Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1480–1489
Marseille, 20-25 June 2022

Multimodal Pipeline for Collection of Misinformation Data from Telegram

Jose Sosa, Serge Sharoff
University of Leeds
Woodhouse Lane, LS2 9JT, Leeds, UK
{scjasm, s.sharoff}@leeds.ac.uk

Abstract

The paper presents the outcomes of AI-COVID19, our project aimed at better understanding of misinformation flow about COVID-19 across social media platforms. The specific focus of the study reported in this paper is on collecting data from Telegram groups which are active in promotion of COVID-related misinformation. Our corpus collected so far contains around 28 million words, from almost one million messages. Given that a substantial portion of misinformation flow in social media is spread via multimodal means, such as images and video, we have also developed a mechanism for utilising such channels via producing automatic transcripts for videos and automatic classification for images into such categories as memes, screenshots of posts and other kinds of images. The accuracy of the image classification pipeline is around 87%.

Keywords: Telegram data collection, COVID-19 misinformation, multimodal classification

1. Introduction

The widespread distribution of COVID-19 misinformation leads to confusion, anti-health policy sentiment, and risk-tolerant behaviour (Chou et al., 2021). AI-COVID19 is our project which aims to use AI tools to understand the dynamics of misinformation spread across several social media platforms, including Twitter, Facebook and Telegram. This paper concentrates on the methods and outcomes of data collection from Telegram, a platform that is currently less active in policing misinformation, so that we can collect a greater variety of COVID-19 misinformation examples.

We designed a pipeline to collect COVID-19 misinformation from Telegram public channels. Then, we used this pipeline to build one of the first multimodal datasets of COVID-19 misinformation; i.e. in addition to the prevailing text data, our dataset includes images, videos, and documents. Overall the dataset comprises almost one million messages from 2k different public channels related to spreading COVID-19 misleading information. In addition, it includes 38k images, 15k videos, and 522 documents (mostly in the PDF and DOCX formats) from those channels. Furthermore, it organises the collected images into three categories: memes, posts and others by means of automatic image classification. Finally, it incorporates a set of transcripts for the collected videos.

We summarised our contributions as follows:

- Automatic pipeline for collecting misinformation from Telegram.
- Joint collection of text and multimedia data. Our pipeline for collecting data allows us to get text and also any multimedia data that can be part of the messages, e.g. images, videos, documents, audios, and stickers.
- A classifier for the collected images. We train a CNN classifier based on AlexNet (Krizhevsky et al., 2012) to identify memes and images from text-based posts.
- A new telegram multimodal dataset on COVID-19 related misinformation.

The complete code for collecting data, the classifier, and the first image of our dataset, including information about Telegram messages, users, media messages, and channels; multimedia data as classified images, videos and their transcripts will be made available for public use.[1]

2. Related work

Since the rise of the pandemic, there have been many studies aimed at collecting resources and creating collections to deal with COVID-19 from different perspectives. Then, the number of datasets around COVID-19 significantly increased, and data sources diversified. Consequently, we can find datasets from purely scientific productions, like papers or specialised medical images, to collections of data extracted from unverified sources, like social networks.

Concerning datasets of scientific productions, the COVID-19 Open Research Dataset (CORD-19) is perhaps the best example. It contains academic articles on COVID-19 and related corona-viruses studies published between 1980 and 2021. CORD-19 represents a joint effort to provide a resource to interdisciplinary scientific communities to identify effective treatments and develop better policies for COVID-19 (Wang et al., 2020b). Another reliable source of COVID-19 information is in (Dong et al., 2020). They developed an interactive web-based dashboard that allows real-time visualisation and tracking of reported COVID-19 cases. This practical tool served as a base for other studies in the context of COVID-19 (Dey et al., 2020).

[1]https://github.com/josesosajs/telegram-data-collection

In addition to collecting texts and statistical information, there are studies which collect COVID-related images, in particular, medical images (Cohen et al., 2020; Yang et al., 2020; Wang et al., 2021). For example, (Xu et al., 2020) collected a dataset of Computer Tomography (CT) images from 110 patients with COVID-19 to train a deep neural network, which then can automatically detect the presence of COVID-19 on new CT images. Other studies rely on the collection and use of x-ray images for similar purposes (Hemdan et al., 2020; Wang et al., 2020a; Apostolopoulos and Mpesiana, 2020).

Automatically collecting and building datasets with those kinds of scientific information is a non-trivial process, principally because of their requirement of intense supervision to create and verify the data. However, with respect to massive amounts of unverified data, social networks represent fruitful sources to build valuable datasets around specific topics. Contrary to scientific productions, those platforms allow users to create and share any content without meticulous verification. Thus, it makes social networks a convenient medium for spreading misleading information. However, it is crucial to collect and analyse unverified data to model and understand the social response against some emerging events, e.g. COVID-19 (Hossain et al., 2020; Alam et al., 2021; Pennycook et al., 2020; Brindha et al., 2020).

