TGDataset: Collecting and Exploring the Largest Telegram Channels Dataset

Massimo La Morgia
Sapienza University of Rome,
Department of Computer Science
Rome, Italy
lamorgia@di.uniroma1.it

Alessandro Mei
Sapienza University of Rome,
Department of Computer Science
Rome, Italy
mei@di.uniroma1.it

Alberto Maria Mongardini
Sapienza University of Rome,
Department of Computer Science
Rome, Italy
mongardini@di.uniroma1.it

Abstract

Telegram is a widely adopted instant messaging platform. It has become worldwide popular because of its emphasis on privacy and its social network features such as channels—virtual rooms in which only the admins can post and broadcast messages to all the subscribers. Channels are used to deliver live updates (*e.g.,* weather alerts) and content to a large audience (*e.g.,* COVID-19 announcements) but unfortunately also to disseminate radical ideologies and coordinate attacks such as the Capitol Hill riot.

This paper introduces the TGDataset, the most extensive publicly available collection of Telegram channels, comprising 120,979 channels and over 400 million messages. We outline the data collection process and provide a comprehensive overview of the data set. Using language detection, we identify the predominant languages within the dataset. We then focus on English channels, employing topic modeling to analyze the subjects they cover. Finally, we discuss some use cases in which our dataset can be instrumental in understanding the Telegram ecosystem and studying the diffusion of questionable news. Alongside the raw dataset, we release the scripts used in our analysis, as well as a list of channels associated with a novel conspiracy theory known as Sabmyk.

CCS Concepts

• **Information systems** → *Information retrieval*; **Social networks**.

Keywords

Dataset, Telegram, Conspiracy Theories, Copyright Infringement

1 Introduction

In today's digital age, instant messaging apps have become ubiquitous, and Telegram stands out as one of the leading platforms. Its focus on user privacy is a significant factor behind its growing popularity [10]. However, as is often the case with platforms that prioritize user privacy, this approach has also attracted malicious users who exploit the platform for illegal activities. While Telegram serves legitimate purposes, such as disseminating weather alerts or government updates [42], it is also misused for spreading radical ideologies [3], orchestrating violent attacks [1], and engaging in market manipulations, such as pump-and-dump schemes [41]. These dual uses underscore Telegram's complex nature and highlight the need for a deeper exploration of its ecosystem.

In this study, we introduce and publicly release [38] the TG-Dataset, a collection of public Telegram channels of over 120,000 channels and 400 million messages. To the best of our knowledge, the TGDataset surpasses existing datasets in scale and scope, offering an unprecedented opportunity to study the Telegram ecosystem comprehensively. Unlike prior datasets focused on specific topics or geographic regions, the TGDataset includes diverse channels covering multiple languages and subjects, enabling a more holistic understanding of the platform. In particular, the TGDataset enables the study of the platform's political leanings and the sharing of questionable content, topics widely explored on other Online Social Networks (OSNs) but not on Telegram. It also includes numerous channels spreading conspiracy theories, as well as those engaged in borderline or illegal activities like carding, inciting violence, and promoting Nazi ideologies. Thus, with its scale and diversity, the TGDataset represents a valuable resource for researchers willing to investigate and shed light on both the legitimate and problematic aspects of Telegram's ecosystem.

The main contributions of this work are the following:

- We present the largest publicly available dataset of Telegram channels, enabling researchers to explore various phenomena on the platform. We perform analyses such as language detection and topic modeling to characterize the dataset's contents, revealing dominant languages, prevalent topics, and temporal trends in channel creation.
- We explore several potential use cases of the TGDataset. One application involves analyzing the diffusion of questionable content and biases across Telegram channels, providing insights into content moderation and information reliability. Another use case focuses on investigating networks that promote conspiracy theories, including emerging phenomena such as the Sabmyk network. Additionally, the dataset can be used to investigate channels engaged in borderline or illegal activities, such as carding, copyright infringement, and extremist propaganda.
- Alongside the dataset, we release scripts for analysis, language detection, and topic modeling, fostering reproducibility and facilitating further research.
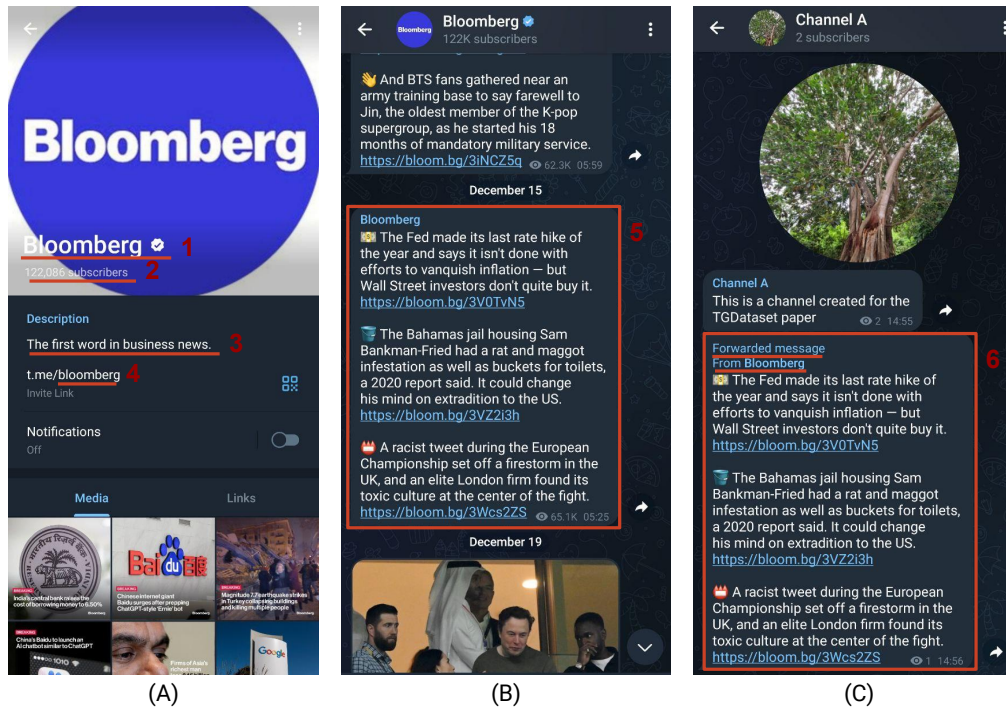
Figure 1: Fig. (A) represents the main information about a channel: the title (1), the number of subscribers (2), the description (3), and the username (4). Fig. (B) shows the posts published by the channel (5). Fig. (C) displays a forwarded message (6).

2 Background: Telegram

Started in 2013, Telegram is now a popular instant messaging platform, with more than 500 million active users in 2021 [7]. Telegram allows users to post content such as text messages, images, audio, stickers, videos, and files weighing up to 2 GB. In addition to private chats (one-to-one communication), Telegram provides two other ways to communicate: Through groups and channels.

