interconnected chats within sampling constraints.

The crawler implemented several specific behaviors to ensure data completeness. For each channel, we also crawled its linked discussion group (if it existed), capturing user comments and interactions that provide crucial context for message interpretation. To access statistics of non-closed polls, we had to cast a vote before recording the results, then subtracted our own vote during post-processing to maintain the original voting distributions. Additionally, we addressed two data integrity challenges. First, when reconstructing reply threads, we could not rely on Telegram's sequential message identifiers, because those shift when messages are deleted to preserve consecutive indexing. Instead, we matched replies using timestamps. Second, we encountered the same problem when identifying the original source of forwarded messages, where we also matched the messages by timestamps. This approach might have led to inconsistencies whenever two messages are posted in the origin chat within the same second.

Telegram imposes rate limits on API requests per account. To scale data collection, we distributed the crawling process across 200 authenticated accounts managed by three worker machines. This parallelization enabled us to bypass per-account limitations while maintaining a single database instance for consolidated storage, running on a fourth separate machine with local NVMe SSD storage drives.

We collected comprehensive metadata including messages, channels, users, polls, and forwarding relationships (SI, Tab. 3). Binary content (images, videos) was excluded due to storage and copyright considerations. Phone numbers were excluded due to privacy concerns. Usernames were hashed upon storage. The resulting dataset was stored in a PostgreSQL relational database optimized for high-volume insert performance.

Ethical Statement

Several ethical and legal considerations arise when collecting and analyzing data from social media platforms, particularly around privacy, informed consent, and data security. Even when data originates from publicly accessible Telegram groups or channels, ethical concerns persist, especially in politically sensitive contexts where users rely on the platform's perceived anonymity and resistance to surveillance (Burnett and Feamster 2015; Olteanu et al. 2019). Simply sharing opinions or experiences in a public forum does not necessarily imply a willingness to have that content analyzed in research or shared in datasets (Crawford and Finn 2015).

Public social media data can still reveal sensitive information, including political views or personal health, and may be traceable to individuals through context, even without explicit identifiers (Ohm 2009). These risks are particularly acute when studying events such as protests or elections, where unintentional exposure can carry real-world consequences (Cohen and Ruths 2013).

To mitigate these risks, we implemented multiple safeguards. First, we do not publish message text, as it can contain sensitive information that is hard to remove at scale. Furthermore, we algorithmically redact common forms of personally identifiable information (PII), such as phone numbers, and guide the crawler towards large, popular chats by weighting download priority by the number of forwards. These measures aim to limit re-identification risk and focus the dataset on content intended for large, public audiences. Nonetheless, the absence of shared community standards for the ethical use of social media data in computational research remains a challenge. As researchers, we advocate for clearer, field-wide guidance to support responsible data sharing and protect user privacy while enabling scientific progress.

Acknowledgments

We would like to thank Jana Lasser for providing early access to the Twitter dataset used in this study. Authors with affiliation "1" received support from the Max-Planck Society. A.G., S.B.M., A.I.G., and V.P. were funded by the German Federal Ministry for Education and Research for the infoXpand project (031L0300A). A.I.G. and A.C.S. were funded by the German Research Foundation – GRK2906 – project number 502807174. U.H. was funded by the Danish National Research Foundation (grant no. DNRF170). J.P.N. was funded by the Austrian Science Fund (DOI:10.55776/P37280). V.P. and A.C.S. were supported by the German Research Foundation under Germany's Excellence Strategy–EXC 2067/1-390729940 (MBExC). This work was further supported by the MWK Niedersachsen via the programs "zukunft.niedersachsen", "Niedersächsisches Vorab" and "Niedersachsen-Profil-Professur".

ChatGPT, Claude, and Qwen3 were used for proofreading and minor stylistic corrections. In the topic modeling analysis, we used BERTopic's ChatGPT integration to generate readable English labels for the identified topics.

References

Angermaier, M.; Hoeldrich, E.; Lasser, J.; and Neto, J. P. 2025. The Schwurbelarchiv: A German Language Telegram Dataset for the Study of Conspiracy Theories. arXiv:2504.06318.

Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020. The Pushshift Telegram Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 840–847.

Bawa, A.; Kursuncu, U.; Achilov, D.; Shalin, V. L.; Agarwal, N.; and Akbas, E. 2025. Telegram as a Battlefield: Kremlin-Related Communications During the Russia-Ukraine Conflict. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2361–2370.

Blas, L.; Luceri, L.; and Ferrara, E. 2025. Unearthing a Billion Telegram Posts about the 2024 U.S. Presidential Election: Development of a Public Dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, 729–732.