Predominately, generic data collection from social media is performed by tracking certain accounts, posts, users, and keywords akin to the topic (Banda et al., 2020; Chen et al., 2020; Aggarwal et al., 2020; Basile et al., 2021; Baumgartner et al., 2020). Twitter is perhaps the most popular social network in this context. Then, we can find several COVID-19 related datasets from this platform. For instance, Banda et al. (2020) released a dataset of more than 150 million tweets associated with COVID-19, which represents one of the largest collections available up to date. Similar, (Chen et al., 2020) have produced a dataset of approximately 50 million tweets. Although it is a smaller dataset, it is more diverse regarding the number of languages, which makes it convenient for studies on languages other than English. Also with respect to collections in languages other than English, Alqurashi et al. (2020) introduce a dataset of almost 4 million Arabic tweets linked to COVID-19.

Analogous to Twitter, Reddit is also a valuable social media platform for building COVID-19 datasets. For example, Aggarwal et al. (2020) have produced a dataset of COVID-19 related posts and comments from Reddit, which comprises a total of 105,000 posts. Similarly, (Basile et al., 2021) presents an interesting collection of Reddit COVID-19 posts from different countries. More recently, the attention has been paid other social networks, e.g. (Zarei et al., 2020) collected a dataset of COVID-19 posts and comments from Instagram. Furthermore, (Medina Serrano et al., 2020) presents a dataset which uses comments from YouTube videos to study misinformation.

In contrast to generic data collection from social media, identifying COVID-19 misinformation is a more difficult task which requires some degree of manual verification. Some approaches managed to create misinformation datasets by a combination of data from different sources (Patwa et al., 2020; Yang et al., 2020; Haouari et al., 2020). (Cheng et al., 2021) created a set of annotated tweets specifically containing COVID-19 misinformation. Furthermore, some misinformation datasets cover languages like Chinese (Yang et al., 2021) and Arabic (Haouari et al., 2020). Alam et al, (2021) have produced a multilingual training set also covering the impact of misinformation, such as harmfulness and topics of their claims, for example, “bad cure”.

Mobile messaging platforms like WhatsApp and Telegram also represent a rich source of fake and legitimate information related to the COVID-19 pandemic. The last one is the most open platform regarding access to its API. Several interesting studies in the recent literature collected and analysed Telegram data to comprehend emerging social problems like immigration movements (Nikkhah et al., 2018), manifestations, and terrorism (Prucha, 2016; Yayla and Speckhard, 2017). However, collecting data from Telegram is still a developing field, so there is a lack of Telegram datasets for some trending topics like COVID-19. Some studies like (Ng and Loke, 2020) collected data from this topic. However, their analysis only covered one Telegram discussion group, which does not represent the diversity of data channels. Contrarily, our approach collects data from an extensive set of Telegram public channels, which are highly related to spreading misinformation about COVID-19.

One of the main drawbacks of many existing COVID-19 datasets from social networks like Twitter, Reddit, Youtube, and specifically Telegram is their focus on solely text data and ignoring multimedia like images and videos, which clearly impact information and misinformation flows and could also be beneficial for COVID communication research. Only few studies explored this, for example (Pramanick et al., 2021) presents an interesting multimodal study to evaluate the harmfulness of COVID-19 oriented memes, which are abundant in most social networks. Thus, differently from other popular Telegram datasets like (Baumgartner et al., 2020) we additionally collected and analysed image and video data. To the best of our knowledge, our dataset is the first with respect to collecting multimodal COVID-19 data with the focus on misinformation.

3. Methodology

In this section, we describe our computational strategy for collecting data from Telegram public channels, together with our approach for analysing the media ele-

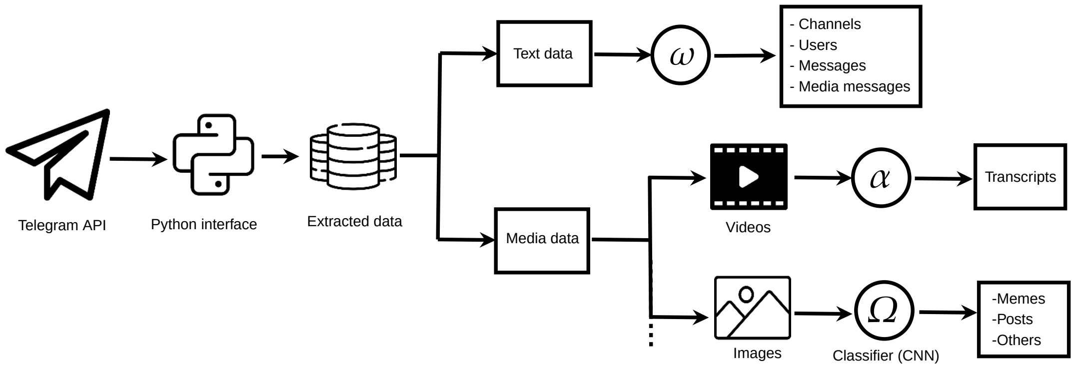

Figure 1: Telegram data collection pipeline. Our pipeline for collecting data starts with the basic block to establish the communication between our algorithm and the Telegram API. Furthermore, we designed two workflows: one for collecting texts of the messages as well as data about the channels and the users. And the second branch is for collecting media, e.g. images and videos. Then, we extended it by classifying the images with a CNN, and extracting transcripts from the videos.

ments on some messages, e.g. images and videos.