Groups: Designed for many-to-many messaging, groups allow up to 200,000 members to communicate with each other by posting content. Groups are identified by a unique username and have a description and a title. They can be public (*i.e.,* can be found with Telegram search, and every user can read its content and join the group) or private (*i.e.,* only members can see the posts, and users need an invite link to join it). Besides, a group presents a list of its members which, if the group is public, is visible to anyone (even non-members).

Channels: They offer one-to-many communication. Thus, differently from groups where every member can post messages, in channels, only administrators can publish posts. Channels have a unique username, a title, a description, and can be private or public. Fig. 1 presents the information displayed by a channel. Contrary to groups, a channel can have an unlimited number of subscribers and does not display any users' information (*i.e.,* nicknames) other than the number of its subscribers. Thanks to these features, Telegram channels have become popular for disseminating content to a large audience. Indeed, several public figures and companies started using official Telegram channels to share news and updates [42]. Nonetheless, this tendency caused an increase in the number of

channels on the platform seeking to impersonate official channels or sell counterfeit products and services.

Scam and Verified mark: To tackle this issue, Telegram has implemented the *verified* and the *scam* marks. The verified mark requires that channels, groups, and bots prove their verified status on at least two social media platforms(*e.g.,* TikTok, Facebook, Twitter, Instagram) [9]. In contrast, the scam status is assigned to channels and groups reported for fraud by multiple users [52].

The forward mechanism: Although private chats, groups, and channels can not communicate directly among them, they are not isolated but may be linked through message forwarding. Leveraging this feature, users and channels can forward messages to other users, groups, or channels displaying the post's original author. Fig. 1 shows an example of message forwarding, where the admin of "Channel A" (Fig. 1 (C)) forwards in his channel a message from the Bloomberg's channel. The message in the red square of Fig. 1 (B)) is the original message, while the one in the red square of Fig. 1 (C) is the forwarded message. As shown in Fig. 1, the forwarded message displays the name of the original channel of the post, which, if clicked, leads to that channel.

3 Related works

Several studies have been conducted on the Telegram platform and related issues. Hashemi et al. [27] collected Telegram Iranian channels and groups to distinguish high-quality groups (*e.g.,* business groups) from low-quality groups (*e.g.,* dating groups). They found that high-quality groups had longer messages and more user engagement. Jalilvand et al. [34] tackled the challenge of ranking

Table 1: Summary of the number of channels, channel messages, and group messages present in both the TGDataset and Pushshift.

Dataset	# channels	# channels msgs	Update to
Nobari et al. [20]	2,556	36,928	October 2016
Pushshift [16]	27,801	192,044,689	November 2019
TGDataset	120,979	498,320,597	July 2022

channels related to a user request on Telegram. Ng et al. [50] analyzed a Singapore-based COVID-19 Telegram group with over 10,000 participants, examining group opinions over time. Their work revealed a peak in engagement when the Ministry of Health raised the disease alert level but a subsequent decline in participation. Nobari et al. [20] performed a structural and topical analysis of Telegram messages, building a dataset of over 2,500 channels and 54 groups and constructing a graph based on mentions. This study found that the PageRank algorithm was not effective in identifying high-quality Telegram channels. Analyzing the dataset released by Nobari et al., the TGDataset contains more than ten times the number of channels and is significantly more recent (July 2022 versus October 2016).

Similar to our work, Baumgartner et al. [16] created the Pushshift Telegram dataset, containing both channels and groups. They collected over 27,800 channels and 317 million messages from 2.2 million unique users updated to November 2019. Differently from [16], the TGDataset boasts a more recent dataset, with the message history of channels aligned to July 2022, thus almost three years of more data. Moreover, while the Pushshift dataset contains both channels and groups, the TGDaset is made of only channels. Thus, narrowing the comparison on the Telegram channels, the TGDataset includes a significantly higher number of channels (approximately four times more) and a greater volume of messages posted in channels (more than double). Table 1 provides the exact number of channels and messages contained in the dataset released by [20], Pushshift, and TGDataset. Finally, another important difference between the two datasets is about the seed choice. Indeed, the Pushshift dataset used as seed channels only those of right-wing extremist politics or cryptocurrencies-related channels, while our dataset starts the collection from seed channels covering heterogeneous topics (see Sec. 4.1). This choice led to building a dataset that should better represent the status of the Telegram channels ecosystem.

Weerasinghe et al. [63] found that Telegram hosts organized groups called "pods" where members boost each other's Instagram account popularity through interaction. Other studies have documented the presence of Telegram channels and groups focused on cryptocurrency pump and dump scams [41, 65]. The work of Urman et al. [61] explored far-right networks on Telegram, revealing that the most prevalent communities are those related to 4chan [12] and Donald Trump's supporters. Moreover, the authors pointed out that the sudden growth of these networks on Telegram corresponds with the bans of numerous far-right figures on other online social networks. Additionally, numerous studies have investigated the
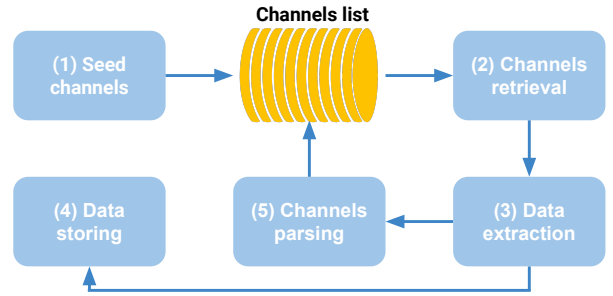

Figure 2: Data collection flow diagram.

use of Telegram by terrorist organizations such as ISIS for content dissemination and proselytism [18, 66].

4 Data collection

Previous Telegram datasets have been created with specific research goals in mind, resulting in a limited scope of channels related to a particular language, country, or topic. For example, [27] only includes Iranian channels, [32] only contains 151 channels related to QAnon, and the PushShift dataset focuses on channels discussing right-wing extremism and cryptocurrency. In contrast, our dataset aims to provide a global snapshot of the Telegram ecosystem. To achieve this goal, we require a dataset that provides an up-to-date representation of Telegram and encompasses a broad range of popular and interconnected channels. This is the motivation behind the creation of the TGDataset.

4.1 Dataset construction

To build the TGDataset, we use a snowball approach as previously done in [16]. We begin with a set of seed channels related to various topics and then extend our dataset by including the source channel of each forwarded message. Fig. 2 represents the flow diagram we use to collect the data. To determine the seed channels (Step 1 in Fig. 2), we utilize Tgstat [6], a popular freemium service that indexes over 150,000 Telegram channels and gathers statistics about them. Tgstat provides information for each channel, such as the subscriber count, topic category, growth rate, and language. Among the other metrics, Tgstat lists the top 100 channels by the number of subscribers. From this rank, we retrieve all the categories these channels belong to, finding the following 18 categories:: *Sales*, *Humor and entertainment*, *News and Mass media*, *Video & Movies*, *Business & Startups*, *Cryptocurrencies*, *Politics*, *Technologies*, *Sport*, *Marketing*, *Economics*, *Games*, *Religion*, *Software and Applications*, *Lifehacks*, *Fashion & Beauty*, *Medicine*, and *Adults*. Finally, we selected the ten most popular channels from each category as the seed channels according to their subscriber count. With this approach, we collect 180 seed channels.

