How to design a collection strategy for monitoring disinformation and hate speech in Telegram

Elisa, Muratore,[1,2] Thomas, Louf,[1] Elisa, Leonardelli,[1,2] Riccardo, Gallotti,[1,2]

1. Bruno Kessler Foundation, Trento Italy 2. Hatedemics consortium, CERV-funded project

Social media platforms have become central spaces for user engagement, but this expansion has also amplified issues like disinformation and hate speech. As research interest in these topics grows, obtaining social media data has been crucial, yet platforms are increasingly restricting such access. In response, researchers are exploring alternative sources, such as Telegram, which offers an open-access API. However, Telegram's data collection dynamics present distinct challenges. Unlike popular platforms like X, Telegram's structure is fragmented, with each channel operating as a self-contained stream. To access content, users must actively navigate to individual channels, making it challenging to identify and select relevant data to collect. Additionally, the lack of keyword search and the risk of data volume explosions when retrieving messages complicate the process, requiring a thoughtful approach to select relevant chats.

In this study, we tackle these challenges by introducing a systematic, and task-driven strategy to effectively collect data from channels linked to disinformation and hate speech. This work is part of the Hatedemics project [1], which aims to combat hate speech and disinformation through AI-based technologies. By leveraging Telegram data, we seek to fill the gap left by increasingly restrictive data policies on other platforms, while also exploring the dynamics of harmful content dissemination in less-regulated digital environments. The final goal of this project is educational, with the aim of creating a didactic and multilingual dataset to train citizens in actively combat hate speech and disinformation.

To facilitate effective data collection from Telegram, we propose a systematic, task-oriented, and target-based strategy consisting of the following steps: seed identification, channel expansion, priority assignment, message retrieval, and evaluation of disinformation and hate speech scores. Given the lack of keyword-based search support in the API, we developed a tailored approach focused on public channels likely to disseminate disinformation and hate speech, particularly targeting minority groups. The process begins with a seed set of known problematic channels curated by experts from the project consortium (Fig. 1a). This set is expanded by identifying channels with similar names and further enriched using a snowballing technique, inspired by [2], which leverages Telegram's channel recommendations based on users' overlap (Fig. 1b). This lightweight channel data is then used to construct the network with channels as nodes and recommendation links as directed edges. Given resource constraints, the more expensive message collection step requires a thoughtful approach. We select the most influential channels based on criteria such as in-degree score, distance from seed, and a custom participation score (Fig. 1c). After normalising these metrics, we compute the priority score through a linear combination of the normalised values. Messages from the top-ranked channels are then retrieved using the Telegram API, resulting in the collection of over 10 million messages (Fig. 1d). These first steps of the collection were implemented using the Telethon library [3] for efficient interaction with the API. To assess content, we compute two key metrics (Fig. 1e): the Infodemic Risk Index (IRI) [4], which gauges the reliability of shared sources, and a hate score (HS) [5], reflecting the proportion of hate speech messages.

We evaluate the effectiveness of our strategy by analysing the Pearson correlation between priority scores and both hate score and IRI (Fig. 1f). Preliminary results indicate that approximately half of the collected channels exhibit relatively high IRI (≥ 0.3) and HS (≥ 0.1), confirming the strategy's potential to successfully identify relevant channels. The most strongly correlated features with these scores include distance from seed channels, eigenvector centrality, in-degree, and participation score—validating the relevance of the selected metrics in computing channel priority. As a very next step, we aim to compare alternative methods for priority score computation by introducing a validation mechanism, allowing us to select the most effective approach for this specific task.

Ultimately, our proposed strategy offers a scalable, task-driven, and data-informed method for collecting Telegram data, addressing the growing challenge of restricted access on mainstream platforms and supporting research on disinformation and hate speech.

References

[1] Hatedemics project, CERV-2023-CHAR-LITI-SPEECH. https://hatedemics.eu/

[2] Baumgartner, Jason et al. (2020). The Pushshift Telegram Dataset. 10.48550/arXiv.2001.08438.

[3] Telethon's Documentation. Telethon. https://docs.telethon.dev

[4] Gallotti, R., Valle, F., Castaldo, N. et al (2020). Assessing the risks of 'infodemics' in response to COVID-19 epidemics. Nat Hum Behav 4, 1285–1293. https://doi.org/10.1038/s41562-020-00994-6

[5] Casula, C., & Tonelli, S. (2024). A Target-Aware Analysis of Data Augmentation for Hate Speech Detection. arXiv preprint arXiv:2410.08053.