Brockers, V. C.; Ehrlich, D. A.; and Priesemann, V. 2025. Disentangling Interaction and Bias Effects in Opinion Dynamics of Large Language Models. arXiv:2509.06858.

Burnett, S.; and Feamster, N. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, SIGCOMM '15, 653–667.

Chernenko, E.; and Dutton, W. H. 2025. Who Trusts Telegram? The Dynamics of Trust and Use of Social Media in Wartime Ukraine. SSRN:5227613.

Cinelli, M.; Cresci, S.; Quattrociocchi, W.; Tesconi, M.; and Zola, P. 2022. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decision Support Systems*, 160: 113819.

Cohen, R.; and Ruths, D. 2013. Classifying Political Orientation on Twitter: It's Not Easy! In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 91–99.

Crawford, K.; and Finn, M. 2015. The Limits of Crisis Data: Analytical and Ethical Challenges of Using Social and Mobile Data to Understand Disasters. *GeoJournal*, 80(4): 491–502.

Di Martino, E.; Galeazzi, A.; Starnini, M.; Quattrociocchi, W.; and Cinelli, M. 2025. Ideological Fragmentation of the Social Media Ecosystem: From Echo Chambers to Echo Platforms. *PNAS Nexus*, 4(9): pgaf262.

Elazar, Y.; Bhagia, A.; Magnusson, I.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, P.; Groeneveld, D.; Soldaini, L.; Singh, S.; Hajishirzi, H.; Smith, N. A.; and Dodge, J. 2024. What's In My Big Data? arXiv:2310.20707.

Gangopadhyay, S.; Dessí, D.; Dimitrov, D.; and Dietze, S. 2025. TeleScope A Longitudinal Dataset for Investigating Online Discourse and Information Interaction on Telegram. *Proceedings of the International AAAI Conference on Web and Social Media*, 19: 2423–2433.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Daumé, H.; and Crawford, K. 2021. Datasheets for Datasets. arXiv:1803.09010.

Ghorbanzadeh, D.; and Saeednia, H. R. 2018. Examining Telegram Users' Motivations, Technical Characteristics, Trust, Attitudes, and Positive Word-of-Mouth: Evidence from Iran. *International Journal of Electronic Marketing and Retailing*, 9(4): 344–365.

Goodman, L. A. 1961. Snowball sampling. *The annals of mathematical statistics*, 148–170.

Grootendorst, M. 2022. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv:2203.05794.

Joulin, A.; Grave, E.; Bojanowski, P.; Douze, M.; Jégou, H.; and Mikolov, T. 2016a. FastText.Zip: Compressing Text Classification Models. arXiv:1612.03651.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016b. Bag of Tricks for Efficient Text Classification. arXiv:1607.01759.

Kargar, S.; and McManamen, K. 2018. Censorship and Collateral Damage: Analyzing the Telegram Ban in Iran. SSRN:3244046.

Keßler, R.; Ventzke, R. D.; Priesemann, V.; and Marzo, G. d. 2026. Network Information Enhances Misinformation Detection on Social Media. Forthcoming.

Kiess, J. 2025. Euroscepticism and Local Far-Right Mobilization via Telegram in Light of the Fundamental Transformation of the Public Sphere. *Political Studies Review*, 23(2): 635–642.

Kireev, K.; Mykhno, Y.; Troncoso, C.; and Overdorf, R. 2025. A Telegram Dataset of Propaganda and Its Moderation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2510–2518.

La Morgia, M.; Mei, A.; and Mongardini, A. M. 2025. TGDataset: Collecting and Exploring the Largest Telegram Channels Dataset. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, 2325–2334.

Lin, H.; Lasser, J.; Lewandowsky, S.; Cole, R.; Gully, A.; Rand, D. G.; and Pennycook, G. 2023. High Level of Correspondence across Different News Domain Quality Rating Sets. *PNAS Nexus*, 2(9): pgad286.

LlmKira. 2025. Fast-Langdetect. https://github.com/LlmKira/fast-langdetect. Accessed on 2025-12-22.

McGregor, S. C. 2019. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, 20(8): 1070–1086.

Ohm, P. 2009. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57: 1701.

Okechukwu, C. 2023. Media Influence on Public Opinion and Political Decision-Making. *International Journal of Political Science Studies*, 1(1): 13–24.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kıcıman, E. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2.

Pfeffer, J.; Matter, D.; Jaidka, K.; Varol, O.; Mashhadi, A.; Lasser, J.; Assenmacher, D.; Wu, S.; Yang, D.; Brantner, C.; Romero, D. M.; Otterbacher, J.; Schwemmer, C.; Joseph, K.; Garcia, D.; and Morstatter, F. 2023. Just Another Day on Twitter: A Complete 24 Hours of Twitter Data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 1073–1081.