For each seed channel, we proceed with downloading all its messages by leveraging the Telethon APIs [8] (Step (2)), an open-source Python tool that provides access to the official Telegram APIs. Specifically, each call to download messages from a channel is performed five seconds after the previous one to avoid flooding the Telegram service. Once we retrieved the data about the channel, we proceeded with the data extraction (Step (3)): in particular, we

Field name	Description
Channel ID	the ID of Telegram channel
Creation date	the timestamp of the creation date of the channel
Username	the Telegram username of the channel
Title	the title of the channel
Description	the description of the channel
Scam	indicates if Telegram marked the channel as a scam
Verified	indicates if Telegram marked the channel as verified
# subscribers	the number of subscribers of the channel
Text messages	
Message	the text of the message
Date	the timestamp of the message publication
Author	the ID of who posted the message
Is forwarded	indicates if the message is forwarded
Forwarded from	the ID from which the message is forwarded
Original date	the timestamp of the first post of the message
Media messages	
Title	the title of the content
Media ID	the ID of the content on Telegram
Date	the timestamp of the content
Author	the ID of who posted the content (int)
Extension	the format of the content (string)
Is forwarded	indicates if the content is forwarded
Forwarded from	the ID from which the content is forwarded
Original date	the timestamp of the first post of the content

Table 2: Information stored about collected channels

retained all the information obtained by Telegram API, excluding media files such as images, videos, or PDFs. The reason behind this decision is to avoid storing copyrighted or illegal content. Anyway, even for media-based messages, we collect the name, author, and ID of the media content using the Telegram API, and we infer the file format based on the extension of the file name. Next, we store the extracted information in MongoDB [48], a NoSQL database (Step (4)). Finally, we parse the channel messages to identify forwarded messages and their original authors (Step (5)). If the author represents a channel that we have never seen, then the channel is added to the list of new channels. Conversely, if the author is a user or represents a group, we ignore the author of the forwarded message. To iteratively expand the TGDataset, we repeat the above procedure of channel discovery (Steps 2, 3, 4, and 5) using the newly found channels as new seeds. We stopped the iteration process on July 31, 2022. Finally, to align our dataset, for each discovered channel, we download all the missing messages until July 31, 2022.

For each channel, we store the following information as listed in Tab. 2: title and if it is marked as verified or scam (item 1 in Fig. 1 (A)), subscriber count (item 2), description (item 3), username (item 4), ID, and creation date. For the messages, we retain the author (the name of the channel), timestamp, and, in the case of forwarded messages, original posting date and original channel.

4.2 Accessing the TGDataset and FAIR principles

We released the TGDataset in alignment with the FAIR (Findable, Accessible, Interoperable, Reusable) principles [64].

Findable. We publicly released the TGDataset [37] on Zenodo [24], an open repository managed by CERN.

Accessible. The dataset is made of 121 files, each of which contains a maximum of 1,000 channels. To ease the retrieval of single channels from the dataset, we store them in alphabetical order. In addition, each file's name describes the index of the channels it contains. Given the large dimension of the dataset, even if compressed (approx. 71GB), we divided the dataset into four parts. In this way, it is possible to download and explore also a single portion of the dataset. Moreover, we released a sample of the dataset on a public GitHub repository [38]. This enables users to explore the dataset's structure without the need to download any portion of it.

Interoperable. All the information is encoded in JSON format so that the data can be easily parsed and manipulated with most programming languages.

Reusable. We have outlined the methodology employed for data collection, and the data is open for anyone to use. We also released on the public GitHub repository [38], Python scripts to load the data into a MongoDB [15] database, as well as scripts that show how to perform basic queries on the database (such as inserting new channels or retrieving channels by username or ID). Furthermore, these scripts can also replicate some of the analysis that we performed in this work, including Language Detection of the channels (Section 5.1) and Topic Modeling (Section 5.2). Finally, we released as CSV files the list of channels belonging to the Sabmyk network and the list of other channels promoting conspiracy theories (Sec. 6), the list of channels annotated with the detected writing language, and the one containing the mapping of inferred topics.

4.3 Dataset overview

The data collection for the TGDataset began on 4 January 2021 and ended on 31 July 2022. The dataset has a total size of about 460 GB and includes 498,320,597 messages and 120,979 unique channels. Out of these channels, 654 (0.54%) have verified status, while 183 (0.15%) are flagged as scam channels. For the purpose of clarity, we will refer to the remaining 120,142 channels, which are neither scams nor verified, as *standard channels*.

To provide a general characterization of the channels contained in the TGDataset, in Fig. 3, we show the CDFs of the number of subscribers, posts published, and the ratio of forwarded messages for verified, scam, and standard channels. As shown in Fig. 3(a), verified channels (green line), in general, have more subscribers than scam (dashed orange line) and standard channels (dotted blue line). Indeed, the verified channels with more than 1,000 subscribers are 98.32%, while the scams and standard are 68.85% and 38.11%, respectively. Concerning the number of messages published, Fig. 3(b) suggests that verified channels tend to post more content than scams and standard ones. The standard channels with at least 1,000 messages are 57.16%, those verified are 83.79%, and the scams are 50.82%. Moreover, we report the extent of use of the forwarding function among channels of the TGDataset. Fig. 3(c) reveals message forwarding is more prevalent in standard and scam channels. The standard channels with at least a 0.1 ratio of their messages that are forwarded are 33.5%, whereas the scams and the verified ones are 32.24% and 13.15%, respectively.

(a) Subscribers

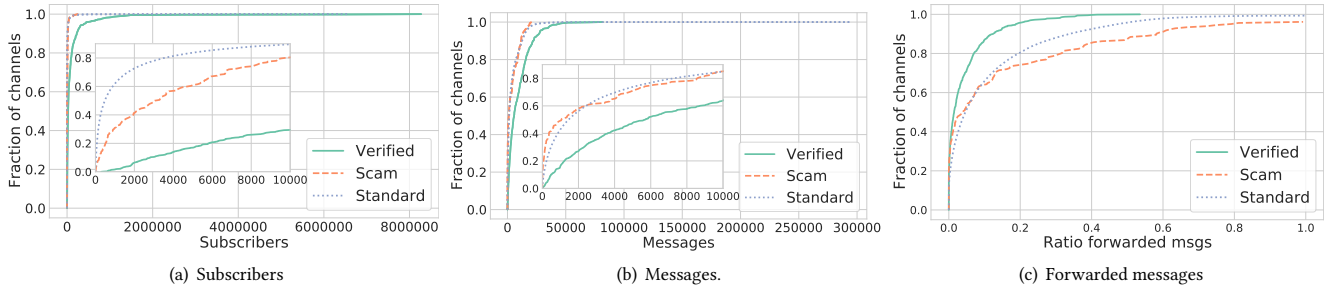(b) Messages.

