Multimodal Pipeline for Collection of Misinformation Data from Telegram

Jose Sosa, Serge Sharoff
School of Computing — School of Languages, Cultures and Societies

UNIVERSITY OF LEEDS

Code available at:
Contact: scjasm@leeds.ac.uk

OVERVIEW

GOAL: Collecting multimodal data from Telegram groups which are active in promotion of COVID-related misinformation.

APPROACH: We designed a pipeline to collect COVID-19 misinformation from Telegram public channels. We used this pipeline to build one of the first multimodal datasets of COVID-19 misinformation. We have also developed a mechanism for producing automatic transcripts for videos and automatic classification for images into such categories:

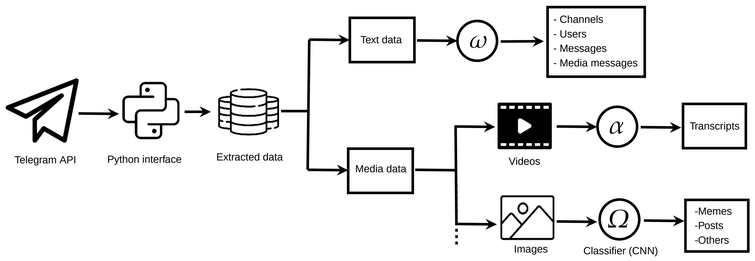
- Memes
- Screenshots of posts
- Other kinds of images

CONTRIBUTIONS

- Automatic pipeline for collecting misinformation from Telegram.
- Joint collection of text and multimedia data, e.g., images, videos, documents, audios, and stickers.
- A classifier for the collected images. We train a CNN classifier to identify memes and images from text-based posts.
- A new telegram multimodal dataset on COVID-19 related misinformation.

METHOD

Pipeline for joint collection of Telegram messages and multimedia data, e.g., images, videos, and documents.



Method diagram

Text data: ω represents the process of extracting and breaking-down the text data. We collected and stored data from the channels, messages, users, and media messages. We stored those in a set of JSON files.

METHOD

Extracting transcripts: α, represents the process of mapping from the input video v_i to its respective transcript t_i. β first extracts the audio a_i from v_i. Then, γ inputs this intermediate representation and map it to t_i, which is the corresponding transcript. The complete mapping is given by:

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

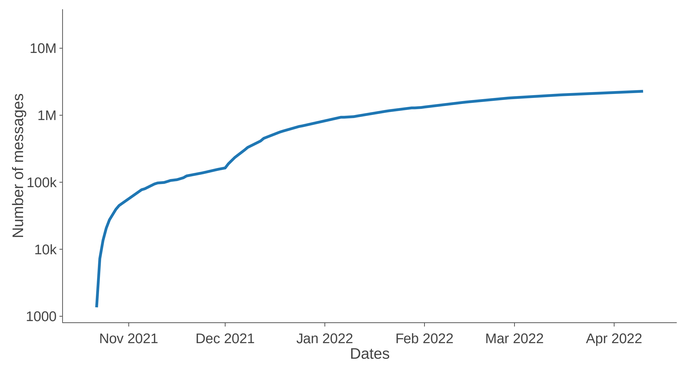
Classifying images: Ω represents our CNN for image classification. It is based on a pre-trained AlexNet. We fine-tuned this model with a COVID-specific training dataset and modified the last fully connected layer to produce outputs for our three classes (memes, posts, and others).

DATA & RESULTS

We adopted a snowball strategy for collecting data, we started with 13 public channels likely related to spreading misinformation about COVID (manually verified).Then, we augmented the list considering the source of forwarded messages. Our 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
- 522 documents (mostly in the PDF and DOCX formats).

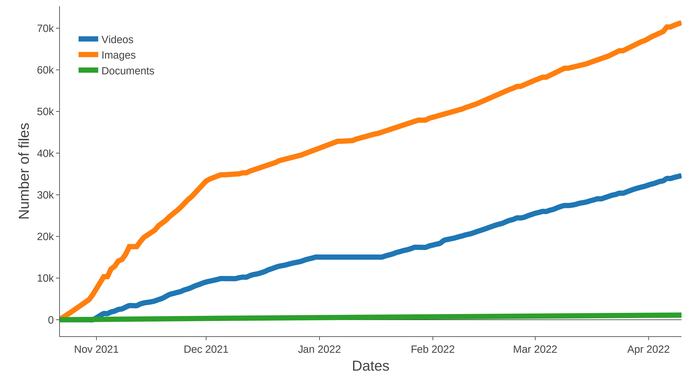
Accumulated text messages: Daily distribution of collected Telegram messages.



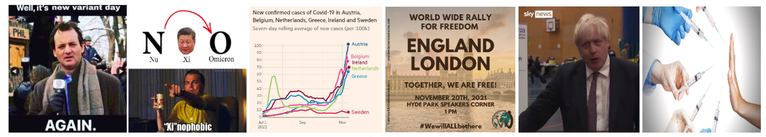
**First version of our dataset, from October 2021 to end of December 2021. Since our collection is still running, we updated some of the graphs.

RESULTS

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



CLASSIFIED IMAGES



Memes Posts Others

TOP-5 WORDS OF TELEGRAM MESSAGES

Word	F1	F2	Log-likelihood
vaccine	132959	26026	179230
vaccinated	26075	9086	71786
vaccines	90899	9471	54248
vaccination	48375	7312	46913
unvaccinated	7366	4063	35175

TOP-5 HASHTAGS AND MENTIONS

Hashtag	(%)	Mention	(%)
#KAG	12.84	@WeTheNews	5.66
#WeAreTheNewMedia	5.43	@PookztA	5.42
#Omicron	1.00	@PatriotArmy	5.42
#COVID19	0.98	@SergeantRobertHorton	4.26
#WWG1WGA	0.84	@disclosetv	4.11

TOP-5 WORDS OF VIDEO TRANSCRIPTS

Word	F1	F2	Log-likelihood
vaccine	132959	5943	44861
vaccinated	26075	2210	19408
people	15749678	19211	18879
vaccines	90899	2759	18751
virus	192947	3157	17664