Pyrogram. 2025. Pyrogram: Elegant, Modern and Asynchronous Telegram MTProto API Framework in Python for Users and Bots. https://github.com/pyrogram/pyrogram. Accessed on 2025-01-13.

Su, C. C.; Chan, M.; and Paik, S. 2022. Telegram and the Anti-ELAB Movement in Hong Kong: Reshaping Networked Social Movements through Symbolic Participation and Spontaneous Interaction. *Chinese Journal of Communication*, 15(3): 431–448.

Telegram. 2024. My Profile, Recommended Channels and 15 More Features. https://telegram.org/blog/my-profile-and-15-more. Accessed on 2026-01-15.

Urman, A.; Ho, J. C.-t.; and Katz, S. 2021. Analyzing Protest Mobilization on Telegram: The Case of 2019 Anti-Extradition Bill Movement in Hong Kong. *PLoS ONE*, 16(10): e0256675.

Ventzke, R. D. 2025. *Understanding Information Diffusion in Online Social Networks Through the Lens of Critical Processes: A Study on the Telegram Messenger Platform*. Master's thesis, University of Göttingen, Göttingen, Germany.

Walther, S.; and McCoy, A. 2021. US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism. *Perspectives on Terrorism*, 15(2): 100–124.

Wijermars, M.; and Lokot, T. 2022. Is Telegram a "Harbinger of Freedom"? The Performance, Practices, and Perception of Platforms as Political Actors in Authoritarian States. *Post-Soviet Affairs*, 38(1–2): 125–145.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes; we took extensive measures to protect user privacy (see Ethics Statement and the section "Dataset overview"), including not publishing message content and pseudonymizing personal information in the metadata. We do not anticipate any ways how our dataset can perpetuate unfair profiling, exacerbate socio-economic divides, or imply disrespect to societies or cultures.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? No; all methods employed are well-established in the field, so we do not see any need for additional justifications.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A; our data do not contain any demographical information or describe any specific populations.

(e) Did you describe the limitations of your work? Yes, see discussion.

(f) Did you discuss any potential negative societal impacts of your work? Yes; we discuss the negative effects that would result from publishing the entire dataset including message text in the Ethics Statement.

(g) Did you discuss any potential misuse of your work? Yes; see Ethics Statement.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes; the section "Dataset overview" includes a detailed description of how we have pseudonymized the dataset. Since it is technically unfeasable to pseudonymize the content of the messages, we only share the content with qualified researchers upon reasonable request to reduce re-identification risks. However, this measure also means that some of the research results are not directly reproducible based on open parts of the dataset.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We include the code for topic modelling and all other experiments. However, the content of the text messages is not included, so a direct reproduction of the results is not possible.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Training splits for topics modelling were not done. The hyperparameters are available in the code.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes; error bars are reported in all figures where reasonable.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? We do not discuss this explicitly, but topic modeling is also not the main focus of our paper.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? For topic modelling, we are only interested in a qualitative picture rather than precise classification of individual messages. Therefore, occasional misclassifications do not substantially affect our high-level findings about general usage patterns across languages.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes.

(b) Did you mention the license of the assets? Yes, our dataset is distributed under the ODC-By license, see Introduction.

(c) Did you include any new assets in the supplemental material or as a URL? Code for all analyses is available on GitHub: https://github.com/Priesemann-Group/telegram_quality_control. A preview of the dataset is available at https://zenodo.org/records/18262126. The full dataset will be released upon acceptance.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No; obtaining individual consent is infeasible given the scale of data collection.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes; see Ethics Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR

(see Wilkinson et al. (2016))? Yes, see section "Usage guide"

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, a datasheet is available at https://github.com/Priesemann-Group/telegram_quality_control/blob/main/DATASHEET.md

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendix

Dataset	Size	Timespan	Topic focus	Format	Includes text	Features
TeraGram	712k chats, 5.95B messages	Sep. 2015 – Nov 2025	General purpose	Parquet	On request	Discussion groups, reply trees, polls, emoji reactions
Blas et. al	43k chats, 1B messages	Aug. 2024 – Feb. 2025	US elections	SQLite	Yes	Link-accessible private chats
TGDataset	120k channels, 400M messages	Jan. 2021 – Jul. 2022	General purpose	JSON	Yes	
Pushshift Dataset	27.8K chats, 317M messages	Sep. 2025 – Nov. 2019	Seed: right-wing extremism, crypto	JSON	Yes	
TeleScope	71k chats, 120M messages	2015 – Oct. 2024	General purpose	JSON, CSV	No	Temporal message posting patterns, extracted entities
Schwurbelarchiv	6k chats, 40M messages, 3M audio files	Oct. 2015 – Jul. 2022	German-language conspiracies	CSV	Yes	Transcribed multimedia content
Kireev et. al	13 channels, 17.3M messages	Oct. 2020 – Jan. 2024	Russian-Ukrainian propaganda	CSV	Yes	Real-time data collection
Bawa et. al	519 channels, 5.2M Messages	Dec. 2020 – Apr. 2023	Pro- vs. anti-Kremlin stances	CSV	Yes	Replies, emoji reactions

Table 2: A comparison of TeraGram and existing Telegram datasets. The number of chats refers to the number of fully downloaded chats.