(c) Forwarded messages

Figure 3: CDFs of the number of subscribers for scam, verified, and standard channels (3(a)), the number of messages posted (3(b)), and ratio of forwarded messages (3(c)).

5 Dataset analysis

In this section, we first study the languages spoken within our dataset, leveraging language detection, and explore the English channels to understand the covered topics using topic modeling. Then, we analyze the temporal aspects of the TGDataset.

5.1 Language detection

As a first step in our analysis, we want to understand the language coverage of our dataset. This study gives hints about the popularity of the Telegram platform in different regions of the world. To automatically retrieve the language used in a channel, we leverage Language Detection [54], a language detection library implemented in Java by Google with over 99% of precision for 53 languages. To get meaningful results dealing with text-based messages, first, the text has to be normalized and polished from useless information [62]. Hence, we applied the following preprocessing steps. First, we excluded messages shorter than 15 characters, as they do not provide sufficient information for accurate analysis [14]. Next, we utilized the RegexpTokenizer developed by NLTK [5] to split text-based messages into tokens. Then, we removed all numbers and emojis from the documents, as they do not contribute to language identification. Also, we discarded references and mentions of groups, channels, or users, as usernames typically do not aid in identifying the language used.

We assign a channel to a specific language only if over 50% of its messages are written in that language. Otherwise, we will not associate a language with that channel. At the end of the analysis, we find that the five most spoken languages in our dataset are Russian with 42,983 channels (35.52%), English with 19,768 channels (16.34%), Farsi with 16,779 channels (13.86%), German 4,950 (4.09%), and Arabic with 2,523 (2.08%). Of particular interest is the strong presence of the Russian and Farsi channels, which likely reflects the popularity of Telegram in Russia and Iran, despite Russia's government banned Telegram from 2018 to 2020 [4], and it is still banned in Iran [13]. Although Iran banned Telegram on 30 April 2018 [2], in our dataset, there are 4,635 out of 16,779 (27.62%) channels in Farsi created after the date of the ban. Instead, looking at the last activity of Farsi's channels, we find that more than 66.53% (11,164) channels continued to operate after the ban. Regarding Russian channels, as shown in section Sec. 5.3, they continue to appear also after the ban of 2018.

5.2 Topic modeling

In this subsection, we investigate the topics covered by the channels in our TGDataset using Topic Modeling [30]. This is a data mining tool that allows finding a brief description of the topic addressed by the messages of a channel. For this analysis, we consider only channels that post English content. Thus, we use as input to the topic modeling 19,768 channels (16.34% of our dataset). As in language detection, a preprocessing phase is necessary. In addition to the deletion of numbers, deletion of emoji, and normalization, we perform other preprocessing steps. We exclude words with one or two characters, as they are common in messages (e.g., y, no, ok) and do not provide meaningful information about the topic. We use the *en_core_web_sm* model developed by spaCy [31] to reduce inflected words to their lemmas, ensuring that different forms of a word are treated as a single term. Lastly, we employ the English stopwords list provided by NLTK [51] to discard common words (e.g., articles and prepositions) that do not carry significant meaning in a sentence [53].

To discover the latent topics addressed within the channels, we use the Latent Dirichlet Allocation (LDA) [17] as the Topic Modeling algorithm. LDA needs as input the number of topics, so we used the UMass measure [47] to select the optimal one. In particular, the higher the coherence of the words representing topics, the closer to 0 the value of UMass. In our case, we calculate the best UMass value reached by selecting the number of topics from 10 to 30.

The best UMass value (-0.69) is obtained with 13 topics. Tab. 3 shows the inferred topics and the top 10 keywords for each of them. Next, we group the channels according to the related topics using LDA. Tab. 5 lists the topics discovered in the TGDataset and the number of channels for each. These emerged topics are quite different from those covered by the seed channels (see Sec. 4). Indeed, some disappeared (*e.g.,* Psychology, Marketing), while other interesting topics came up. This shift in topic distribution is due to the snowball approach, which adds new channels linked by message forwarding to the dataset. As a result, channels from larger, more interconnected networks with higher forwarding activity are more likely to be included. This method naturally shifts the dataset toward more viral or widely shared topics.

Among the new topics, one is related to carding, the practice of selling full details of stolen credit cards or selling prepaid cards or other goods purchased with them. Similarly to what happens

Topic	Top 10 keywords
Video-game modding	pubg, esp, lvl, max, mod, login, password, aimbot, vip, recoil
Carding	iphone, premium, netflix, amazon, samsung, paytm, giveaway, hacking, tutorial, vpn
Entertainment	reddit, artist, submit, edition, meta, til, hire, score, anime, league
Indian education	indian, upsc, exam, affair, copyright, infringement, unavailable, wildlife, prelim, batch
Religion	jesus, lord, christ, spirit, pray, holy, shall, church, prayer, bless
US news	trump, biden, vaccine, election, covid, joe, donald, court, military, mandate
Social	twitter, reuter, tweet, album, telegram, ivurl, utc, radio, hashtag, youtube
World News	russia, ukraine, independent, minister, military, coronavirus, europe, foreign, israel, court
Software	android, rupee, enroll, web, udemy, proxy, linux, software, premium, mod
COVID-19	vaccine, covid19, covid, medical, vaccination, vaccinate, pfizer, disease, patient, australia
Crypto	btc, bitcoin, crypto, usdt, binance, usd, token, trading, blockchain, profit
Extremists and radicals	violate, defense, jews, jewish, attacker, defend, jew, defender, hitler, antifa
Adult content	pornographic, leak, t.me, xxx, meme, wanna, sharp, iphone, porn, teen

Table 3: Top 10 keywords within LDA topics.

in dark web forums [35], carders (the people who own the stolen credit cards) use Telegram channels to place gift cards or goods for sale. In the TGDataset, we find 1,489 channels (7.53% of the English part of the dataset) offering this service. In particular, Telegram marked 45 carding channels as scams, suggesting that some of them do not deliver the service they offer.

Another unusual cluster containing 989 channels (5% of English channels) is the one about extremists and radicals. Here, several channels promote white supremacy, as well as other Nazi ideologies and conspiracies against the white race (the title of one of these channels is *White Genocide Immigration Anti-White Agenda*). Interestingly, there are four channels in this category having the verified status. One channel (*"ISIS Watch"*) just publishes daily updates on banned terrorist content. Instead, the other three channels are related to official news agencies of the Russian government and are no longer accessible on Telegram. Moreover, some of their messages we collected are obscured by the platform. To investigate this aspect deeper, we analyze the messages obscured by Telegram as they incited violence, published illegal pornographic content, or shared content protected by copyright. In this case, Telegram replaced some or all of the channel's messages with text explaining the reasons for obfuscation. Interestingly, the channel itself and the metadata (*e.g.,* the posting date) related to the original messages are still available. As shown in Tab. 4, the service message most used by Telegram to obscure a post is the one about the violation of Telegram's Terms of Service (428,793 messages). Follow the one about the spread of pornographic content (263,643 messages), the
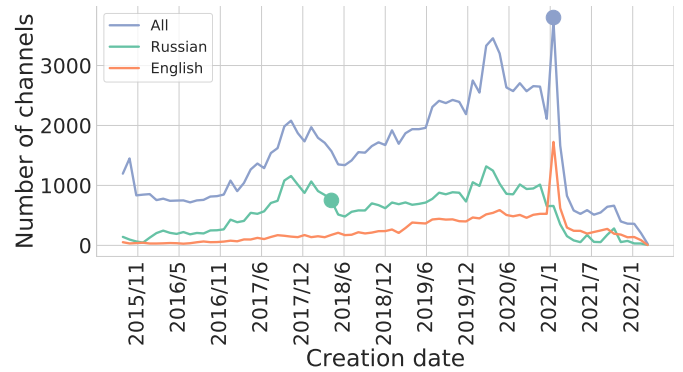