3.1. Collecting Data

A common approach to obtaining data from messaging platforms like Telegram is by exploring the public channels. Contrary to other social networks like Twitter or Facebook, where user activity is commonly available, Telegram is not open regarding access to the messages history for specific users. Then, in line with prior approaches, our data collection from Telegram is channel-based. Similarly to Baumgartner et al. (2020) and Wich et al. (2021), we adopted a snowball sampling strategy, in which a set of seed channels leads to its augmentation by selecting channel names in messages forwarded from other channels. We started gathering data from a manually extracted list of approximately 13 public channels likely related to spreading misinformation about COVID. Then we augmented it to 70, which represented the seed for the snowball sampling strategy. Note that we manually verified that most messages from those selected channels were related to COVID misinformation.

Currently, we are collecting data from Telegram public channels daily. At the beginning of our collection process, we were retrieving messages from the 70 items in the seed list. Then, we augmented it by considering the source of forwarded messages. We repeated this process daily using the inflated collection of channels. For future iterations, due to the growing number of channels and the limitations of the Telegram API, we randomly shuffled our list and got messages from the first n channels. After collecting some data, we decided to change the sorting criteria of our list of channels. Then, we ordered them by their contributions to the dataset. Thus, we assured the collection start with the n elements that more messages provide to the dataset. Note that n is constrained by the number of requests sent to Telegram API. On average, we are getting data from 200 public channels every day.

A fundamental element of our data collecting process is the interface to establish the communication between Telegram API and our algorithm. For this task, we used Telethon[2], a python package that allows us to configure and control the requests to the Telegram API. Then, we collected and stored data from the channels, messages and users. Furthermore, we identified and collected the messages containing media when possible. Thus, the output of our collection process is a set of four JSON files: channels, messages, users, and media messages (as illustrated on Figure 2); and a folder with multiple types of media data. Having a flexible file format as the JSON files allows deploying those to any relational or non-relational database scheme.

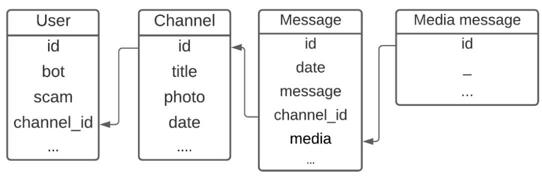

Figure 2: Structure and relationship of the JSON files for each collected Telegram entity. A complete list of the fields for each entity is include in the documentation of Telethon's API: `https://tl.telethon.dev/constructors/`. The only extra field that we added is an internal UUID for each element of every entity which allows to establish the relationship between them.

[2]https://docs.telethon.dev/en/latest/

3.1.1. Extraction of multimodal data

Additionally to collecting text data, we added a second branch to our pipeline to download the media from some messages, e.g. images, videos, audios, and documents. We saved those items despite their nature. However, we are particularly interested in analysing pictures and videos.

3.1.2. Video Data

We collected and stored videos from a subset of telegram messages. Our early analysis of video data aims to reduce their dimensionality. Then, we extracted the transcript from each video and used the text data to summarise the video content. A transcript represents a more convenient resource for future experiments because it is less complex to analyse and classify a transcript than the video itself. For extracting the transcripts, we used α, which represents the process of mapping from the input video v_i to its respective transcript t_i. However this mapping does not take place directly, i.e. a function β first extract the audio a_i from v_i. Then, γ inputs this intermediate representation and map it to t_i, which is the corresponding transcript representation. Thus, the complete mapping is given by:

$$\alpha(v_i) : \beta(v_i) \rightarrow \gamma(a_i) \rightarrow t_i \tag{1}$$

We implemented the audio extractor β and the transcript generator γ by means of standard python libraries [3,4].

After having the automatically generated transcripts t_i, we manually evaluated their accuracy, i.e. a human annotator checked if a sample of the transcript texts matches what the videos are saying. Using the accuracy indicator, we can then classify the accurate transcripts. However, now this is still a work in progress. From a random sample of 60 transcripts, our early analysis suggests that most of those are accurate. According to the annotator 75% of the transcripts in the subset are fairly accurate (errors are caused by such factors as background music) and around 33% of them present COVID-19 misinformation. Then, we are expecting our future annotation follows a similar trend.