Type	Fraction (%)	Content Included	Reason for Omission (if any)
TEXT	30.95	Yes, upon legitimate request	-
PHOTO	27.97	No (caption on request)	Binary media excluded; only textual metadata retained
VIDEO	7.92	No (caption on request)	Binary media excluded
WEB_PAGE	5.07	Yes (link text and metadata)	-
AUDIO	1.62	No (caption on request)	Binary media excluded
STICKER	1.55	No (identifier only)	Media omitted; only human-readable emoji label retained
DOCUMENT	1.27	No (caption on request)	Binary media excluded
ANIMATION	1.17	No	Binary media excluded
VOICE	0.44	No (caption on request)	Binary media excluded
VIDEO_NOTE	0.41	No (caption on request)	Binary media excluded
POLL	0.25	Yes (question and options)	Vote counts may be partially hidden
DICE	<0.01	No	Content depends on server-side evaluation; not reproducible
LOCATION	<0.01	No	Location data excluded due to privacy concerns
CONTACT	<0.01	No	Contact information excluded due to privacy concerns
VENUE	<0.01	No	Venue/location data excluded due to privacy concerns
GAME	<0.01	No	Server-side logic unsupported; content not retrievable
PAID_MEDIA	<0.01	No	Binary/media excluded; restricted content
GIVEAWAY	<0.01	No	Binary/media or server-side logic excluded

Table 3: Message types in the dataset. For each type, we show the total count, share of the dataset, what information was retained, and a brief justification for any omission. Binary media and interactive or privacy-sensitive content were excluded to ensure compliance and scalability.