Figure 4: Creation date of Russian, English, and all the channels. The green dot represents the date when Russia banned Telegram, while the azure one indicates WhatsApp's announcement about privacy policy.

violation of local laws (145,448 messages), and copyright infringement (48,402 messages). Finally, the other topics are more aligned with the ones covered by the seed channel of the TGDataset.

5.3 Temporal aspect

Another interesting aspect is the creation date of the collected channels since, from this information, it is possible to understand how Telegram evolved through time. Fig. 4 shows the monthly creation rate of the channels contained in the TGDataset. In particular, the blue line represents the overall number of channels created on a certain date, while the green and the orange lines the number of Russian and English channels, respectively. As we can see, the number of channels created is steadily increasing, indicating that the Telegram platform is becoming more popular daily. Indeed, also many governments choose to use Telegram channels to communicate information regarding COVID-19 [42]. Until the beginning of 2018, the growth of channels is led by Russian-speaking users. However, in the first months of 2018, it is possible to note a sudden drop in daily created channels, likely due to the Russian ban of Telegram (green dot in Fig. 4). Nonetheless, Russian channels, even if at a lower rate, continue to appear every day. Finally, in the first months of 2021, there was an abrupt increase of English-language spoken channels that overtook the Russian ones for the first time in Telegram history, according to our data. This event coincides with the change of WhatsApp's privacy policy (azure dot in Fig. 4) and the migration of many users from WhatsApp to Telegram [11].

6 Use cases

The TGDataset has already been utilized in various studies [33, 39, 40]. For instance, La Morgia et al. [39] analyzed the problem of fake accounts on Telegram using the TGDataset, while Imperati et al. [33] employed the TGDataset to identify 17,829 channels disseminating conspiracy-related content. In the following, this section describes other scenarios that can be explored and analyzed using the TGDataset.

Political leaning and presence of questionable contents. Previous studies show the presence of moderation in OSNs (*e.g.,*

Message description	Number of messages	Number of channels
This channel can't be displayed because it violated Telegram's Terms of Service.	428,793	2,546
This channel can't be displayed because it was used to spread pornographic content.	263,643	969
This channel can't be displayed because it violated local laws.	145,448	847
This channel is unavailable due to copyright infringement.	48,402	1,633

Table 4: Number of posts obscured by each service message and number of channels presenting a specific service message.

Table 5: Number of channels, scams, and verified ones belonging to each discovered topic.

Topic	# channels	# scam	# verified
Religion	4,725 (23.90%)	0	5
US news	2,948 (14.91%)	38	51
Video-game modding	1,957 (9.90%)	14	0
Covid	1,716 (8.68%)	0	5
Carding	1,489 (7.53%)	45	0
Entertainment	1,440 (7.28%)	0	3
World news	995 (5.03%)	0	17
Extremists and radicals	989 (5.00%)	0	4
Indian education	939 (4.75%)	1	6
Software	871 (4.41%)	2	14
Porn	830 (4.20%)	5	0
Crypto	563 (2.85%)	5	10
Social	306 (1.55%)	0	4

Twitter) produces a significant reduction of questionable content, with respect to not moderated OSNs (*e.g.,* Gab) [21], and that the information diffusion is biased by the political leaning of the community [19]. However, very little is known about the loosely moderated Telegram platform. Thus, we use the TGDatset to perform the first high-level investigation on Telegram. To this end, as done in [21, 22], we leverage Media Bias/Fact Check (MBFC) [45], an independent fact-checking organization that rates the bias of news outlets. In particular, MBFC classifies the political bias of news outlets with one of the following labels: *Extreme Left, Left, Left-Center, Least-Biased, Right-Center, Right, Extreme Right*. Additionally, MBFC provides a second label to quantify the reliability of the outlet, categorizing outlets as *Reliable* and *Questionable*.

For each English channel in the TGDatset, we label the links it contains according to the Media Bias/Fact Check (MBFC) categorization. We then assigned each channel the most frequently chosen label among its categorized links. This approach classifies 9,045 channels out of 19,768; the remaining channels did not share news or contain sources not present in MBFC. Upon aggregating channels into left-leaning (Left-Center, Left, Extreme Left) and right-leaning (Extreme Right, Right, Right-Center) groups, we found a nearly balanced distribution: 4,491 channels shared right-leaning news, and 3,900 shared left-leaning news. However, 2,748 channels were identified as spreading extreme right content, while none were found to be spreading extreme left content. Among the top four topics with the most categorized channels (World News, US News, Religion, and Covid), World News is the most left-leaning, with over 60% of the channels associated with left-leaning outlets. Channels on the topic of Religion were almost evenly split between left

and right. Conversely, channels dealing with US News and Covid were predominantly right-leaning. Specifically, the extreme right was underrepresented in World News (76 channels) but predominant in Religion (590 channels), US News (1,144 channels), and Covid (601 channels). These three topics—Religion, US News, and Covid—also had the highest number of channels sharing content from questionable sources, with 1,353, 2,320, and 1,277 channels, respectively. Analyzing the reasons behind their questionable classification, we find that 55% of Religion, 80% of US News, and 85% of Covid questionable channels shared content from outlets associated with conspiracy theories by MBFC.

This analysis reveals a significant presence of Telegram channels disseminating news from unreliable sources. Thus, the TGDataset could be a valuable resource for researchers aiming to understand the spread of misinformation, fake news, and conspiracy theories. This use case involves using the content of the text messages posted by the channels (represented by the *Message* field) to extract links for further processing, and the forwarding information, including whether the message was forwarded and its source (contained in the *Is Forwarded* and *Forwarded From* fields).

Study of channels spreading conspiracy theories. In the previous use case, we discovered that there are channels that spread news related to conspiracy theories. Here, we attempt to identify them through community detection. A community in a graph is a subset of nodes that are densely connected to each other and weakly connected to nodes in other communities.

For this study, we represent our dataset as a directed graph $G = (V, E)$ in which nodes in V are the channels, and edge $u \rightarrow v$ in E represents the presence in channel u of a message originally posted in v and forwarded to u by the admin of channel u. Since the users of channel u can navigate the forwarded message and land on channel v, the edge represents in a natural way the possible flow through channels of users following forwarded messages. To build and analyze our graph, we use the NetworkX library [26].