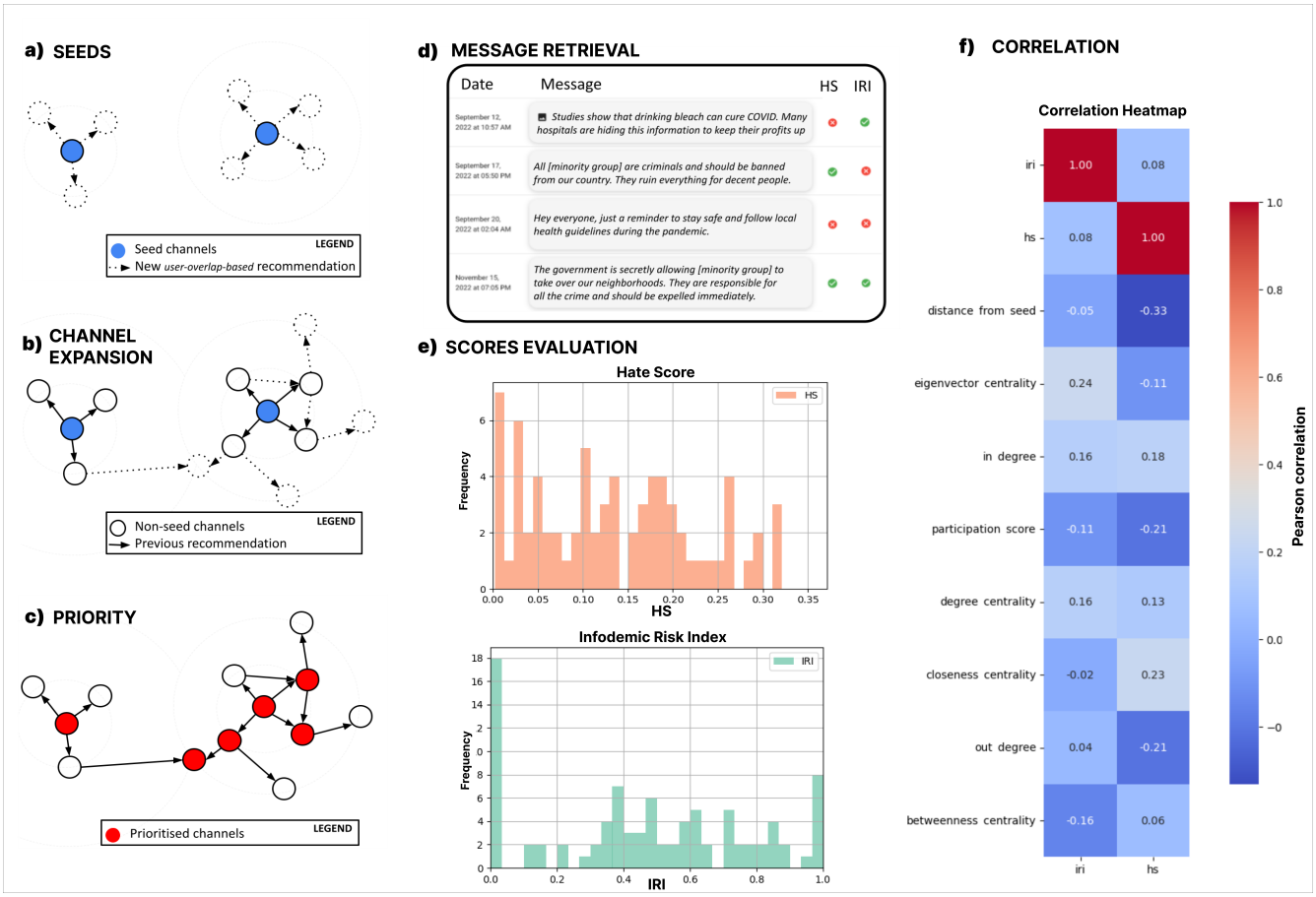

Figure 1: Step of the data collection strategy. **a)** Identification of seeds and initial expansion. **b)** First step of channels snowballing through channel recommendations. **c)** Top-ranked channels by priority scores for message retrieval. **d)** Example of retrieved messages, anonymised via HMAC algorithm and NLP techniques. **e)** Histograms displaying the IRI and the HS of the considered channels. **f)** Pearson correlation heatmap showing the relationship between IRI, HS and various metrics.