3.1.3. Image Data

As a first approach to exploit the images, we trained a Convolutional Neural Network (CNN) Ω to distinguish between three categories: memes, posts, and others, (see Figure 3 for examples) with following definitions of the categories:

- **Memes**: A meme is an image with a short piece of text, typically aimed at exciting humorous or amusing response. At this stage, the task is purely a visual classification, i.e. we are not looking for the meaning or context of the text. However, considering the nature of the channels we are following, we can assume that the collected images are highly related to COVID-19 misinformation.
- **Posts**: In this category, we included all the images that show posts from social media, most of them pictured as screenshots from news websites, Facebook, Twitter, and WhatsApp. Shared posters are also considered as members of this category.
- **Others**: This category includes all the other images that are not memes or posts. For example, we can find images of people, vaccines, masks, world leaders, pets, objects, and landscapes.

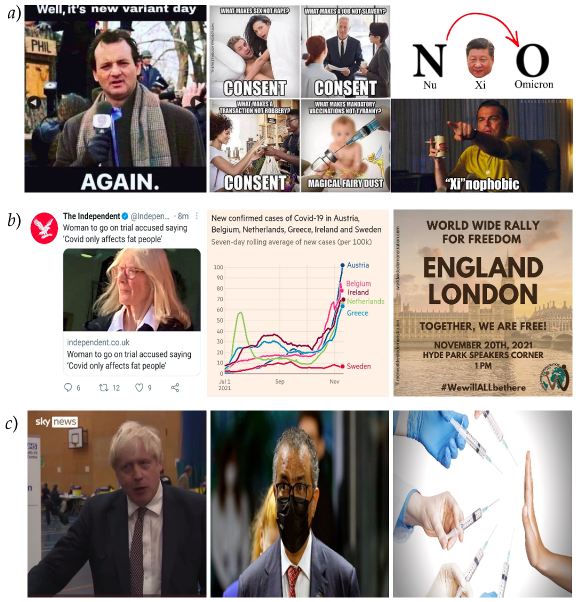

Figure 3: Random examples of classified images for each one of the three categories: a) memes, b) posts, and c) others.

We are particularly interested in the first two categories, i.e. memes and posts, because we can do post-processing (e.g. text extraction) and input those to existing pipelines for multimodal classification (Pramanick et al., 2021).

3.1.4. Image Classifier Architecture

We based our CNN for image classification Ω on a pre-trained AlexNet (Krizhevsky et al., 2012). We fine-tuned this model with a COVID-specific training dataset and modified the last fully connected layer to produce outputs for our three classes. For training the network Ω, we utilised a subset of a publicity available dataset of Twitter images together with a dataset of COVID-specific memes(Singh et al., 2020). We selected approximately 3k pictures for each class and divided those in a 90/10 ratio for training and validation. Our fine-tuned classifier Ω achieved an overall accuracy of 87.4% on a small test set from our Telegram collection, the confusion matrix and the scores are shown in Figure 4 and Table 1.

Similarly to the process with the videos, we extracted

[3]https://zulko.github.io/moviepy/ref/AudioClip.html
[4]https://pypi.org/project/SpeechRecognition/

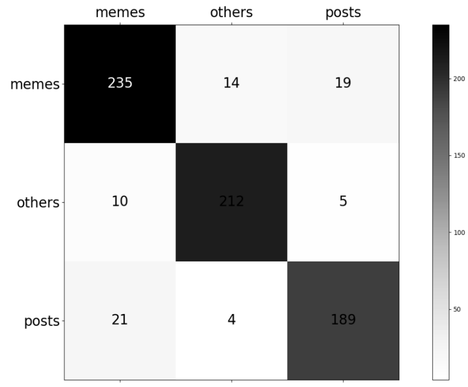

Figure 4: Confusion matrix for the three classes.

Class	Precision	Recall	F1-score
Memes	0.88	0.92	0.89
Others	0.88	0.93	0.88
Posts	0.88	0.93	0.89

Table 1: Precision, Recall and F1-Score for each class.

the text from a subset of memes and posts images I_i when possible. The extraction of the text tx_i from an image I_i is given by $\lambda(I_i) = tx_i$, where λ is an standard function for Optical Character Recognition (OCR)[5]. However, at this stage we are not aiming for a deep analysis of the text on the memes and posts. We prioritised having an accurate classification of the images of our dataset. Then, we can use, particularly the memes as inputs for existing multimodal approaches which evaluated their harmfulness(Moens et al., 2021; Zhou et al., 2021).

4. Results

Channels. We started collecting messages from a list of approximately 13 channels highly related to contain misinformation regarding COVID-19. We augmented this list everyday as described on section 3.1. Then, the first version of our dataset contains messages from $2,602$ different Telegram channels. Because the limitations of Telegram API, we are not able to collect data from our full list of channels, which includes $11,161$ channels until now. Hence, after some days of collecting data, we selected a subset of channels which most contributed to our misinformation dataset, based on the number of messages and prioritise those for future data collection. This is reflected as constant line on the plot of Figure 5. Starting with 13 predefined channels, then 77, and from that point we increased the amount of channels automatically. In average we are collecting data from approximately 200 channels daily. There are some values below the average, those are because some unexpected issues with the collection, which normally stopped our script earlier.