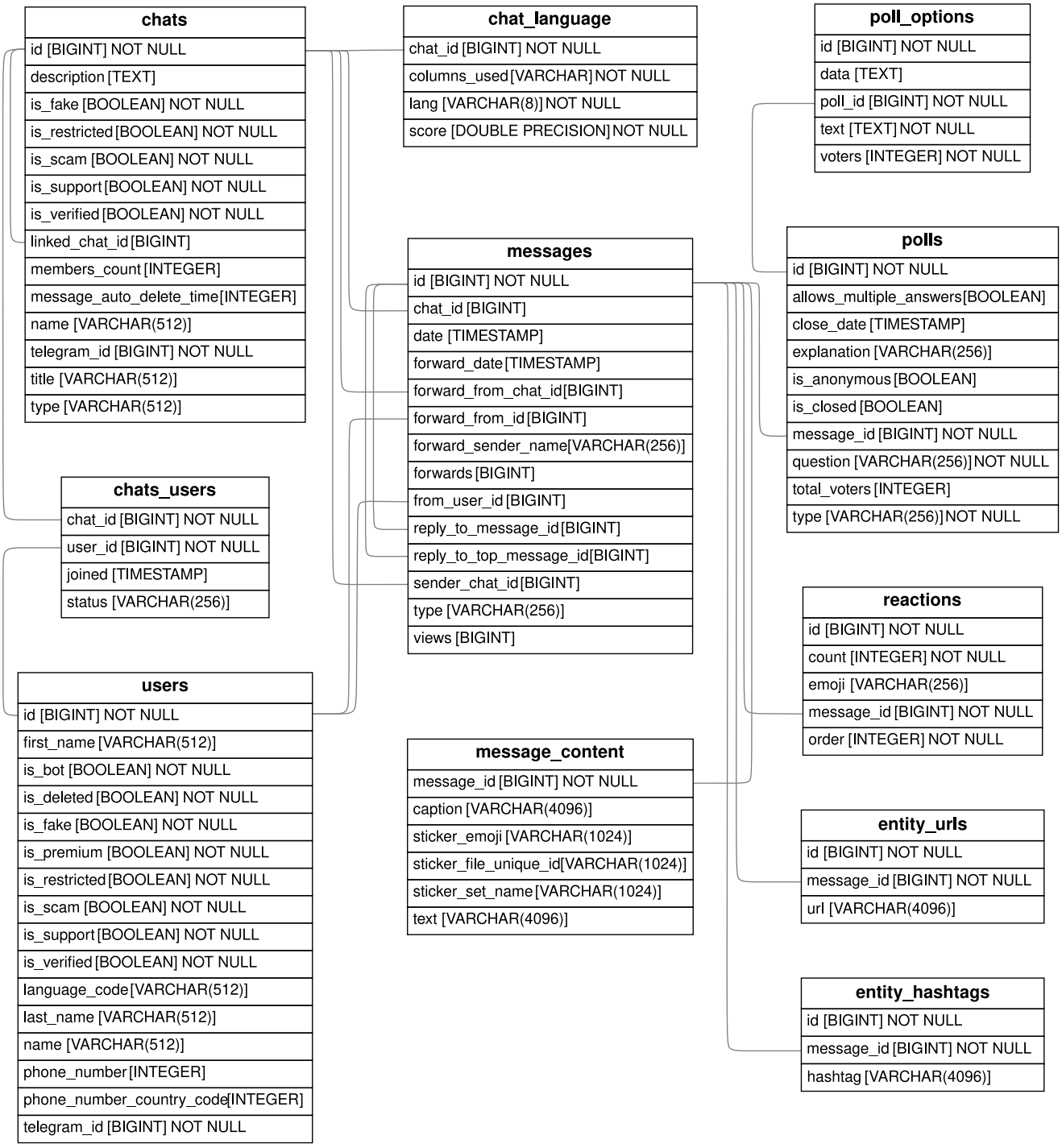

Figure 8: Entity-relation diagram of the SQL database.

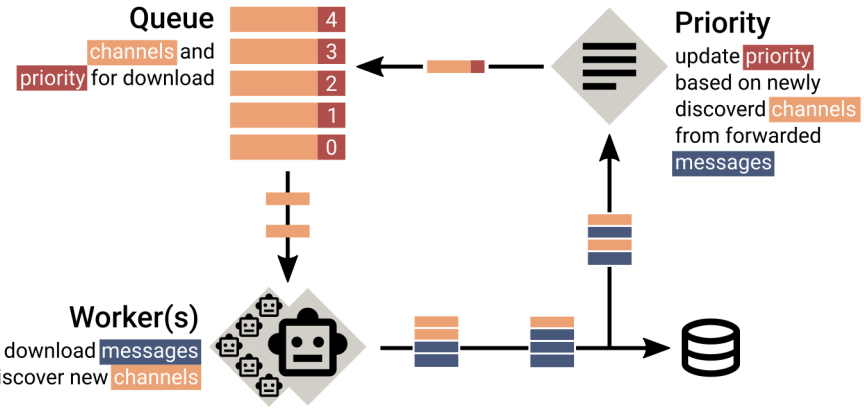

Figure 9: A sketch of the crawling algorithm. The crawler discovers new chats through forwarded messages. The chats are then prioritized for download based on their out-degree, i.e., the number of forwarded messages from this chat into already downloaded chats.

Unigram	Approximate count
the	303780793
to	194300070
of	139954153
and	138496907
a	108742672
in	106484729
on	86053545
is	78172495
this	77054100
be	67760360

Table 4: Top-10 unigrams.

3-gram	Approximate count
due to copyright	36559249
one of the	1739521
- - -	1557765
the united states	1247866
this is the	1185023
a lot of	1051740
this is a	998812
part of the	871766
be able to	846558
as well as	833266

Table 5: Top-10 3-grams. We removed all 3-grams that are substrings of the Telegram redaction message on copyright violation except the first one.

10-gram	Approximate count
message couldn't be displayed on your device due to copyright	36553503
- - - - - - - - - -	1045726
| gettr | truthsocial | rumble | telegram 🇺🇸 🇮🇱	302400
north atlantic treaty organization by @USERNAME a @USERNAME project -	278543
cant be displayed because it violated telegram's terms of service.	207879
subscribe 👉 https://t.me/USERNAME franksocial | gettr | truthsocial | rumble	185394
help us spread truth! ✝️ 🌎 🇺🇸 main: t.me/USERNAME news: t.me/USERNAME videos:	181626
support team is here to help with your trb product-related	153010
this channel cant be displayed because it violated local laws.	104086
‘’‘’‘’‘’‘’‘’‘’	89937

Table 6: Top-10 10-grams. Many 10-grams discovered by the analysis were shifted substrings of the same message. In such cases, we show only the first occurrence for readability. Text in capital letters marks redacted content.