Then, we use the Leiden algorithm [60], an algorithm for community detection that improves the Louvain algorithm [25]. To ensure the accuracy of our partition, we used modularity, a validation metric that measures how much better our partitioning is compared to a random partition. Thus, we compute the optimal number of communities with respect to modularity, achieving a high score of 0.78, which indicates a strong community structure remarkably better than random partitioning. This approach finds a partition with 311 communities, 47 of which with more than one node. One of the communities, which consists of 236 channels and where English is the common language, has the peculiarity that all posts and messages forwarded from one channel to the other are about a new conspiracy theory called Sabmyk. Sabmik is a conspiracy theory

that proposes itself as a better alternative to QAnon. It promotes a quasi-religion centered around a messianic figure known as Sabmyk [56]. This community of channels has, collectively, more than 1 million subscribers. The most popular channel in the community is *Great Awakening Channel* with 119,103 subscribers. According to our data, most of these channels were created at the beginning of 2021—130 of them (55.08%) between January and February—with the latest in our dataset created in February 2022. The nature of these channels is very different: Some of them are fake channels of celebrities (*e.g.,* Mel Gibson, Keanu Reeves, Kanye West), while others target news outlets or official channels of national bodies (*e.g.,* U. S. Marines Channel, U. S. Navy Channel) and others conspiracy theories like QAnon. Regardless of the name of the channel in the community, the content is always the same. Indeed, they recycle and forward messages among them: Out of the 1,203,986 messages published by these 236 channels, only 65,602 are unique. With this analysis we scratch the surface of the Sabmyk network. Still, more investigation are needed to understand how Sabmyk lures such a large number of subscribers, its ultimate goal, and its connection to other conspiracy theories.

Performing this study requires several key elements of the dataset. By analyzing the content of the *Message* field, it is possible to identify the topics within conspiracy theories. With the *Is Forwarded* field, we can understand if the message has been forwarded, and in the case it is, which was the source channel leveraging the *Forwarded From* field. This information makes it possible to build a graph representing connections between channels and to detect the communities as we did for our analysis.

Carding and underground markets. In Section 5.2, we examine the main topics within our dataset. Upon investigating the channels associated with these topics, we discovered several engaging in borderline activities, such as running underground marketplaces. These channels sell electronics, Netflix accounts, and hacking tools for low prices. These types of marketplaces are often associated with carding. Consequently, it is worth delving further into these Telegram channels to determine the types of goods being sold, their origins, whether the channels are just an attempt to fraud subscribers or if they are actually engaged in more illegal activities, and whether there is any correlation with Dark Web marketplaces.

For this analysis, the *Message* field is crucial for identifying the types of goods sold within the channels, such as electronics, Netflix accounts, and hacking tools. Analyzing the content of the messages and extracting the URLs makes it possible to determine whether the channels are involved in fraudulent activities or illegal transactions.

Copyright Infringement and personal content. Recently, Telegram hit the news several times for its use for the distribution of copyrighted content (*e.g.,* movies and software). For instance, Italian authorities seized 545 channels for copyright infringement [44]. However, the issue extends beyond copyright violations. Indeed, the platform is frequently misused to distribute personal content without consent, such as the unauthorized distribution and sale of an Indian teacher's course material [55]. Even more troubling, Telegram is also used to share leaked nude photos [43] or facilitate revenge porn [36]. Joining the information about the topic of the channels, and the description of the messages removed by the platform (Tab. 4), we can observe that the four categories in which is most present the diffusion of copyrighted material are: Carding

(448 channels), Indian Edu (219 channels), Video-game modding (208 channels), Adult content (157 channels). Starting from these channels, it is possible to study the diffusion of the phenomenon, understand how these channels are organized, and how they monetize. For this analysis, the *Message* field is essential for identifying illicit content, such as copyrighted material, while the *Description* field could provide additional context about channel activities.

Cryptocurrencies frauds. We also found several (563) channels related to the cryptocurrency world. It is well known that Telegram channels are exploited by Pump and Dump groups [49, 65] that perpetrate frauds on the crypto-market. We notice that some of the Pump and Dump channels monitored in [41] are also in our dataset. Previous works monitored the channels vertically, focusing on the fraud and the groups' mechanics. Thus, using the *Message*, *Forwarded from*, and *Description* fields, the TGDatast can help understand how these channels are connected, if they operate together, how they promote their services, and likely discover new channels that carry out similar activities.

Violence and extremism. According to several newspapers, the Capitol Hill riot of January 2021 was planned months before also leveraging Telegram channels and groups [29, 46]. In response to this event, the platform has intensified its monitoring and obscured dozens of public channels promoting calls to violence [28]. Nonetheless, despite the commitment of Telegram to obfuscate these channels, within our dataset, there are several public channels promoting the spread of neo-Nazi ideologies or calls to violence (*e.g., White Aryan Woman*, *Feuerkrieg Division **OFFICIAL***). Therefore, the problem is still far from being solved. The TGDataset can help study and characterize these channels. Indeed, the main problem with these channels is the absence of a dataset containing them, as they are obscured or blocked once discovered. Instead, the TGDataset includes several unblurred channels of this typology that can be used to analyze their features and build a machine-learning model able to detect them automatically. This study requires the *Message* and *Description* fields to analyze harmful content and obtain context on the purposes of these channels, and the *Forwarded from* field to trace the channel behind it.

Temporal analyses and information propagation. The TGDataset represents a valuable resource for exploring the temporal dynamics of information spread by leveraging message timestamps and channels' creation dates. Message timestamps (*Date* field in Text messages and Media messages) facilitate granular analyses of dissemination patterns, enabling researchers to track how content spreads within and across channels. Instead, channels' creation dates (*Creation date* field) enable the analysis of the correlation between a channel's longevity and its influence on information propagation. These temporal attributes support a variety of studies. For example, if combined with the forwarding information (*Is forwarded* and *Forwarded from* fields in Text messages and Media messages), they allow for tracing the rise and evolution of trending topics, assessing the velocity of information diffusion for different content types, identifying peak activity periods for specific topics or channels, and examining shifts in information flow patterns over time. Moreover, beyond individual channels, the communities identified in the TGDataset present opportunities for studying inter-community information diffusion. Investigating these larger

network structures could provide valuable insights into how information propagates across interconnected groups within Telegram, unveiling the mechanisms that drive broader dissemination trends.

7 Ethical Considerations

This paper presents the TGDataset, a new dataset that includes 120,979 Telegram channels and over 400 million messages. The data collection was conducted carefully to only include information from Telegram channels and exclude personal data such as usernames, phone numbers, and subscribed channels.

Our data collection process complies with Telegram's Terms of Service [58] and Telegram API's Terms of Service [57], as there are no clauses that prohibit the collection of public chat data. Additionally, we have taken care to avoid flooding the platform with requests during data collection. Moreover, our research complies with the academic research exemptions outlined in Article 85 of the GDPR [59], which provide certain flexibilities for processing publicly available data in the interest of freedom of expression and academic purposes. As our work exclusively involves publicly accessible content and does not process private user data, it falls under the legitimate or public interest exemptions. Furthermore, we adhere to the principle of data minimization, as emphasized in the GDPR's guidelines [23]: We limit our analysis to admin-generated public posts, ensuring that only the information strictly necessary for our research objectives is processed.