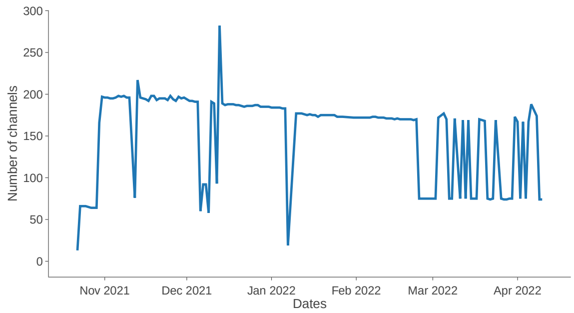

Figure 5: Number of channels extracted per day.

Users. In addition to the data about the channels, we collected their users. On average our dataset contains $159,905$ unique users as shown on Figure 6. Although the information available regarding Telegram users is quite restrictive, there are some practical indicators, e.g. if a user is a bot. Then, we can use it to determine the sort of members for a given channel.

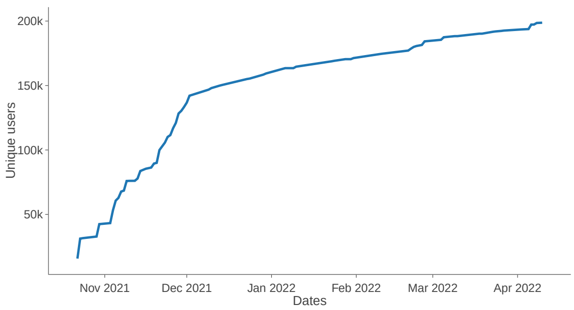

Figure 6: Unique users from all channels.

Messages. Overall the first snapshot of our dataset contains $1,131,560$. After ignoring the empty messages, which represent service messages from Telegram or messages with only media, we got a total of $812,196$ messages that have some text (See Figure 7). We started to collect data from October 22^{nd}, its first release version includes data until December 31^{st} of 2021, with the graphs reported in this paper updated until April 2022. As our data collection is still running, we expect our dataset to continue growing.
We used a language classifier (Lui and Baldwin, 2012)[6] for automatic language identification for each message. Although English is the predominant language in our pipeline, our collection results contain a significant proportion of messages in other languages as illustrated in Table 2. In average the length of messages in English is 256 characters, which is quite long, i.e. almost twice the length of text data obtained from other social networks like Twitter.

[5]https://github.com/madmaze/pytesseract

[6]https://github.com/saffsd/langid.py

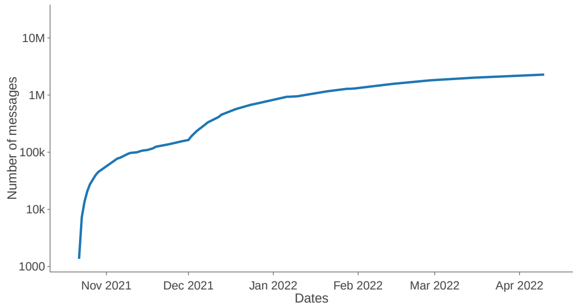

Figure 7: Accumulated telegram messages.

Language	Messages(↓)	Percent
English (en)	549,126	67.61
German (de)	58,308	7.18
Chinese (zh)	35,051	4.32
Spanish (es)	24,821	3.06
Russian (ru)	21,098	2.60
Others	106, 630	15.23

Table 2: Languages distribution of all the messages in our dataset.

Additionally, we calculated the frequency for each word in our corpus, only including English Telegram messages. Then, we used the frequency list of Open-WebTex(Sharoff, 2020) as a reference corpus for computing the log-likelihood (LL). This metric helps to identify the most indicative words in our corpus when compared against the reference. Table 3 shows the top 20 words sorted by their log-likelihood score.

Hashtags and mentions. Furthermore, following previous works (Baumgartner et al., 2020), we looked for hashtags and mentions within the messages, obtaining a list of $9,127$ unique hashtags, from a set of $53,728$ elements and $127,248$ mentions, from which $5,468$ are unique mentions. We observed some evident hashtags from the list on Table 4 directly related to COVID-19. For example, #Omicron, #COVID19, #Vaccines, and #Covid. For the case of mentions in Table 5, most of them correspond to channels dedicated to shared news, predominately fake ones. However, two channels of the list got the Telegram verified badge; @disclosetv and @EpochTimes, which are neutral respecting their shared content.