Category	Domain
Social media	facebook.com, twitter.com, instagram.com, linkedin.com, youtube.com, youtu.be, twitch.tv, reddit.com, redd.it, pinterest.com, tiktok.com, snapchat.com, whatsapp.com, telegram.com, telegram.org, discord.com, viber.com, line.me, kakaotalk.com, wechat.com, qq.com, weibo.com, t.me, t.co, x.com, substack.com, parler.com, rumble.com, gab.com, truthsocial.com, 4chan.org, bitchute.com, gettr.com, dlive.tv, odysee.com, wa.me
Search engines	google.com, bing.com, yahoo.com, duckduckgo.com, baidu.com, yandex.com, goo.gl
E-commerce	amazon.com, amzn.to, amazon.in, paypal.me, paypal.com
Link shorteners	bit.ly, rebrand.ly
Misc.	wordpress.com, lbry.com, minds.com, change.org, gmx.net, soundcloud.com, doi.org, github.com, ift.tt, linktre.ee

Table 7: Blacklist of domains excluded from URL classification.

Figure 10: Topics in Farsi and Arabic chats identified using BERTopic.

Topic id	Topic	Message count	Fraction
-1	political economy russia	711168	0.711168
0	israel gaza conflict	8263	0.008263
1	fires incidents	5402	0.005402
2	cultural events	4767	0.004767
3	road accidents	4580	0.00458
4	social interactions	4377	0.004377
5	books giveaways reading	4166	0.004166
6	personal struggles	3946	0.003946
7	cultural critique	3580	0.00358
8	anonymous messages	3580	0.00358
9	personal journeys	3477	0.003477
10	family conversations	3472	0.003472
11	cats and kittens	3319	0.003319
12	stray dogs management	3264	0.003264
13	healthcare system	3125	0.003125
14	uzbek society	3038	0.003038
15	winter weather	2702	0.002702
16	album music	2516	0.002516
17	interactive experience	2359	0.002359
18	drawing art	2247	0.002247
19	space exploration	2242	0.002242
20	vaccination covid	2136	0.002136
21	battlefront updates	2111	0.002111
22	power outages	1817	0.001817
23	prayer and faith	1809	0.001809
24	official receptions	1757	0.001757
25	armenian azerbaijani conflict	1713	0.001713
26	channel interaction events	1625	0.001625
27	weather forecast	1617	0.001617
28	artificial intelligence	1600	0.0016
29	fraud alerts malware	1543	0.001543
30	job vacancies food service	1536	0.001536
31	sleeping struggles	1504	0.001504
32	humanitarian aid	1497	0.001497
33	corruption convictions	1413	0.001413
34	clothing sales	1384	0.001384
35	military aggression casualties	1333	0.001333
36	outstanding scientists	1303	0.001303
37	sanctions strategy	1260	0.00126
38	china russia relations	1253	0.001253

Table 8: Top-40 Russian topics

Topic id	Topic	Message count	Fraction
-1	social issues	646264	0.646264
0	sports leagues iran	55357	0.055357
1	israel gaza conflict	12168	0.012168
2	music videos	6586	0.006586
3	weather forecast	5953	0.005953
4	covid19 statistics	5242	0.005242
5	automobile pricing	4650	0.00465
6	fashion apparel sales	4375	0.004375
7	higher education admissions	4163	0.004163
8	poetry of love	3478	0.003478
9	elections irán	3298	0.003298
10	religious texts	3069	0.003069
11	love relationships	3046	0.003046
12	real estate apartments	2707	0.002707
13	job opportunities iran	2527	0.002527
14	vaccine distribution	2485	0.002485
15	love expression	2121	0.002121
16	water management	2119	0.002119
17	cultural heritage tourism	2110	0.00211
18	morning love	2102	0.002102
19	nutrition healthcare diabetes	2020	0.00202
20	fire incidents	1926	0.001926
21	electricity consumption	1844	0.001844
22	space exploration	1787	0.001787
23	ukraine russia conflict	1758	0.001758
24	apple iphone updates	1688	0.001688
25	earthquake activity	1648	0.001648
26	kiss and affection	1496	0.001496
27	traffic accidents	1493	0.001493
28	poultry market pricing	1435	0.001435
29	morning prayers	1397	0.001397
30	nuclear negotiations	1328	0.001328
31	cooking recipes	1299	0.001299
32	school closures	1284	0.001284
33	iranian cinema festival	1276	0.001276
34	hijab regulations iran	1192	0.001192
35	film series downloads	1143	0.001143
36	condolences messages	1140	0.00114
37	snow fall winter	1108	0.001108
38	rainy moments	1097	0.001097