Lastly, since the dataset includes channels that discuss controversial topics, some controversial messages may be present in the dataset. To comply with ethical standards, we did not download any images or include links in the public dataset as they may contain adult content or copyrighted material. Hence, according to our IRB's policy, we did not require explicit authorization to conduct our experiments.

8 Limitation

To create the TGDataset, we began with 180 seed channels and then employed a snowball technique to grow the dataset. This approach is limited to only reaching channels that are linked by message forwarding. Despite this constraint, we were able to uncover over 100,000 channels and could potentially discover even more through continued iterations. Nonetheless, there may be groups of channels that remain inaccessible from our seed channels.

In the TGDataset, channels are represented as a snapshot of their current state. Although some historical details about the channels are available in the dataset, such as their creation date and timestamp of messages, other information is not present. For instance, a channel during time can change its name, description, and of course, the number of subscribers. This information could be useful for reconstructing the evolution of the channels. In the current release of the TGDataset, we do not retrieve the comments posted by users on channels' messages or the reaction emojis they may have used (such as the 'like' button on Facebook). While we have observed that many Telegram channels do not use these features, they can still offer valuable insights for future analyses.

9 Conclusions and Future works

Telegram has gained significant popularity in recent years. As a result, it is crucial to study and understand the activity taking place on the platform.

In this work, we present the TGDataset [38], a collection of more than 120,000 public Telegram channels that, to the best of our knowledge, is the largest collection of channels publicly available. After characterizing the main quantitative aspect of the TGDataset, we performed language detection to understand which are the most popular languages on the dataset. Then, we investigate the main topics covered by the English channels. Additionally, we publicly released the script we used to analyze it and the labeling we obtained (language and topic of the channels). In this paper, we investigate a few possible use cases in which our dataset can be extremely useful. In particular, our preliminary study of the TGDataset revealed some Telegram channels spread questionable content and conspiracy theories. Moreover, we observed the presence of several borderline activities and channels dealing with dubious ethical content. With the release of the TGDataset, we aim to provide a valuable resource to researchers that enables further investigation into these areas, leading to a more refined understanding of Telegram and helping to mitigate potential risks to users.

As future work, we plan to continue running our data collector and further enlarge the TGDataset, releasing new versions of the dataset at regular intervals of time. Moreover, we intend to overcome the actual limitations by recording channel updates (channel's name, subscribers, and description), adding to the list of channels to monitor also those referenced by link, inserting the possibility to add new seed channels, and collecting replies to messages.

10 Acknowledgments

References

[1] 2017. Telegram to block terror channels after Indonesian ban. https://www.bbc.com/news/business-40627739.
[2] 2018. Iran has banned Telegram after claiming the app encourages 'armed uprisings'. https://www.theverge.com/2018/5/1/17306792/telegram-banned-iran-encrypted-messaging-app-russia.
[3] 2019. Telegram the latest safe haven for white supremacists. https://www.adl.org/blog/telegram-the-latest-safe-haven-for-white-supremacists.
[4] 2020. Russia lifts ban on Telegram messaging app after failing to block it. https://www.reuters.com/article/us-russia-telegram-ban-idUSKBN23P2FT.
[5] 2021. NLTK RegexpTokenizer. https://www.nltk.org/_modules/nltk/tokenize/regexp.html.
[6] 2021. Telegram Analytics. https://tgstat.com/.
[7] 2021. Telegram FAQ. https://telegram.org/faq#q-what-is-telegram-what-do-i-do-here.
[8] 2021. Telethon's Documentation. https://docs.telethon.dev/en/latest/.
[9] 2022. Page Verification Guidelines. https://telegram.org/verify.
[10] Jan. 2021. WhatsApp delays privacy policy changes after users defect to rivals Signal and Telegram. https://fortune.com/2021/01/15/whatsapp-delays-privacy-policy-changes-after-users-defect-to-rivals-signal-and-telegram/.
[11] May 2021. WhatsApp privacy policy change: Telegram, Signal saw massive spike in January, shows data. https://indianexpress.com/article/technology/social/whatsapp-privacy-policy-change-telegram-signal-saw-massive-spike-in-january-shows-data/.
[12] 4chan. 2023. 4chan. https://www.4chan.org/.

[13] Azadeh Akbari and Rashid Gabdulhakov. 2019. Platform surveillance and resistance in Iran and Russia: The case of Telegram. *Surveillance & Society* 17, 1/2 (2019), 223–231.

[14] Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics.* 229–237.

[15] Kyle Banker, Douglas Garrett, Peter Bakkum, and Shaun Verch. 2016. *MongoDB in action: covers MongoDB version 3.0.* Simon and Schuster.

[16] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Telegram Dataset. In *Proceedings of the International AAAI Conf. on Web and Social Media*, Vol. 14. 840–847.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

[18] Zhenfeng Cao, Minzhang Zheng, Yulia Vorobyeva, Chaoming Song, and Neil Johnson. 2017. Dynamical patterns in individual trajectories toward extremism. *Available at SSRN 2979345* (2017).

[19] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[20] Arash Dargahi Nobari, Negar Reshadatmand, and Mahmood Neshati. 2017. Analysis of Telegram, an instant messaging service. In *Proceedings of the 2017 ACM on Conf. on Information and Knowledge Management.* 2035–2038.

[21] Gabriele Etta, Matteo Cinelli, Alessandro Galeazzi, Carlo Michele Valensise, Walter Quattrociocchi, and Mauro Conti. 2022. Comparing the Impact of Social Media Regulations on News Consumption. *IEEE Transactions on Computational Social Systems* (2022), 1–11. https://doi.org/10.1109/TCSS.2022.3171391

[22] Gabriele Etta, Alessandro Galeazzi, Jamie Ray Hutchings, Connor Stirling James Smith, Mauro Conti, Walter Quattrociocchi, and Giulio Valentino Dalla Riva. 2022. COVID-19 infodemic on Facebook and containment measures in Italy, United Kingdom and New Zealand. *PloS one* 17, 5 (2022), e0267022.

[23] Your Europe. November 19, 2024. Data protection under GDPR. https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_en.htm.

[24] European Organization For Nuclear Research and OpenAIRE. 2013. Zenodo. https://doi.org/10.25495/7GXK-RD71

[25] Sayan Ghosh, Mahantesh Halappanavar, Antonino Tumeo, Ananth Kalyanaraman, Hao Lu, Daniel Chavarria-Miranda, Arif Khan, and Assefaw Gebremedhin. 2018. Distributed louvain algorithm for graph community detection. In *2018 IEEE international parallel and distributed processing symposium (IPDPS).* IEEE, 885–895.