To verify our snowball strategy, we assessed random 50 messages from each of the original set of 13 channels, the 70 channel extended set and from our current collection. The rate of misinformation messages drops from 74% to 60% to 42%, while still keeping the majority of messages in our target category.

4.1. Multimodal data

Media messages. From a total of $1,131,560$ Telegram messages, $888,810$ include some media data, either accompanying the text or solely the media itself (See Figure 8). Because the limitations of the number of request send to the Telegram API, we are not able to download all the media from those messages. However, we built a JSON file for saving them and keep the reference to the file, which we can use those to download it if still available in the future.

The media messages are distributed into nine categories

Word	F1	F2	LL(↓)
vaccine	132959	26026	179230
vaccinated	26075	9086	71786
vaccines	90899	9471	54248
vaccination	48375	7312	46913
unvaccinated	7366	4063	35175
jab	21719	4708	33285
pandemic	19085	3646	24948
vax	2222	2501	24268
coronavirus	1204	2097	21489
mandates	32631	3339	19013
virus	192947	5061	16019
jabbed	2345	1620	14590
jabs	8847	2005	14335
booster	35224	2669	13700
chronology	11605	2016	13447
adverse	75885	3223	13057
ivermectin	810	1207	12145
vaxxed	1068	1184	11465
deaths	331266	4854	10387
passports	32779	2007	9501

Table 3: Frequencies and Log-likelihood scores (LL) of representative words appearing on the Telegram messages.

Hashtag	Proportion(%)
#KAG	12.84
#WeAreTheNewMedia	5.43
#Omicron	1.00
#COVID19	0.98
#WWG1WGA	0.84
#ShutItDown	0.66
#FightBack	0.49
#CrimesAgainstHumanity	0.39
#UnitedWeStand	0.36
#MAGA	0.34
#ReclaimTheLine	0.32
#China	0.31
#DoNotComply	0.30
#SaveTheChildren	0.30
#Ukraine	0.30
#Vaccines	0.29
#TheDefender	0.29
#UndergroundWarReport	0.28
#Covid	0.25
#Antifa	0.24

Table 4: Most common hashtags from our corpus of Telegram messages. Note that we performed this analysis exclusively on English messages.

Mention	Prop.(%)
@WeTheNews	5.66
@PookztA	5.42
@PatriotArmy	5.42
@SergeantRobertHorton	4.26
@disclosetv	4.11
@ZeroHedgeTyler	2.28
@AreWeAllBeingPlayed	2.09
@EpochTimes	1.73
@HoCoMDPatriots	1.66
@KanekoaTheGreat	1.40
@leagueofextraordinarypepes	1.38
@TGNewsU	1.35
@No_BS_News	1.23
@OneRepublicNetwork	1.21
@HATSTRUTH	1.20
@ChiefNerd	1.07
@CBKNEWS	1.07
@ExposeThePEDOSendTheCABAL	0.98
@awakenspecies	0.87
@GitmoTV	0.84

Table 5: Most common mentions from our corpus of Telegram messages. Similar to the hashtags, we only used English messages.

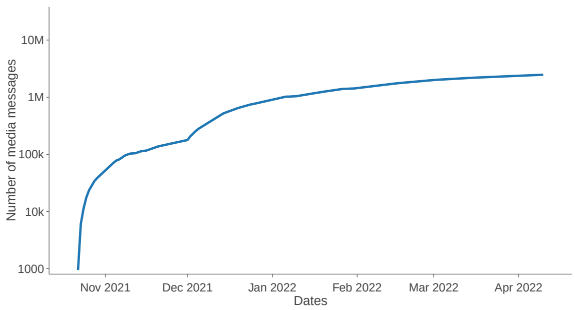

Figure 8: Accumulated media messages.

determined by Telegram: *Photo, Document, WebPage, Poll, Invoice, Unsupported, Game, Dice,* and *Contact*. According to Figure 9, *Photo*, *Document* and *WebPage* are the dominant categories on our dataset. Note that Telegram includes videos under the category of *Document*.

We downloaded the files from media messages when possible despite their category, which extended our dataset. Currently, it comprises $40,882$ images, $15,040$ videos, and 522 documents (.pdf, .doc, etc). Figure 10 shows the number of collected files every day. At this stage, we are particularly interested in the analysis of videos and images.

Images. After removing duplicates from the whole set of images, we used our trained classifier described in section 3.1.4 to classify the remainder $37,616$ into three classes: Memes, posts, and others. As shown in Figure 11 the amount posts (16,546) and others (13,999) are quite similar, while the number of memes

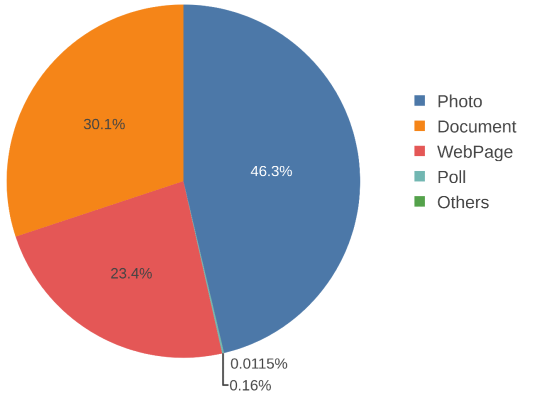

Figure 9: Media messages types distribution. Others includes: Invoice, Unsupported, Game, Dice, and Contact.