Table 9: Top-40 Farsi topics

Topic id	Topic	Message count	Fraction
-1	religious practices	703045	0.703045
0	football matches	13133	0.013133
1	mourning and memorials	10789	0.010789
2	prophetic narrations	7544	0.007544
3	women marriage guidelines	5899	0.005899
4	israeli hamas negotiations	5894	0.005894
5	resistance activities	4820	0.00482
6	ramadan fasting rules	4736	0.004736
7	bot channels access	4353	0.004353
8	gaza shelling	4108	0.004108
9	morning inspiration	3725	0.003725
10	religious guidance	3579	0.003579
11	bot subscription update	3483	0.003483
12	love expressions	3405	0.003405
13	ukraine russia conflict	3157	0.003157
15	telegram channels	3017	0.003017
14	fashion apparel sales	3017	0.003017
16	education exams schedule	2678	0.002678
17	female perspectives	2601	0.002601
18	islamic books collection	2458	0.002458
19	elections and voting	2345	0.002345
20	servant of God	2285	0.002285
21	shabl aljawaber channels	2254	0.002254
22	inner conflict	1903	0.001903
23	martyrs and resistance	1877	0.001877
24	faith in God	1853	0.001853
25	fires incidents	1829	0.001829
26	gaza prayers support	1607	0.001607
27	armed conflicts	1439	0.0014
28	moon and space	1424	0.001424
29	weather forecast	1385	0.001385
30	job openings saudiarabia	1362	0.001362
31	seeking knowledge	1353	0.001353
32	gaza airstrikes victims	1340	0.00134
33	remembrances of god	1318	0.001318
34	apps and updates	1310	0.00131
35	account management	1277	0.001277
36	prisoner exchange	1251	0.001251
37	seeking forgiveness	1186	0.001186
38	sudan political developments	1085	0.001085

Table 10: Top-40 Arabic topics

Topic id	Topic	Message count	Fraction
-1	political conspiracy	644734	0.644734
0	football leagues updates	11075	0.011075
1	vaccine debate	10577	0.010577
2	taiwan china relations	7139	0.007139
3	christian faith	5121	0.005121
4	negative perceptions actor	4718	0.004718
5	trudeau protests canada	4676	0.004676
6	election fraud allegations	4646	0.004646
7	aviation incidents	4098	0.004098
8	trump rallies events	4061	0.004061
9	vaccine documentary watching	3688	0.003688
10	islamic teachings	3610	0.00361
11	biden presidency debate	3597	0.003597
12	female characters analysis	3227	0.003227
13	racial discourse	3204	0.003204
14	elon musk twitter	2868	0.002868
15	wildfires and firefighters	2591	0.002591
16	transgender issues	2513	0.002513
17	brazilian political crisis	2333	0.002333
18	epstein maxwell scandal	2206	0.002206
19	australia freedom movement	2202	0.002202
21	nasa spacex missions	1909	0.001909
20	friends with affection	1909	0.001909
22	antisemitic narratives	1888	0.001888
23	inflation rates policies	1749	0.001749
24	nato news updates	1729	0.001729
25	music expressions	1709	0.001709
26	france political protests	1661	0.001661
27	game releases indie	1590	0.00159
28	mass shootings	1570	0.00157
29	samsung smartphones android	1568	0.001568
30	pope francis death	1505	0.001505
31	drone warfare operations	1489	0.001489
32	anthro canine nsfw	1469	0.001469
33	5g emf radiation	1398	0.001398
34	climate change hoax	1372	0.001372
35	abortion debate	1347	0.001347
36	bjp elections india	1342	0.001342
37	jfk assassination conspiracy	1336	0.001336
38	royal family news	1323	0.001323