[26] Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. *Exploring network structure, dynamics, and function using NetworkX.* Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

[27] Ali Hashemi and Mohammad Ali Zare Chahooki. 2019. Telegram group quality measurement by user behavior analysis. *Social Network Analysis and Mining* 9, 1 (2019), 1–12.

[28] Taylor Hatmaker. 2021. Telegram blocks 'dozens' of hardcore hate channels threatening violence. https://techcrunch.com/2021/01/13/telegram-channels-banned-violent-threats-capitol/.

[29] Rebecca Heilweil. Jan. 2021. How Trump's internet built and broadcast the Capitol insurrection. https://www.vox.com/recode/22221285/trump-online-capitol-riot-far-right-parler-twitter-facebook.

[30] Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 42, 1 (2001), 177–196.

[31] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[32] Mohamad Hoseini, Philipe Melo, Fabricio Benevenuto, Anja Feldmann, and Savvas Zannettou. 2023. On the globalization of the QAnon conspiracy theory through Telegram. In *Proceedings of the 15th ACM Web Science Conf. 2023.* 75–85.

[33] Vincenzo Imperati, Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Francesco Sassi. 2025. The Conspiracy Money Machine: Uncovering Telegram's Conspiracy Channels and their Profit Model. In *34th USENIX Security Symp. (USENIX Security 25).*

[34] Asal Jalilvand and Mahmood Neshati. 2020. Channel retrieval: finding relevant broadcasters on Telegram. *Social Network Analysis and Mining* 10, 1 (2020), 1–16.

[35] Alex Kigerl. 2020. Behind the Scenes of the Underworld: Hierarchical Clustering of Two Leaked Carding Forum Databases. *Social Science Computer Review* (2020), 0894439320924735.

[36] Rachel Kraus. October 29, 2020. Telegram's massive revenge porn problem has made these women's lives hell. https://mashable.com/article/nudes-revenge-porn-crime-telegram.

[37] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. 2024. TG-Dataset. https://zenodo.org/records/7640712.

[38] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. 2024. TG-Dataset repository. https://github.com/SystemsLab-Sapienza/TGDataset.

[39] Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Jie Wu. 2023. It'sa Trap! Detection and Analysis of Fake Channels on Telegram. In *2023 IEEE International Conf. on Web Services (ICWS).* IEEE, 97–104.

[40] Massimo La Morgia, Alessandro Mei, Alberto Maria Mongardini, and Jie Wu. 2024. Pretending to be a VIP! Characterization and Detection of Fake and Clone Channels on Telegram. *ACM Transactions on the Web* (2024). https://doi.org/10.1145/3705014

[41] Massimo La Morgia, Alessandro Mei, Francesco Sassi, and Julinda Stefa. 2020. Pump and Dumps in the Bitcoin Era: Real Time Detection of Cryptocurrency Market Manipulations. In *2020 29th International Conf. on Computer Communications and Networks (ICCCN).* IEEE, 1–9.

[42] Dymples Leong. May 2020. Telegram, the powerful COVID-19 choice of communications by many governments. https://www.channelnewsasia.com/commentary/coronavirus-covid-19-government-telegram-whatsapp-fake-news-info-936061.

[43] Hannah Gelbart Maria Korenyuk Lucy Swinnen, Jack Goodman and Juliana Gragnani. 16 February 2022. Telegram: Where women's nudes are shared without consent. Telegram:Wherewomen'snudesaresharedwithoutconsent.

[44] Andy Maxwell. October 10, 2022. Telegram Piracy: Police Target 545 Channels & Eight Suspected Admins. https://torrentfreak.com/telegram-piracy-police-target-545-channels-eight-suspected-admins-221010/.

[45] MBFC. 2023. Media Bias Fact Check. https://mediabiasfactcheck.com/.

[46] Jemima McEvoy. Jan. 2021. Capitol Attack Was Planned Openly Online For Weeks—Police Still Weren't Ready. https://www.forbes.com/sites/jemimamcevoy/2021/01/07/capitol-attack-was-planned-openly-for-weeks-police-still-werent-ready/?sh=1b06696376e2.

[47] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing.* 262–272.

[48] MONGODB. November 19, 2024. MongoDB. https://www.mongodb.com/.

[49] Massimo La Morgia, Alessandro Mei, Francesco Sassi, and Julinda Stefa. 2021. The doge of wall street: Analysis and detection of pump and dump cryptocurrency manipulations. *ACM Transactions on Internet Technology (TOIT)* (2021).

[50] Lynnette Hui Xian Ng and Loke Jia Yuan. 2020. Is this pofma? Analysing public opinion and misinformation in a COVID-19 Telegram group chat. *arXiv preprint arXiv:2010.10113* (2020).

[51] NLTK Project. 2023. NLTK stopwords. https://www.nltk.org/search.html?q=stopwords&check_keywords=yes&area=default.

[52] Jack Ricle. Aug. 2022. Scammers in telegram and how to report. https://www.telegramadviser.com/scammers-in-telegram-and-how-to-report/.

[53] Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational linguistics* 24, 1 (1998), 97–123.

[54] Nakatani Shuyo. 2010. Language Detection Library for Java. http://code.google.com/p/language-detection/

[55] Manish Singh. November 30, 2022. Telegram shares users' data in copyright violation lawsuit. https://techcrunch.com/2022/11/29/telegram-shares-data-of-users-accused-of-copyright-violation-following-court-order/.

[56] Joe Sommerlad. March 2021. Sabmyk Network: Founder of bizarre new religion targeting QAnon believers 'unmasked' by Hope Not Hate. https://www.independent.co.uk/news/world/europe/sabmyk-network-qanon-conspiracy-theories-b1820639.html.

[57] Telegram. November 19, 2024. Telegram API Terms of Service. https://core.telegram.org/api/terms.

[58] Telegram. November 19, 2024. Telegram Terms Of Service. https://telegram.org/tos/.

[59] GDPR TEXT. November 19, 2024. Article 85 GDPR. Processing and freedom of expression and information. https://gdpr-text.com/read/article-85/.

[60] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* 9, 1 (2019), 1–12.

[61] Aleksandra Urman and Stefan Katz. 2022. What they do in the shadows: examining the far-right networks on Telegram. *Information, communication & society* 25, 7 (2022), 904–923.

[62] Alper Kursat Uysal and Serkan Gunal. 2014. The impact of preprocessing on text classification. *Information Processing & Management* 50, 1 (2014), 104–112.

[63] Janith Weerasinghe, Bailey Flanigan, Aviel Stein, Damon McCoy, and Rachel Greenstadt. 2020. The pod people: Understanding manipulation of social media popularity via reciprocity abuse. In *Proceedings of The Web Conf. 2020.* 1874–1884.

[64] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9.

[65] Jiahua Xu and Benjamin Livshits. 2019. The anatomy of a cryptocurrency pump-and-dump scheme. In *28th {USENIX} Security Symp. ({USENIX} Security 19).* 1609–1625.

[66] Ahmet S Yayla and Anne Speckhard. 2017. Telegram: The mighty application that ISIS loves. *International Center for the Study of Violent Extremism* (2017).