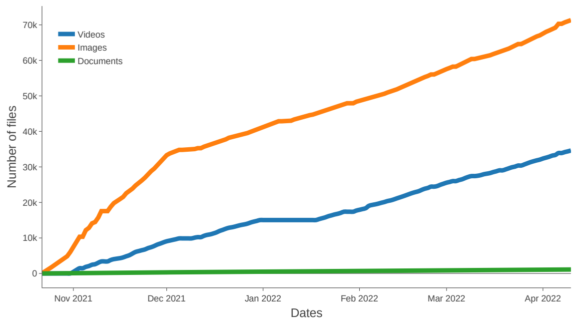

Figure 10: Accumulated media. Daily distribution of collected images, videos, and documents.

(7,071) represents almost half of them.

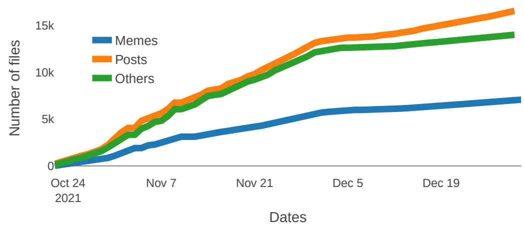

Figure 11: Distribution of classified images.

Videos and transcripts. We extracted the transcripts from the collected videos and built a corpus of $3,231,184$ words from them. Then, we calculated the log-likelihood score, using the same process as for the messages. Table 6 lists the top 30 words from this corpus, based on their log-likelihood score. We only removed function words[7]. Note that most of the words in the list are related to COVID-19, which is a positive

[7]https://www.nltk.org/

indicator to assume the context of the videos is in the COVID19 domain without explicitly looking at them.

Word	F1	F2	LL(↓)
vaccine	132959	5943	44861
vaccinated	26075	2210	19408
people	15749678	19211	18879
vaccines	90899	2759	18751
virus	192947	3157	17664
f******	2482	1324	15992
coronavirus	1204	1063	13653
pandemic	19085	1083	8676
vaccination	48375	1277	8332
pfizer	12224	748	6101
booster	35224	865	5522
5g	40781	852	5171
ivermectin	810	404	4836
myocarditis	384	364	4712
graphene	34240	748	4605
immune	146846	1099	4499
lockdown	19436	628	4344
oxide	27241	660	4196
omicron	706	344	4105
adverse	75885	843	4086
children	2504217	3483	4075
pcr	16345	506	3458
flu	64332	704	3392
protein	213869	1015	3291
viruses	63719	687	3291
hydroxychloroquine	122	237	3283
unvaccinated	7366	411	3279
disease	574392	1502	3275
vaccinations	18834	481	3107
cells	521885	1395	3095

Table 6: Log-likelihood score (LL) of words appearing on the transcripts.

5. Conclusions and Future Work

In this paper we described our pipeline for collecting multimodal data from Telegram. Futhermore we detailed the first version of our misinformation dataset, to the best of our knowledge it is the first Telegram dataset to include multimodal misinformation data about COVID-19. Our dataset includes almost one million Telegram messages from approximately 2K channels. Additionally, it comprises around 60k multimedia files, distributed between images, videos and documents. Furthermore, we report an automatic classifier for the image categories, and a transcript extraction tool for the collected videos.

Our dataset represents a valuable resource for researchers from different disciplines. For example, our collection of memes could augment similar existing datasets for multimodal analysis. Similarly, our corpus of messages and video transcripts could serve to study the flow of COVID-19 misinformation in social networks. This early version of the dataset represent one of the most complete and structured Telegram collections around COVID-19 misinformation. We plan to continue collecting data in the same way and expand our current dataset within the next months. Furthermore, we will perform an extensive analysis of the images, videos and transcripts using multimodal analysis frameworks (Knight and Adolphs, 2020).

6. Acknowledgements

Project has been funded via the EPSRC IAA3125 grant: AI tracing tools for detecting COVID19 misinformation in collaboration with the Yorkshire&Humber Academic Health Sciences Network and the British National Health Service.

7. Bibliographical References

Aggarwal, J., Rabinovich, E., and Stevenson, S. (2020). Exploration of gender differences in COVID-19 discourse on Reddit. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., and Nakov, P. (2021). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Alqurashi, S., Alhindi, A., and Alanazi, E. (2020). Large arabic twitter dataset on COVID-19. *arXiv preprint arXiv:2004.04315*.