Table 11: Top-40 English topics

Chat name	Language	Characteristic topic
INSIDERR_POLITIC	English	brics expansion 2023
vanek_nikolaev	Russian	military air operations
INSIDER_USA_NEWS	English	child protection legislation
vv_volodin	Russian	ukraine crisis mentions
Russica2	Russian	russia geopolitics conflict
dmitrynikotin	Russian	ukraine celebration media
real_DonaldJTrump	English	election fraud allegations
trump_magacommunity	English	justice fight victims
zarubinreporter	Russian	news updates telegram
vatnoeboloto	Russian	poverty in russia
panchenkodi	Russian	ukraine crisis mentions
stalin_gulag	Russian	poverty in russia
slvn_pomet	Russian	ukraine conflict updates
project_veritas	English	undercover investigation
PatriaDigital	Spanish	brics expansion 2023
realKarliBonne	English	trump debate topics
JamesOKeefeIII	English	election fraud allegations
Alertas24	Spanish	protests in Venezuela
realx22report	English	congressional events
PepeMatter	English	child protection legislation
qthestormrider777	English	deep state exposure
CharlieKirk	English	gop election audit
RealGenFlynn	English	veteran honoring events
LauraAbolichannel	English	stock market crisis
TrumpChannel	English	biden hunter laptop
stewpeters	English	military degeneracy issues
RealDonaldoTrumpo	English	political resignation votes
BellumActaNews	English	protests in Venezuela
ResisttheMainstream	English	biden trump transition
ShadowofEzra	English	twitter censorship health
rattletrap1776	English	space force initiatives
DirtRoadDiscussion	English	satanic rituals war
TuckerFans	English	trump election interference
sanidadgob	Spanish	health and wellbeing
DDGeopolitics	English	ukraine conflict updates
bioclandestine	English	freedom fight hope
Jack_Posobiec	English	sexual abuse cases
SantaSurfing	English	taliban takeover afghanistan
trottasilvano	French	green energy sustainability
DBongino	English	dan bongino streaming
liusivaya	Spanish	ukraine crisis mentions
Richardcitizenjournalists	English	news updates telegram
VigilantFox	English	meat health supplements
rsbnetwork	English	trump debate topics
FearlessReport	English	biden hunter laptop
AntiSpiegel	German	eu regulations analysis
SGTnewsNetwork	English	child protection legislation
JFK_Q17	English	truth unleashed
geopolitics_live	English	ukraine air defense
CaptKylePatriots	English	elite crime exposures
cruel_historyy	English	iraq protests violence
Slavyangrad	English	artillery operations
DanScavinoFORCE	English	election fraud allegations
HATSTRUTH	English	spiritual struggles prayer
ReinerFuellmichEnglish	English	twitter censorship health
andweknowLT	English	disinformation alerts

Chat name	Language	Characteristic topic
MichaelJLindell	English	election fraud allegations
IntelRepublic	English	russia european relations
candlesinthenight	English	citizenship exchange service
QNewsOfficialTV	Dutch	meat health supplements
FaktenFriedenFreiheit	German	medical controls criticism
GitmoTV	English	elite crime exposures
ScottRitter	English	citizenship exchange service
JamesWoodsFans	English	biden trump transition
worlddoctorsalliance	English	olivia health issues
stormypatriotjoe21	English	luciferian satanism
projectcamelotKerry	English	military intelligence news
patriotstreetfighter	English	event tour schedule
NewsmaxTV	English	newsmax podcasts
followsthewhiterabbit	English	military air operations
ElectionHQ2024	English	election fraud allegations
RealMarjorieGreene	English	blm antifa violence
AMGNEWS2022	English	elite crime exposures
GarrettZ	English	election fraud cronyism
Qanon_storm_incoming	English	spiritual struggles prayer
robinmg	English	eu regulations analysis
BennyJohnson	English	biden trump transition
police_frequency	English	protests in Venezuela
LizCrokinReport	English	epstein scandal
PrincessDiana_Q	English	biden trump transition
TheTrumpist	English	lebron james crash
BenShapiroGang	English	congressional events
Haintz	German	health and wellbeing
ConservativeBrief	English	biden trump transition
DonaldTrumpOffice	English	multilateral dialogue
NTDNews	English	green energy sustainability
hsretoucher17	English	trump debate topics
zeeemedia	English	injection injuries athletes
TrueGreatAwakening	English	secret control influence
BannonWarRoom	English	eu regulations analysis
WesternJournal	English	lebron james crash
techno_fog	English	epstein scandal
liltalkshow	English	whistleblower dominion bios
TheRealKimShady7	English	luciferian satanism
PepeDeluxed	English	democracy peril corporation
theprofessorsrecord	English	election fraud allegations
WhiteHatsQ	English	deep state exposure
wendyrogersaz	English	senate legislation minors
DrDavidMartin	English	democracy peril corporation
jordansather	English	content moderation

Table 12: Initial seed list of Telegram channels used to bootstrap the snowball crawler. We picked the 100 biggest channels in the political category according to https://telegramchannels.me/ranking (retrieved on 2025-05-15). The characteristic topic is the topic that is most overrepresented in messages from the given chat, compared to the baseline of messages from all seed chats.

Algorithm 1: Snowball crawling of Telegram channels. The algorithm begins with an initial channel, extracts forwarded messages to discover new channels, and repeats until no new channels are found. Distribution across multiple workers and queue prioritization by out-degree are omitted for clarity.

Input: Initial channel C_{initial}
Output: List of discovered Telegram channels

$q \leftarrow \{C_{\text{initial}}\}$
while q is not empty **do**
  $C_{\text{current}} \leftarrow \text{Dequeue}(q)$
  $m \leftarrow \text{GetMessages}(C_{\text{current}})$
  $m_f \leftarrow \text{GetForwardedMessages}(m)$
  $C_{\text{new}} \leftarrow \text{DiscoverChannels}(m_f)$
  $\text{Enqueue}(q, C_{\text{new}})$
end while