Apostolopoulos, I. D. and Mpesiana, T. A. (2020). COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, 43(2):635–640.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., and Chowell, G. (2020). A twitter dataset of 150+ million tweets related to COVID-19 for open research. *Type: dataset*.

Basile, V., Cauteruccio, F., and Terracina, G. (2021). How dramatic events can affect emotionality in social posting: The impact of COVID-19 on reddit. *Future Internet*, 13(2):29.

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, pages 840–847.

Brindha, M. D., Jayaseelan, R., and Kadeswara, S. (2020). Social media reigned by information or misinformation about COVID-19: a phenomenological study. *Alochana Chakra Journal*, 9(5):585–602.

Chen, E., Lerman, K., and Ferrara, E. (2020). COVID-19: The first public coronavirus twitter dataset. *Type: dataset*.

Cheng, M., Wang, S., Yan, X., Yang, T., Wang, W., Huang, Z., Xiao, X., Nazarian, S., and Bogdan, P. (2021). A COVID-19 rumor dataset. *Frontiers in Psychology*, 12.

Chou, W.-Y. S., Gaysynsky, A., and Vanderpool, R. C. (2021). The COVID-19 misinfodemic: Moving beyond fact-checking. *Health Education & Behavior*, 48(1):9–13.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020). COVID-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*.

Dey, S. K., Rahman, M. M., Siddiqi, U. R., and Howlader, A. (2020). Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach. *Journal of medical virology*, 92(6):632–638.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Haouari, F., Hasanain, M., Suwaileh, R., and Elsayed, T. (2020). Arcov19-rumors: Arabic COVID-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*.

Hemdan, E. E.-D., Shouman, M. A., and Karar, M. E. (2020). Covidx-net: A framework of deep learning classifiers to diagnose COVID-19 in x-ray images. *arXiv preprint arXiv:2003.11055*.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). COVIDLies: Detecting COVID-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December. Association for Computational Linguistics.

Knight, D. and Adolphs, S. (2020). Multimodal corpora. In *A practical handbook of corpus linguistics*, pages 353–371. Springer.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Lui, M. and Baldwin, T. (2012). langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July. Association for Computational Linguistics.

Medina Serrano, J. C., Papakyriakopoulos, O., and Hegelich, S. (2020). NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Moens, M. F., Huang, X.-J., Specia, L., and Yih, W.-t. (2021). Findings of the association for computational linguistics: Emnlp 2021. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Ng, L. H. X. and Loke, J. Y. (2020). Analyzing public opinion and misinformation in a COVID-19 telegram group chat. *IEEE Internet Computing*, 25(2):84–91.

Nikkhah, S., Miller, A. D., and Young, A. L. (2018). Telegram as an immigration management tool. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 345–348.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., and Chakraborty, T. (2020). Fighting an infodemic: COVID-19 fake news dataset. *arXiv preprint arXiv:2011.03327*.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., and Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780.

Pramanick, S., Sharma, S., Dimitrov, D., Akhtar, M. S., Nakov, P., and Chakraborty, T. (2021). Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Prucha, N. (2016). Is and the jihadist information highway–projecting influence and religious identity via telegram. *Perspectives on Terrorism*, 10(6):48–58.

Sharoff, S. (2020). Know thy corpus! robust methods for digital curation of web corpora. *arXiv preprint arXiv:2003.06389*.

Singh, A., Goswami, V., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., and Parikh, D. (2020). Mmf: A multimodal framework for vision and language research. `https://github.com/facebookresearch/mmf`.

Wang, L., Lin, Z. Q., and Wong, A. (2020a). Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R. M., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D. A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D. M., Weld, D. S., Etzioni, O., and Kohlmeier, S. (2020b). CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July. Association for Computational Linguistics.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., Cai, M., Yang, J., Li, Y., Meng, X., et al. (2021). A deep learning algorithm using ct images to screen for corona virus disease (COVID-19). *European radiology*, pages 1–9.

Wich, M., Gorniak, A., Eder, T., Bartmann, D., Cakici, B. E., and Groh, G. (2021). Introducing an abusive language classification framework for telegram to investigate the german hater community. *arXiv preprint arXiv:2109.07346*.

Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J., et al. (2020). A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129.

Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., and Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about COVID-19. *arXiv preprint arXiv:2003.13865*.

Yang, C., Zhou, X., and Zafarani, R. (2021). Checked: Chinese COVID-19 fake news dataset. *Social Network Analysis and Mining*, 11(1):1–8.

Yayla, A. S. and Speckhard, A. (2017). Telegram: The mighty application that isis loves. *International Center for the Study of Violent Extremism*, 9.

Zarei, K., Farahbakhsh, R., Crespi, N., and Tyson, G. (2020). A first instagram dataset on COVID-19. *arXiv preprint arXiv:2004.12226*.

Zhou, Y., Chen, Z., and Yang, H. (2021). Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE.