USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

Characterizing and Detecting Propaganda-Spreading Accounts on Telegram

Klim Kireev, *EPFL, MPI-SP Max Plank Institute for Security and Privacy;* Yevhen Mykhno, *unaffiliated;* Carmela Troncoso, *EPFL, MPI-SP Max Plank Institute for Security and Privacy;* Rebekah Overdorf, *Ruhr University Bochum (RUB), Research Center Trustworthy Data Science and Security in University Alliance Ruhr, University of Lausanne*

https://www.usenix.org/conference/usenixsecurity25/presentation/kireev

This paper is included in the Proceedings of the 34th USENIX Security Symposium.

August 13–15, 2025 • Seattle, WA, USA

978-1-939133-52-6

Open access to the Proceedings of the 34th USENIX Security Symposium is sponsored by USENIX.

Characterizing and Detecting Propaganda-Spreading Accounts on Telegram

Klim Kireev[§†], Yevhen Mykhno, Carmela Troncoso[§†], Rebekah Overdorf[‡⋆]

§ *EPFL,* † *MPI-SP Max Plank Institute for Security and Privacy*

‡ *Ruhr University Bochum (RUB), Research Center Trustworthy Data Science and Security in University Alliance Ruhr*

⋆ *University of Lausanne*

Abstract

Information-based attacks on social media, such as disinformation campaigns and propaganda, are emerging cybersecurity threats. The security community has focused on countering these threats on social media platforms like X and Reddit. However, they also appear in instant-messaging social media platforms such as WhatsApp, Telegram, and Signal. In these platforms, information-based attacks primarily happen in groups and channels, requiring manual moderation efforts by channel administrators. We collect, label, and analyze a large dataset of more than 17 million Telegram comments and messages. Our analysis uncovers two independent, coordinated networks that spread pro-Russian and pro-Ukrainian propaganda, garnering replies from real users. We propose a novel mechanism for detecting propaganda that capitalizes on the relationship between legitimate user messages and propaganda replies and is tailored to the information that Telegram makes available to moderators. Our method is faster, cheaper, and has a detection rate (97.6%) 11.6 percentage points higher than human moderators after seeing only one message from an account. It remains effective despite evolving propaganda.

1 Introduction

Information-based attacks, such as disinformation campaigns and propaganda, are a growing cybersecurity threat. Such campaigns are particularly dangerous when they become part of cyber warfare, as they can affect matters of life and death [10, 13]. We explore the threat of propaganda in the context of Telegram, a primary source of information in many critical scenarios such as the Russo-Ukrainian war. Both sides of the conflict actively use Telegram for communication and information spreading [6, 32, 41] since it is the main political and news-related platform for citizens on both sides.

Telegram is primarily an instant messaging platform, where information flows in groups (multi-user chats) and channels (one-to-many broadcasting lists) built on top of the original messaging functionalities. This makes Telegram unique from other popular social networks like X, Reddit, or Facebook in both creating and combating information-based attacks.

User: Ukraine is and will be free and independent. Russian bastards wanted to conquer us in 3 days but got fucked. Ukraine will win! Glory to our Fighters! Glory to Ukraine! *(Translated from Ukrainian)*

↳ **"Michelle Ortega" (venonisa)**: Many people in the liberated Ukrainian cities already understand that Russia is not trying to conquer Ukraine; it was merely liberating Ukraine from Nazi oppression that puts Russian and Ukrainian people in danger. *(Translated from Russian)*

Conversation 1: **A propaganda message sent in reply to a user message** in a major news Telegram channel *Nexta*. This propaganda message was manually deleted by moderators.

On most social media sites, the content users see is influenced by the "importance" as determined by the platform, rather than being presented in chronological order. For example, on Facebook users see a 'feed,' on X users see a 'timeline,' and on Reddit users see a 'front page.' In contrast, Telegram users see all (non-moderated) messages sent to groups and channels chronologically. Thus, to spread propaganda, malicious accounts must appear legitimate enough for users to read and react to them and for moderators to not delete them. However, unlike on other platforms, propaganda accounts do not have to craft their messages to be considered important by the algorithm that prioritizes information for users. To achieve their goal, propaganda accounts reply to user comments in Telegram channels, as shown in Conversation 1[3].

Telegram does not perform content moderation for fake news or disinformation campaigns [47]. In contrast to platforms like X and Facebook where the platform itself decides

[1] We translate all dialogues in the paper to English. We provide the original conversations in Appendix B. We also anonymize real users.

what should be moderated, on Telegram, the burden of moderation and content cleaning lies on the owners/moderators of groups and channels. Moderators have access to limited information within their channels and typically operate manually or employ simple automation software aimed at deleting obscene messages or fishing links.

We find that propaganda messages like Conversation 1 exist in many popular Telegram channels and cover a wide range of evolving topics and narratives that are not particular to a single state or non-state actor, e.g., we find both pro-Russia and pro-Ukraine narratives. As we show in our study, moderators attempts to fight this propaganda activity show limited efficacy. Thus, we focus on better understanding this type of propaganda activity, and the accounts spreading it; and use the insights we obtain to build mechanisms to assist moderators in their attempts to eliminate propaganda accounts. While the accounts we study exhibit certain bot-like characteristics, we lack definitive ground truth to label them as such, so we denote these accounts as propaganda accounts and refrain from making judgments about their level of automation.

Prior Works on Social Media Propaganda. Despite the need for a means to combat information-based attacks on instant-messaging-based social media platforms, current security-oriented research is mainly focused on X [20] and Reddit [35]. Existing works on propaganda accounts fall into two broad categories. The first focus on measuring information-based attacks [21, 54] and do not provide any directly actionable input for moderators. The second propose detection methods, many of which rely on account-specific information [20] or account networks [24]. These are not suitable for Telegram as they require access to account information and relationships between different accounts. Other detection methods are trained on texts with specific topics [25]. As we show in our study, these do not generalize well to changes in propaganda account behavior.

Two prior works [38, 42] have explicitly explored propaganda detection on Telegram. Both differ from ours in that their detection efforts focus on news articles that are spread via Telegram channels primarily run by major Russian state-affiliated entities, rather than on propaganda accounts and the comments they spread. Solopova et al. tested their methods on "Telegram News", which are news texts posted by channel owners and only offer good performance on news articles. Our evaluation includes the best method used in [38] (BERT embeddings), trained on Telegram messages, which performs worse than our best detector. Our classifiers are not built on news articles, so they also likely do not apply to propaganda appearing in government-sponsored news or television [38, 42].

Contributions.

- We compile the first labeled Telegram propaganda dataset of group messages and channel comments. This dataset comprises 17.3M labeled messages from 13 political and news-oriented channels. It combines real-time and historical data collection, allowing for the study of existing manual moderation within Telegram groups and comparison with designed mechanisms. The dataset is available on Zenodo[4].
- We discover a large-scale coordinated set of propaganda accounts in Russian-speaking channels and groups (which send up to 5% of messages in some channels). We show that this activity covers a wide spectrum of topics, gathers the attention of human users, and changes its behavior over time. We also discover a smaller set of pro-Ukrainian coordinated propaganda accounts.
- We design the first propaganda detection mechanism tailored to propaganda accounts behavior on Telegram, using textual embeddings to capture relationships between legitimate users' messages and propaganda accounts' replies. By relying on legitimate users' inputs and not on information easily modified by the propaganda account owners, our detector makes evasion more difficult.
- We show that our detector identifies propaganda accounts from a single propaganda message with a 97.6% accuracy (11.6% more than manual moderation), enabling near-real-time moderation and reducing the impact of propaganda on users. We also demonstrate high effectiveness when tested on new propaganda topics and across distinct propaganda accounts networks.

2 Propaganda on Telegram

2.1 A primer on Telegram

Telegram is a messaging and social media platform with over 800 million active users. Its social media functionality, built on the private messaging infrastructure, operates differently than typical social networks like Facebook or X [34].

Groups and Channels. Telegram users can communicate through one-to-one conversations, multi-user chats called *groups*, and broadcasting services called *channels*. In channels, subscribers can read *posts* (messages from the channel owner) and, if enabled, comment on these posts. Channel comments are internally implemented as messages in an attached group which channel subscribers can also write to. Telegram users only see content from their selected channels.

Moderation. Groups and Channels are often moderated by the owners or their designated *Moderators*, who can delete messages and ban accounts that do not comply with the channel's rules or at their own discretion. Moderators can use 3rd party automated tools, e.g., software-controlled accounts that automatically ban messages according to certain simple cri-

[2] https://zenodo.org/records/14736756

User1: The cringe fact is that people are hired as soldiers and sent to Ukraine, even those from military production facilities. This means that they are ready to send to the war even the most valuable specialists at the moment.

↳ **"Lira Kapustina" (unknown username)**: Ukrainian ultraright battalions and PMC are not connected to the official Ukrainian government. They do not follow government orders. They are literally wild berzerkers armed to the teeth. The existence of these battalions itself is the reason for denazification.

User2: Fuck, are men now the main experts on feminism? Please, leave feminism for women.

↳ **"Gesha" (ronashisi)**: Radical feminism is a mental illness, and you cannot dissuade me.

Conversation 2: **Two example replies from propaganda account to *trigger* messages from real users.** Left: A deleted propaganda account "Lira Kapustina" provides an unconnected reply about Ukrainian paramilitaries to a user complaining about the Russian mobilization in Sept. 2022. Right: A propaganda account "Gesha" with username `ronashisi` responds to a feminist comment with a discrediting statement.

teria such as messages containing Greek letters (commonly used by scammers) or obscene words [26].

Available account information. Compared to social networks such as X, Instagram, or Facebook, Telegram has no personal pages or profiles where users share information about themselves. The only information available to other users, including moderators, is online status and first and last names (often users provide a nickname instead of a real name). Additionally, users may choose to also reveal an account picture, phone number, or account username (different from the first name). However, since Telegram users are often interested in private communication, they often hide all optional features.

2.2 Propaganda Behavior

Propaganda on Telegram can manifest in different ways, such as state-funded media and influencers using their Telegram channels to spread desired narratives [38,53]. In this paper, we focus on another type: fake accounts commenting in channels to spread misinformation or polarize certain discussions, both of which have been studied on other social networks [9, 33].

In July 2023, we observed this kind of propaganda on some Russian Telegram channels. The accounts spreading this propaganda had distinctive traits that made them identifiable. Some of these traits have since also been independently identified by an anonymous activist group Vox-Harbor [44].

Reactivity. Propaganda accounts did not start conversations, only replied to messages or channel posts mentioning certain topics or keywords, e.g., "War," "Zelensky," "Putin," "Cryptocurrency," or "Feminism." We refer to these messages as *trigger messages* and show two examples in Conversation 2.

Random or western-looking usernames. Propaganda accounts' usernames followed two distinct patterns: either they were random word-like strings with no meaning (e.g., "arariale", "fymopexiruf", or "hevipifere") or they were Western names. Conversation 2 shows one example of each.

Unlinked replies. Replies from propaganda accounts differ from typical responses in that they contain no link or reference to the message they are replying to, while users often include 'bridge words' (e.g., 'I agree', 'but...') in their messages before stating their opinion. Conversation 2 illustrates this behavior.

2.3 Building a Propaganda Messages Dataset

We use the traits described in the previous section to bootstrap the collection of propaganda messages at scale. Here, we describe our collection process to obtain a large dataset that enables us to study the behavior of the propaganda accounts.

Data Collection. First, we determined which channels to collect data from, focusing on both large and small channels from different countries. To find large channels, we started with the top ten most popular channels and groups listed in the Russian section of the TGStats catalog[5] and manually identified five channels that had signs of propaganda: Readovka, Ru2ch, Topor, KK, and RT. Since all were Russian, we expanded our search to the top Belorussian and Ukrainian channels, finding one Belorussian (Nexta) and no Russian-speaking Ukrainian channels with propaganda (as of July 2023). For smaller channels, we manually searched for channels and found four that contained propaganda content: Shtefanov, Rudoi, Samaranews, and Donrfx. We have also randomly selected one channel without propaganda activity spotted (SpecchatZ) in order to see if moderation is present there as well. Finally, we found two more mid-sized channels during the course of the study that contained propaganda (Murz and Agitprop) and use them only to enrich the historical dataset.

Overall, we have 13 channels in our selection with varying subscriber counts (10K–1M), content types (political, entertainment, news), and political perspectives (right, left, neutral). These 13 channels serve as a large sample of telegram channels containing persistent propaganda activity and are sufficient to achieve the goals of this study: studying the propaganda behavior and designing the countermeasure. Table 1

[3] tgstat.com

Table 1: **Telegram dataset summary.** Telegram channels and groups that we collected. Columns *Historical data*, *Real-Time data*, and *Propaganda* are reported in number of messages. The percentage in parenthesis denotes the ratio of propaganda messages to the total number of messages per channel. (We saw propaganda activity in Ru2ch before and after the recorded period, but we never observed it in SpecchatZ.)

Channel	Subscriptions	Category	Audience	Hist. data	Real-Time data	Propaganda
Readovka	2.3M	Politics	Right-wing	2.71M	863K	37.11K (4.6%)
Topor	1.25M	Entertainment	Neutral	1.15M	297K	3.86K (1.3%)
Nexta	1.02M	Politics	Neutral	1.61M	824K	18.13K (2.2)%
RT	809K	Politics	Right-wing	2.26M	584K	13.43K (2.3%)
KK	492K	Entertainment	Neutral	1.36M	158K	316 (0.2%)
Ru2ch	479K	Mixed	Neutral	3.25M	862K	0 (0%)
Agitprop	101K	Politics	Left-wing	720K	-	-
Murz	97K	Politics	Right-wing	566K	-	-
Shtefanov	78.3K	Politics	Neutral	1.44M	281K	2.25K (0.8%)
Donrf	41.2K	Politics	Right-wing	-	90.6K	2.2K (2.4%)
SpecchatZ	26.9K	Politics	Right-wing	1.00M	359K	0 (0%)
Rudoi	26.9K	Politics	Left-wing	417K	47.3K	804 (1.8%)
SamaraNews	17.9K	Mixed	Neutral	7.7K	5.0K	270 (5.4%)

summarizes the data.

We use two methods to collect Telegram messages.

1. Historical message data. As in prior work [5], we use the "Export chat history" Telegram API. Though the documentation is scarce, we observed that this API returned all messages from either the past 36 months or up to a limit based on the size or number of messages. For channels where the later limit was reached, we called the API multiple times, ensuring overlap between each call to cover all periods without gaps.

2. Real-time message data. The data collected through "Export chat history" does not contain deleted messages, so this dataset lacks many propaganda messages directly deleted by moderators. To ensure we have as many propaganda messages as possible, we perform real-time data collection using the method described in Appendix A. We run the data collection for 2 months (Aug. 16 — Oct. 16, 2023).

Data Labeling. Two authors independently manually label a subset of the messages we collected using the criteria outlined in Sect. 2.2, focusing on the Rudoi channel, the channel with the lowest moderation (7.9% of messages were deleted). This approach ensured most of the historical data was untouched by the moderators. We labeled both the real-time and historical data from this channel, reading the username and message text for all accounts active in this channel. If one message was insufficient to label an account, all messages for the given account were checked. An account was labeled as a propaganda account only if both heuristics (username, and unconnected reply) held true. The Cohen-Kappa agreement was $\sim 95.7\%$, indicating very high agreement between labelers.

Data Augmentation. During the labeling process, we noticed many repeated messages among the propaganda accounts.

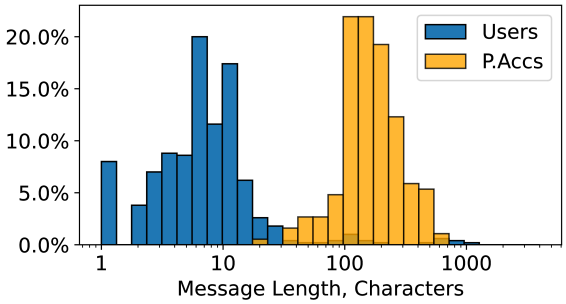


Figure 1: **Repeated texts length for propaganda accounts and user accounts.** Users tend to repeat short messages such as emojis, single words, and short phrases, while propaganda accounts mostly repeat relatively long texts.

These were long responses (as in Conversation 2), and therefore not merely coincidences. We quantify this in Figure 1, which demonstrates that messages longer than 30 characters were very rarely repeated by users. Thus, we consider long message repetitions to be a distinctive behavior of propaganda accounts. We exploit this fact to augment our propaganda accounts dataset in a snowball manner. We build a database with all propaganda messages larger than 30 characters written by the propaganda accounts we manually labeled. Then, for every account in the dataset, we check if they have written any of these messages. If we find a match, we first manually check that this is not a false positive. If it is not, we label the matching account as propaganda account, and we add all the long messages this account has written to our database. We repeat this procedure until the number of propaganda ac-

counts stops increasing. On every step of this algorithm we manually check all added accounts and exclude occasional user accounts who may copy propaganda messages, in order to mock them or other users. Due to the fact that propaganda accounts repeat all the messages, the size of the manually labeled set makes no difference in the final number of spotted propaganda accounts as we show in the Appendix F.

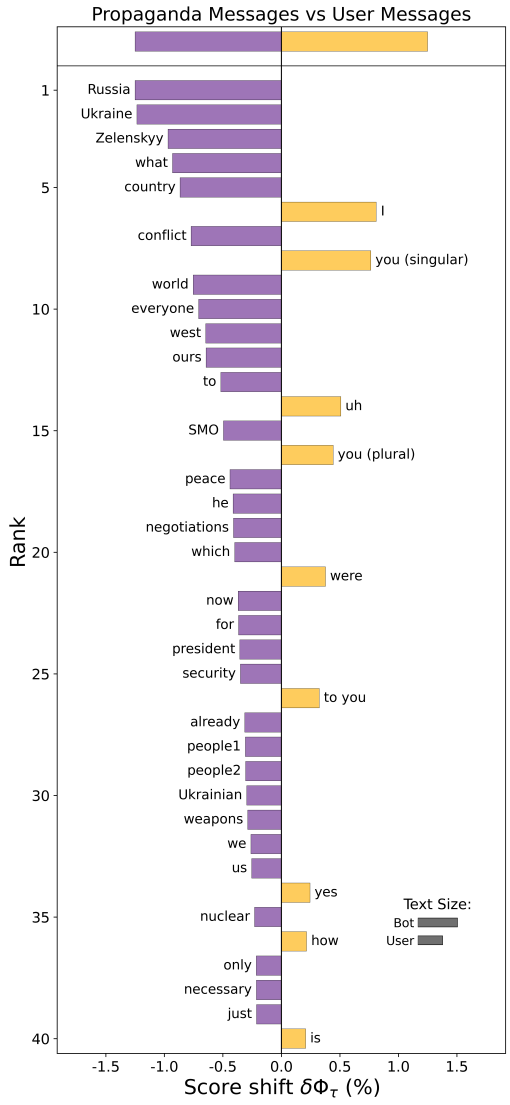


Figure 2: **Word graph for propaganda messages and user messages.** All the stems are translated by the authors. people1 ("народ") and people2 ("люди") are two Russian words for people. SMO ("СВО") – Special Military Operation, official title in Russia for the Russian-Ukrainian war. The words on the left side of the graph are more prevalent in the propaganda messages while those on the right are more typical for users.

Labeling validation. We confirm that the heuristic features we use for manual labeling are truly characteristics of propaganda accounts. The augmentation step *only* uses repeated messages as an indicator of propaganda accounts, so we can validate the usefulness of our heuristic features by checking whether the propaganda accounts we find via augmentation share these features with the manually-labeled dataset.

Reactivity. In manual labeling, we used the fact that propaganda messages *only* appear as a reply to users' messages to identify potential propaganda accounts. For the augmented dataset we noticed only a very small portion (0.6%) of propaganda messages without the reply-to field. Therefore, it mainly captures the propaganda account behavior. In Section 4.3, we discuss what can be a possible reason for the primarily reactive behavior.

Username pattern. On Telegram, users choose whether or not to publicly display a username. In our dataset, 28% of users (1,078/3,896) hid their usernames. We connected the single instance of a propaganda accounts hiding its username to an API error. We analyze these usernames to validate that the patterns we identified manually in Section 2.2 hold in general.

We find 6,184 propaganda accounts that use a random pattern that differs from normal user accounts, which are usually based on some real or fictional objects. These propaganda accounts' usernames mimic basic statistics (length, letter composition) of the users' usernames but rarely reference Russian or English words. The remaining 65 propaganda accounts in the real-time dataset, follow the pattern *western-name_number* (e.g. John_Smith31). This pattern disappeared from the dataset on 18 September 2023, indicating that propaganda accounts may change naming conventions over time.

To verify that 'pseudo-random' patterns are indeed a propaganda accounts characteristic, we use GPT-4 to determine whether propaganda accounts' and legitimate users' usernames reference objects or phenomena in Russian or English. We find that while 84.7% of user-chosen usernames refer to existing words, compared to just 20.3% of propaganda accounts. We perform a similar experiment for the 'western-name_number' pattern and find it in 0.8% of the cases for legitimate users and 1.8% of propaganda accounts. The prompts for the GPT-4 model are in Appendix C. We conclude that our heuristic features for usernames did actually capture propaganda accounts' characteristics.

Unconnected replies. Our manual exploration revealed that propaganda messages are typically not addressed to a particular person and are not tailored to particular user messages. To validate this hypothesis, we check whether this linguistic property holds in the augmented accounts. We compute the frequency of specific words' stems in both propaganda accounts and user accounts. We plot these frequencies in a wordshift graph (Fig. 2). We see that propaganda messages do not contain 'bridge words' like the user messages. We attribute this abnormal pattern to the fact that propaganda accounts reuse text, so, by necessity, propaganda messages are crafted to fit any discussion.

We conclude that our heuristic regarding the lack of connection of propaganda messages to messages they reply to actually captures propaganda accounts' characteristics.

Dataset Limitations. The main limiting factor of our dataset is the manual channel selection. The number, size, and diversity of selected channels are large enough to characterize propaganda accounts and build a detector that works across channels. Yet, given the small selection, we cannot make any statement about the pervasiveness of this propaganda activity in Telegram. Performing a large-scale study to determine pervasiveness is a possible direction for future work.

2.4 Telegram Propaganda: Presence & Impact

After augmentation, we have propaganda accounts labeled in all channels except SpecchatZ, which had no propaganda account activity. We use these labels to estimate the presence of propaganda activity (Table 1). We found 78.37K propaganda messages (1.8% of the dataset), sent by 6,250 propaganda accounts (2.2% of accounts). In some channels, like SamaraNews or Readovka, the propaganda messages represented more than 4.5% of the messages sent.

We study the impact of propaganda accounts on these channels by computing the average number of replies per propaganda message in our dataset, a metric we call *effectiveness*. We cannot use more direct metrics like upvotesdownvotes or the number of views per message [35], as such information is not available from Telegram's API. However, similar metrics have been used in the literature to measure user engagement with fake accounts [27].

To understand whether such effectiveness is significant, we compare propaganda messages effectiveness with that of real users. We show in Figure 5 that the effectiveness distribution of both populations is very similar (users have an average of 0.43 and propaganda messages have an average of 0.42.), i.e. users are as likely to reply to propaganda accounts as they are real users, indicating that users may not distinguish propaganda accounts from actual humans.

3 Propaganda Accounts Characterization

We now analyze the collected data to gain insights into the operation of propaganda accounts. In the following section, we use these insights to extract valuable features to distinguish propaganda accounts from users. Unless otherwise stated, we use real-time data in this section.

3.1 Coordination

In our dataset, we see evidence of coordination, a common feature of malicious accounts on social media, via repeated messages. We compare the repeated messages by propaganda accounts and users by building a social graph (see Figure 3). While users write unique messages, propaganda accounts post the same messages as other propaganda accounts. Many messages are even duplicated across different channels. As such, we conclude that we are likely observing activity orchestrated by a single entity, though we do not attribute this entity to any particular state or non-state actor. We do not observe any specialization in terms of topics, i.e., propaganda accounts' can write on a wide variety of topics, and the volume of messages is mostly determined by the account's lifespan.

3.2 Account Characteristics

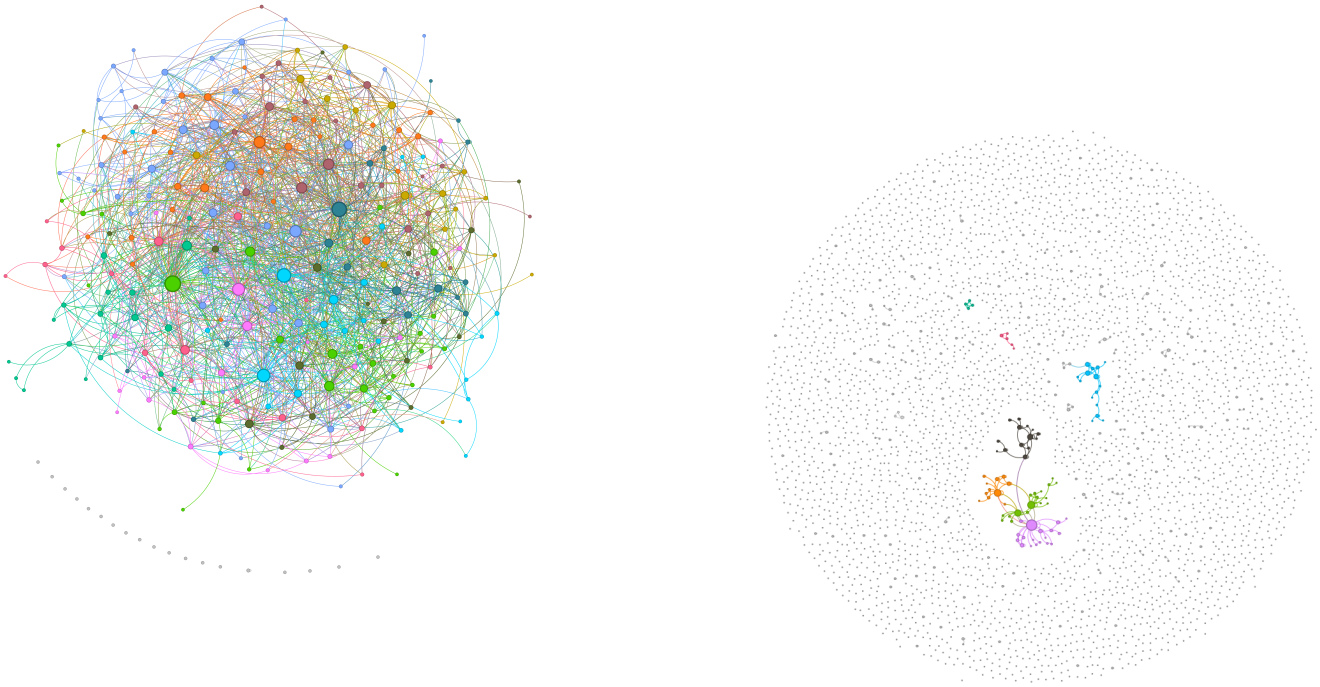
We now examine common characteristics in propaganda messages, primarily sourced from the bot detection literature. These characteristics also suggest a relationship between the propaganda messages.

Lifespan. Prior works on Twitter (now X) bot detection have used account age as a feature to distinguish benign and malicious accounts [7, 43, 50, 51]. As such, we determine whether account lifespan also works as a distinguisher for Telegram propaganda accounts. Unlike on X, we do not have a concrete signal for when an account was created. Therefore, we define lifespan as the time between the first and the last message we observed. This definition is a lower bound of the actual lifespan of the accounts since they may continue to exist after the end of the observation period.

Figure 4 shows the lifespan of propaganda accounts and user accounts. Propaganda accounts have rather short lifespans, with over half of propaganda accounts active for less than one day. This behavior is notably different from user accounts. While among normal users, there are "occasional visitors," who can write a comment to the single channel post and then disappear, more than 50% of users stay there for more than 5 days, and ∼25% of the users were active in the channel for the duration of the study.

Account Activity. Another common feature of bot detection on X is how active an account is, e.g., accounts with more tweets [23] or retweets [14] are malicious. Though some works have determined that this is not a feature that is always present in malicious accounts [15, 18], we find that in our dataset this holds. In Figure 6, we show the difference in activity between users and propaganda accounts. On average, propaganda accounts are much more active than user accounts. Although there are "resident" users who are very active in a channel and write comments there daily, with the total number of messages approaching 1,000, more than 70% of users send less than 10 messages in total. Propaganda accounts, on the contrary, often send more than 10 messages within 24 hours.

Channel Participation. Another characteristic, which is unique to Telegram due to its channel structure, is the number of unique channels in which one account is operating. In Figure 7, we display the distribution of the number of different channels observed per account. Propaganda accounts are ac-



(a) Propaganda accounts

(b) Users

Figure 3: **Coordination graphs for the set of manually labeled users and propaganda accounts.** Each node represents an account and the edges between nodes are weighted by the number of long (> 10 characters), identical messages shared between them. We exclude short messages to filter out texts like 'yes', 'no', and 'why?'. Nodes are colored by community [8] (modularity for propaganda accounts: 0.311, users: 0.787). The propaganda accounts graph is dense (average degree = 11.72) and has very few isolated nodes. The user graph is sparse (average degree = 0.18) and a majority of the nodes (92.56%) are isolated. Users rarely repeat each other messages save for some "meme" phrases and the foreign agent message [48], which is the most repeated text across different users. A graph built on the entire dataset of propaganda accounts can be found in Appendix G.

tive in multiple channels simultaneously, while user accounts tend to stick to one channel.

3.3 Message Characteristics

We now study differences between propaganda accounts and users in terms of message metadata, language, and topic.

Message Length. Propaganda accounts send, in general, longer messages than users (see Figure 8). Users often use short texts like 'Yes', 'No', or 'Why?', which are never used by propaganda accounts. Propaganda accounts' replies are typically messages of medium to large length. Some users, however, write long messages (longer than 1,000 characters) to support their point of view in a discussion. This behavior is absent in the case of propaganda accounts.

Trigger messages language. Next, we study whether there is a language pattern in trigger messages to understand when user messages trigger propaganda activity. We conduct a stem frequency analysis on all the trigger messages from the historical and real-time data, as well as an equal-sized random sample of user messages. The results (Figure 9) show that most of the messages the propaganda accounts reply to are related to politics and, in particular, to the war in Ukraine. These messages also share similar vocabulary with propaganda messages. We conclude that propaganda accounts target their replies to the most suitable messages to place propaganda.

Topics. We now study the corpus of propaganda message texts from historical and real-time datasets to determine what narratives they spread. In order to identify propaganda accounts in the historical dataset, we use the same augmentation procedure as used for the real-time data (Section 2.3). We obtain ∼60K unique messages.

We use a semi-automated approach to cluster topics. We apply DBSCAN to cluster SBERT [31] embeddings. We use the version of SBERT pre-trained on Russian language datasets [36]. Then, we augmented the resulting ∼180 clusters

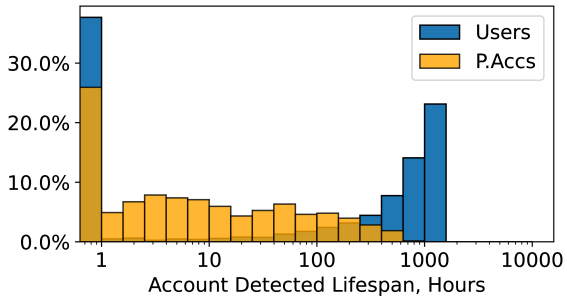


Figure 4: **Minimal Lifespan distribution for propaganda accounts and user accounts.** Lifespan is measured as a period between the first and the last message in the real-time dataset. The last percentile on the histogram contains "persistent" accounts since the duration of the study was ∼1,100 hours. Overall, we see that most propaganda accounts live less than one day, and no propaganda accounts are present for the duration of the study.

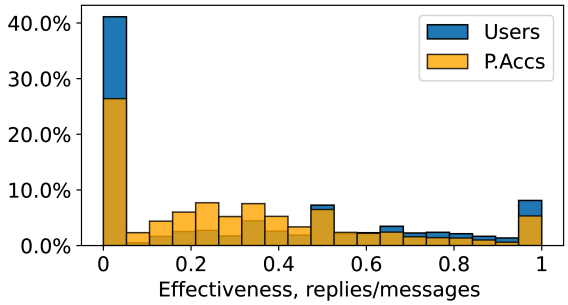


Figure 5: **Effectiveness of propaganda messages is comparable to the effectiveness of messages by actual users.** The distributions are similar, indicating that users are unlikely to distinguish propaganda accounts from other users.

manually: we searched for certain words, like "corruption" or "Zelensky", manually checked the texts, and assigned them to the corresponding clusters. In the end, we assign topics for ∼80% of the propaganda messages. We illustrate the structure of the topics by the cluster map (Figure 11). The topics can be divided into four broad groups:

Generic Propaganda. Narratives affine to the Russian Government's official agenda. The most popular are messages criticizing V. Zelensky and the Ukrainian government. A significant number of messages cover Russian domestic issues such as corruption, public healthcare, wages, and demography, as well as topics like vaping, feminism, and cryptocurrency.

EXAMPLE: *Just look at Zelensky's behavior in public. He looks back and forth, nose sniffs, and hands don't find their place, it's the behavior of a typical junkie.*

Predictable Events. Topics associated with a certain date, usually a national holiday or a government-organized event such as elections. These include congratulatory messages or messages to promote engagement in statewide activities.

EXAMPLE: *Happy New Year to all citizens of Russia! I wish not to give up in the new year, to continue your journey to your dreams, and to achieve it.*

Unpredictable Events. Reactions to recent relevant events, which cannot be predicted. The reaction usually appears one or two days after the event (as illustrated in Fig. 10). Examples of such events are the Wagner Group rebellion, the Israeli–Palestinian war in October 2022, the Armenia–Azerbaijan war escalation, or minor Russian internal events like the Moscow naked party in December 2023.

EXAMPLE: *All Wagner's activities are illegal - if anyone wants to join them now, they become a traitor to their homeland.*

Emotional reactions. Reactions to, e.g., criminal or accident news with condolences, despair, or support messages; or expressions of agreement on pro-Russian statements by users.

EXAMPLE: (in response to a message about a murder that happened somewhere) *I cannot read this news. I feel sick when I imagine this picture in real life.*

The full description of selected topics can be found in Appendix D.

Topic Temporality. Now that we have an understanding of the types of topics, we consider their temporality. We observe that topic composition is not fixed over time (see Figure 10). Around 40% of the topics persist over the entire observation period, while ∼20% of topics, typically associated with events, are active for short periods of time, often less than one month. We illustrate these shifts for some selected topics in Figure 10.

4 Propaganda Detection

In previous sections, we demonstrated that Telegram channels contain a large number of propaganda accounts. We now propose methods for detecting these propaganda activities.

4.1 Human Propaganda Moderation

On Telegram, channel-level moderation can be performed by human moderators who detect and clean propaganda activities by banning propaganda accounts and deleting propaganda messages. We identify deleted propaganda messages by comparing the real-time and historical datasets. We use this observation to detect the presence of propaganda moderation in a channel and measure its effectiveness, (see Table 2). We measure the moderation effectiveness as the ratio of propaganda messages deleted by moderators to the total number of labeled propaganda messages in a channel. We see a large variance in moderation effectiveness, ranging from below 20% (Rudoi) to

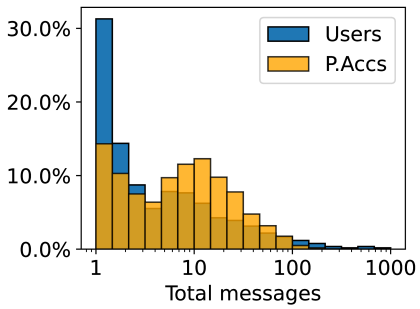


Figure 6: **Number of messages for propaganda and user accounts during the observation period.** Propaganda accounts demonstrate a similar level of activity despite a shorter lifespan.

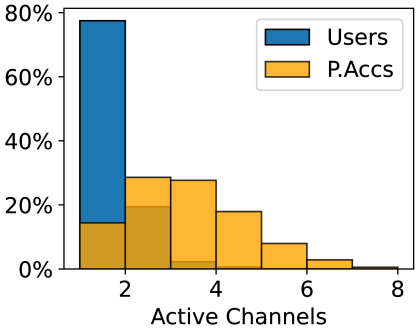


Figure 7: **Number of active channels for propaganda and user accounts.** Most users stick to one channel, while propaganda accounts are active in multiple channels simultaneously.

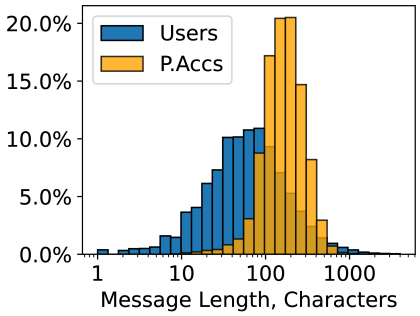


Figure 8: **Message length distributions for propaganda and user accounts.** Propaganda messages range from 10 - 1k characters, while users may send a single emoji or write a long post.

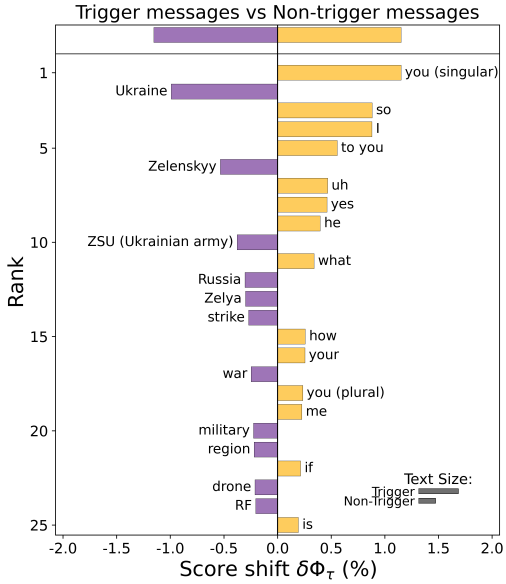


Figure 9: **Word graph for trigger and non-trigger messages.** All the stems are translated by the authors. Zelya("Зеля") – diminutive form for Zelenskyy, RF("рф") – Russian Federation. On the left side of this graph are the stems that are more prevalent for the trigger messages, while on the right side are the words typical for non-trigger messages.

more than 80% of propaganda messages removed (RT, Nexta, and Shtefanov). Comparing these ratios with the total moderation rate, i.e., the total number of deleted messages divided by the total number of messages in a given channel, we see that in some channels, the moderation of propaganda messages is much more aggressive than other messages. By contacting the moderation team of the Shtefanov channel (87% of propaganda messages deleted), we confirmed that their high success rate is due to their moderation policy's focus on the detection and deletion of propaganda messages alongside a substantial human effort into checking every message.

Table 2: **Propaganda moderation effectiveness** measured as the ratio of deleted propaganda messages to the total number of propaganda messages in a channel. Higher percentages indicate more aggressive moderation. We also report the ratio of all deleted messages to the total number of messages.

Channel	Size	Propaganda Moderation	Total Moderation
RT	584K	94.7%	9.3%
Shtefanov	281K	87.5%	15.2%
Nexta	824K	84.1%	19.6%
SamaraNews	5.0K	64.1%	56.8%
KK	158K	45.2%	13.3%
Readovka	863K	38.5%	16.5%
Topor	297K	29.6%	26.2%
Rudoi	47.3K	19.9%	7.9%
SpecchatZ	359K	-	18.5%

We conclude that some Telegram channels have a strong interest in propaganda moderation, and they do so mostly using manual detection and deletion. Manual detection has several drawbacks. First, it requires a dedicated staff (either hired or volunteered). Second, moderators are not always online, and their reaction time is limited by their concentration and reading abilities. Third, a non-negligible portion of propaganda messages (5-15%) remain visible to the users and attract interactions from them. Finally, human moderators are exposed to propaganda content, which may result in psychological issues, similar to hate speech or violent content moderation [39].

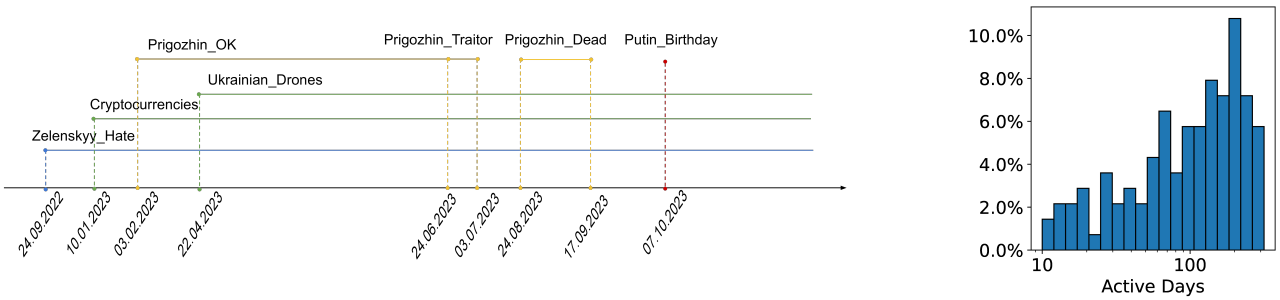


Figure 10: **Topics temporality**. *Left: Timeline of Selected Topics* Note that the Prigozhin rebellion [49] started on the 23rd of June, but the corresponding messages appeared only the next day, on the 24th. *Right: Topic Longevity* - Total activity time for different topics in days. Almost half of the topics are ephemeral (their lifespan is less than 100 days).

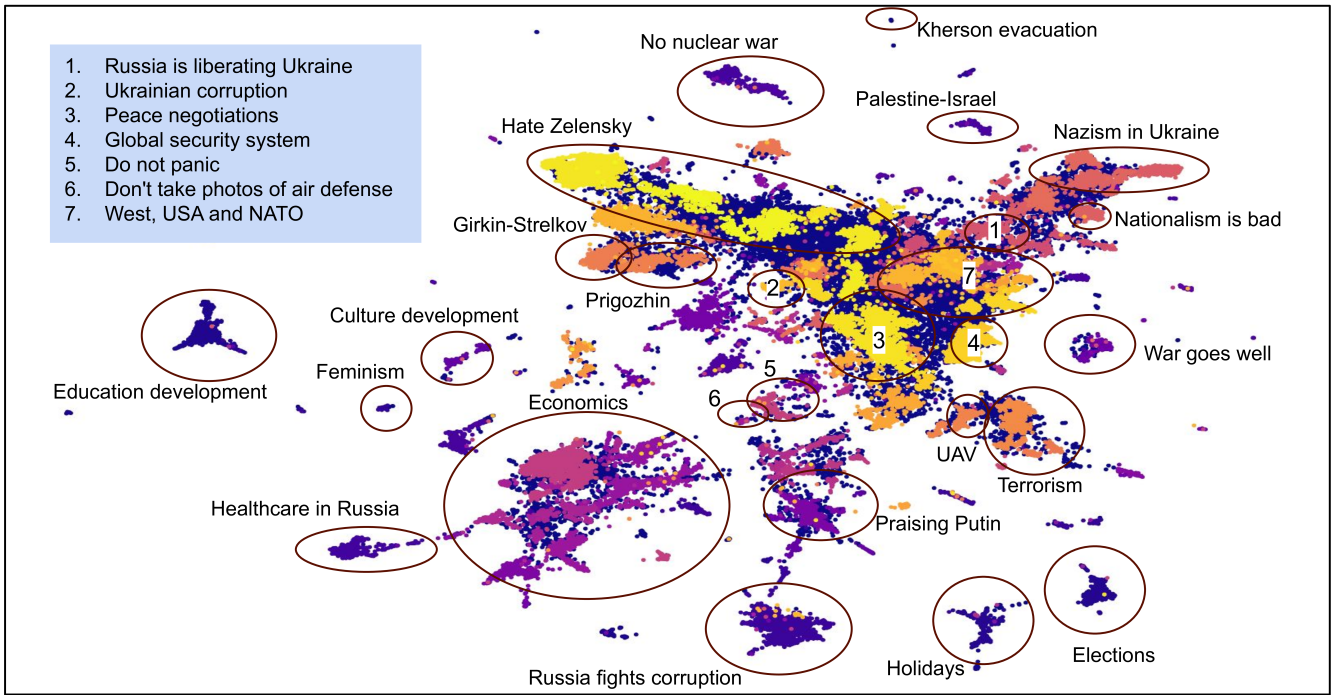


Figure 11: **Cluster map for the propaganda messages.** Cluster map is generated using UMAP. Different colors represent different clusters. Some clusters and groups of clusters are annotated in order to illustrate main topics and narratives. The central area consists of multiple clusters, mostly about the Russian-Ukrainian war, these small clusters are denoted with numbers. The largest fraction of the corpus is made up of messages dedicated to the War in Ukraine. The second largest group of clusters are topics related to the internal policies of the Russian Federation.

4.2 Automated Propaganda Detector

Automating propaganda detection would remove the need for an online dedicated staff, as it provides a better detection rate more quickly and cheaply than human moderators. To remain more effective than human moderators over time, it should also be at least as robust to changes in propaganda behavior, e.g., associated with new topics.

In line with work on other platforms (e.g., X [28] and Reddit [35]), we consider detection solutions in the form of a machine-learning classification model, which can be deployed by channel owners via the bot API on Telegram.

The classification model must only use information available via the API, i.e., the information moderators have access to. This consists of account information (First Name, Last Name, Username), message metadata (date and time, size), and message content (message text, trigger message text). As such, current SOTA approaches based on relations between different accounts (e.g., based on connections in the social graphs [1,16,29]) are also not possible. We do not use account information because it can be easily hidden or manipulated.

Table 3: **Detection Performance.** We report the average effectiveness over channels that have aggressive propaganda deletion policies, i.e., *Nexta*, *RT*, and *Shtefanov*. For comparison, we include the effectiveness of human moderators (see Table 2), where we report the precision due to the lack of false positive data for the accuracy estimation.

Method	Overall Accuracy	New Topics Accuracy	Validation
Human Moderators	86.0%	81.2%	-
Handcrafted features	83.8%	88.6%	80.6%
Trigger embeddings	79.0%	41.3%	54.0%
Propaganda embeddings	96.8%	89.4%	81.2%
Trigger-Propaganda ensemble	96.5%	77.5%	73.4%
Trigger-Propaganda embeddings	**97.4%**	**93.0%**	**88.8%**

We implement the following approaches:

Handcrafted features. We use handcrafted features computed on messages' metadata and content, like those widely used in the bot [17, 20], trolls [35], and spam [12, 52] detection literature. Concretely, we select the following textual features from [45]: *message length*, *number of words*, *number of URL links*, *number of emojis*, *number of exclamation marks*, *number of question marks*, *message time in seconds*, *latency between message and reply in seconds*. We use these features to train XGBoost [11], RandomForest [22], LightGBM [55], Decision Tree, Logistic Regression, and DNN classifiers. We only report results for the XGBoost model, since it achieved the best performance on our tasks.

Propaganda embeddings. The content of propaganda messages is different from user messages in terms of specific word frequency and style (see Sect. 3.3). Content-based detection in the literature is based on the use of n-grams [30], or textual embeddings [19, 25, 46]. In this work, we use the latter since textual embeddings are an n-grams generalization. Concretely, we use the same pre-trained SBERT embeddings that we use in Sect. 3.3 for clustering as they show good performance on other Russian language NLP tasks. We try several classifiers, including XGBoost, trained on these embeddings, and report results on the best performing one: a simple 3-layer DNN.

Trigger embeddings. Trigger messages are distinct from user messages in terms of their word frequencies (Sect. 3.3). An additional advantage over propaganda messages is that they cannot be manipulated by propaganda creators. We use the same embeddings and classifiers as for propaganda messages.

Trigger-Propaganda ensemble. We combine the results of the classifiers trained on trigger embeddings and propaganda embeddings in an ensemble. We take as output the rounded sum of the output of these two detectors.

Trigger-Propaganda embeddings. Using an ensemble does not capture any relationship between trigger messages and propaganda messages. Yet, we know that these pairs often appear distinct from normal conversations (see Conversation 2). To capture this mismatch, we assume that the textual information is partially preserved in the embedding, and feed the concatenation of embedding pairs of triggers and the corresponding propaganda replies to the same DNN-based classifier.

4.3 Evaluation

We evaluate all approaches with respect to the requirements in Section 4.1, and report the results in Table 3.

Automated detection performance. We first evaluate whether automated detection can obtain better performance than human moderators. We evaluate the automated detection approaches by training the classifiers on messages collected between August 16 and September 18, 2023 and testing their performance on messages collected between September 18 and October 16. This separation splits the data to train and test evenly and mimics a realistic scenario in which moderators deploying the detector can only train on labeled data from the past. To ensure that the evaluation is fair in terms of accuracy, we create a balanced dataset in terms of propaganda messages and users' messages.

The results of this evaluation (2nd column in Table 3) show that using trigger-propaganda embeddings, trigger-propaganda ensemble, and propaganda embeddings as input to the classifier outperforms human labelers. Among these, trigger-propaganda embeddings performs the best, closely followed by only using the content of propaganda messages. Notably, while the trigger-propaganda ensemble and trigger-propaganda embeddings use the same input data, there is a large difference in terms of performance. We interpret that this is because explicitly capturing the relation between trigger messages and their replies is important for detection. Also, trigger-propaganda ensemble does not improve over just using the propaganda messages, indicating that the trigger itself carries little information for detection.

Performance on unseen topics. Propaganda messages may refer to events or facts that are not present in the training period. In this section, we study how the different automated

Table 4: **Worst topic accuracy.** Red numbers indicate that the topic is in the top 5 worst topics per detector. Bolding indicates the best detector for each topic. We also report human moderators' performance in moderated channels. For example, the 2nd worst topic for the trigger-propaganda embeddings was Terrorism and the handcrafted features performed the best on that topic. For human moderators, we use "-" when topics have less than 50 messages in their channels.

Topic	Trigger-Propaganda emb.	Propaganda emb.	Handcrafted features	Human Moderators
Roads Developing	60.0%	60.0%	**74.0%**	-
Terrorism	68.1%	66.7%	**73.9%**	-
Alcoholism	**77.5%**	64.1%	23.3%	-
Holidays	**77.7%**	40.7%	63.0%	-
Education Developing	79.7%	**80.5%**	63.2%	72.5%
Sadness Emotion	**89.9%**	69.5%	1.2%	81.0%
Cryptocurrencies	91.1%	**97.0%**	11.8%	-
Sad News Emotion	**97.2%**	81.7%	23.4%	82.4%
Despair Emotion	**96.5%**	91.2%	45.6%	82.7%
Ukrainian Refugees	**100.0%**	**100.0%**	98.9%	77.7%
Palestine-Israel	**88.8%**	81.5%	81.9%	77.7%
Russia Helps	**99.3%**	97.8%	**98.5%**	77.8%
Culture Developing	99.3%	**99.7%**	75.3%	79.4%

approaches perform in such circumstances. We observed five new topics on the test set:

1. *Road Development*: The growth of Russian's road network.
2. *Alcoholism*: The decrease in alcohol consumption in Russia, thanks to the introduction of new laws.
3. *Putin Birthday*: President birthday wishes (Oct. 7).
4. *Armenia-Azerbaijan*: The Nagorno-Karabakh war and Armenia-Azerbaijan relations (from Oct. 20 after the escalation that started on Oct. 19).
5. *Palestine-Israel*: The Israel-Hamas war (from Oct. 9 after the events on Oct. 7).

The 3rd column in Table 3 shows the average accuracy across new topics. Trigger-Propaganda embeddings provide the best capability to adapt to new topics (93.0%), followed by using propaganda embeddings (89.4%). Trigger messages yield very poor results, likely due to overfitting. Whether alone or in the ensemble, trigger messages reduce the detector performance considerably since they vary greatly in language and length, unlike propaganda messages which follow certain style patterns that can be captured by embeddings.

The performance of all automated approaches decreases on new topics, and so does the performance of human moderators (5% decrease). We conjecture that human moderators also need to "learn" these new topics, to efficiently delete the propaganda messages associated with them.

Error analysis. We now study whether detection performance degradation is due to the appearance of new topics or if certain topics are inherently more difficult to classify than others.

We plot the distribution of accuracy over topics of handcrafted features, propaganda embeddings, and trigger-propaganda embeddings in Figure 12. Using trigger-propaganda embeddings demonstrates the most consistent results across topics. Using handcrafted features results in poor generalization, with some topics being particularly difficult to identify even if they exist in the training set.

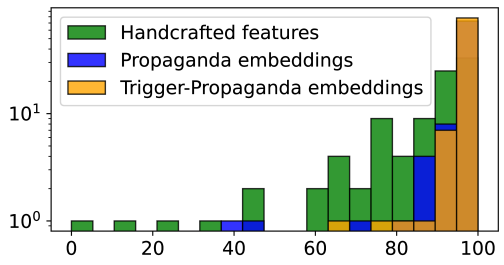


Figure 12: **Accuracy distribution for different topics.** Trigger-Propaganda embeddings demonstrate the most consistent performance across topics. Handcrafted features offer very bad performance on some topics.

We explore this issue in Table 4, which shows each detector's performance on its worst five topics. Handcrafted features perform poorly on emotional topics, cryptocurrency-related messages, and alcoholism because these topics tend to have short messages (e.g., *"I do not understand what crypto is."*), and message length is one of the handcrafted features. The *Holidays* topic is difficult for all approaches since holiday-related messages, such as *"Happy New Year!"* also appear in the user messages. Still, trigger-propaganda embeddings yield

good results when the propaganda message is inconsistent with the trigger message (see Conversation 3).

Finally, we study whether human moderators make the same errors as automated detectors. Figure 13 shows that most of the trigger-propaganda embeddings errors are also made by handcrafted features, making it almost strictly superior, while the ML-based methods share few errors with humans. Table 4 shows that most topics problematic for humans are not for ML-based detectors. We found no clear reason why some topics are easier for humans. An interesting case is the *Putin_Birthday* topic where human moderators have a precision of 94.2% while the best ML-based detector (trigger-propaganda embeddings) only achieves 91.5%.

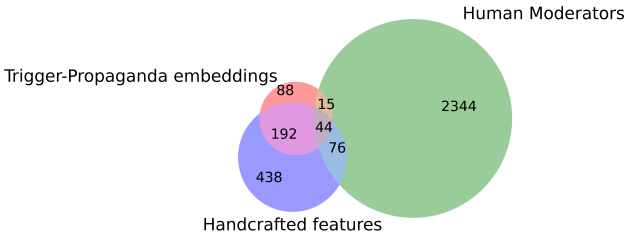


Figure 13: **False negatives in ML-based detectors and human moderators.** A Venn Diagram of human errors, handcrafted features and Trigger-Propaganda embeddings.

Validation on Pro-Ukrainian propaganda network. Though automated detection generalizes to unseen topics, it is not clear if the detection methods can retain their performance if propaganda accounts change their behavior more radically. Here, we evaluate the performance of the detectors on a second network that we discovered during our evaluation.

Pro-Ukrainian propaganda network. During error analysis, we found some false positives had clear propaganda purposes but distinct content and account behavior. These messages contain pro-Ukrainian propaganda targeted at a Russian-speaking audience. We call this network *pro-Ukrainian*, as

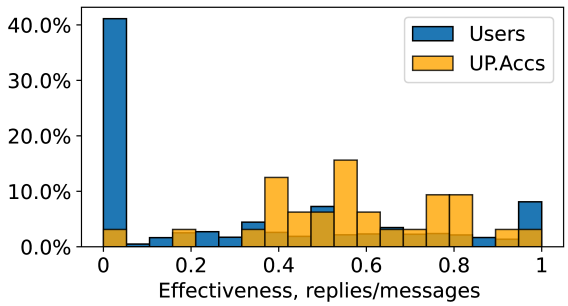


Figure 14: **Effectiveness of pro-Ukrainian propaganda messages** The pro-Ukrainian messages tend to attract more replies than messages written by user accounts.

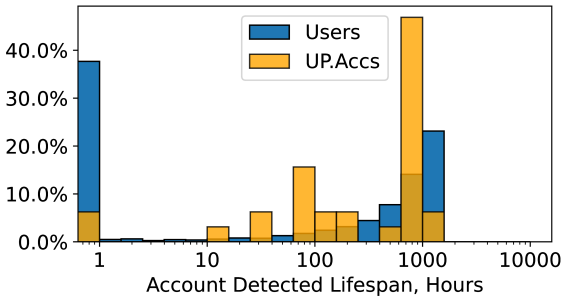


Figure 15: **Lifespan of the pro-Ukrainian propaganda account accounts.** The lifespan pattern differs significantly from the pro-Russian account lifespans (Fig. 4): they are active longer than users, despite lower presence in total.

opposed to the *pro-Russian* network presented in the previous sections. The pro-Ukrainian accounts repeat messages from each other but not from accounts in the pro-Russian network, indicating a unique, second network. We did not find these channels at the beginning of the study because these accounts were not active in the channels we use for hand-labeling.

The examples of messages used by the pro-Ukrainian network include:

EXAMPLE: *"Under Putin's leadership, Russia has witnessed systematic violations of human rights, restrictions on freedom of speech, and the suppression of opposition."*

EXAMPLE: *"Ukraine has all the signs of a sovereign state: its own constitution, economy, army, and it represents the interests of its people in the international arena."*

EXAMPLE: *"The concentration of power in just one man's hand can slow down the decision-making process and lead to an insufficient response to challenges and changes in society and the world."*

Pro-Ukrainian propaganda accounts use fictional nicknames for first and last names (e.g., "Atlanta" or "Az Air."), instead of common names, and hide their usernames. Their topics and style also differ from the pro-Russian accounts. The activity of these accounts is sporadic: they are active for short periods (1-2 days), disappear (20-30 days), and then reappear, demonstrating a relatively long total lifespan (Fig. 15). Unlike the pro-Russian accounts, we also observe that they react (e.g., like) each other's comments. Despite these differences, pro-Ukrainian accounts, as pro-Russian ones, often post replies unconnected with the trigger messages, demonstrating the same lack of 'bridge words' (Fig. 16). It is hard to compare the effectiveness of these networks since they operate mostly in different channels and time periods. However, pro-Ukrainian accounts surpass user accounts in the average number of replies (Fig. 14, 0.65 vs 0.43), indicating that this network may also draw significant user attention. We report additional information about this network in Appendix E.

Evaluation. To evaluate automated detection on the pro-Ukrainian network, we repeat the labeling and augmentation

User1: One may ask, What does Putin's birthday have to do with it?

↳ **"daniil" (ahanthuda)**: Vladimir Vladimirovich, Happy Birthday! May everything go well for you!

User2: Happy 71st birthday to Putin! He was born in Leningrad in 1952 on October 7, but despite his age, V. Putin is still as handsome as ever!

↳ **"Mark" (xiverelaroya)**: Our leader is strong! I wish you a happy birthday, Vladimir Vladimirovich!

Conversation 3: **Examples of errors for propaganda embeddings detector (left) and for both propaganda embeddings and Trigger-Propaganda embeddings detectors (right)** Left: A propaganda account does not catch the irony and provides an unconnected reply. the Trigger-Reply system spotted the mismatch between the user message and reply, while the detection system using only the message information failed. Right: The conversation is completely normal, the reply matches the message, and even human labelers cannot detect a propaganda account based on this conversation.

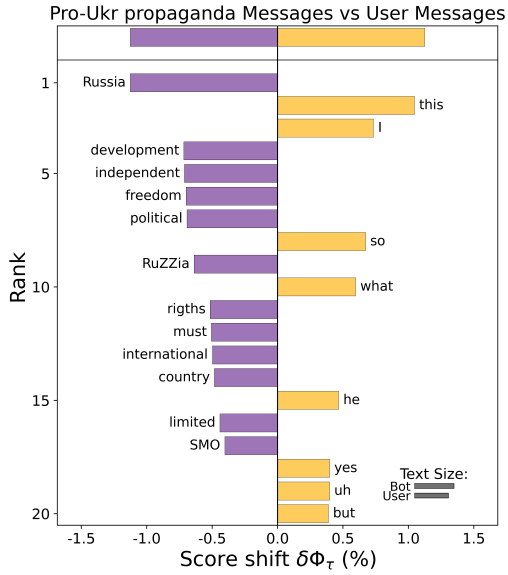


Figure 16: **Word graph for pro-Ukrainian propaganda messages and user messages.** All the stems are translated by the authors. SMO ("СВО") – Special Military Operation, official title in Russia for the Russian-Ukrainian war.

from Sect. 2.3 and obtain 2.7K propaganda messages, which we balance with an equal number of user messages.

We observe a performance degradation for all approaches (Table 3, 4th column). The embeddings-based methods show the most significant drop (11-15%). Handcrafted features drop just 6%, since message length, its main heuristic, remains useful for long pro-Ukrainian messages. Trigger-Propaganda embeddings still perform the best (88.8%), likely due to its capacity to capture the relationship between triggers and replies.

Robustness to Evasion. In addition to robustness against topic and domain shifts, robustness against intentional evasion is essential. Here, we show that detector evasion is non-trivial without *manually* altering behavior. Even if the adversary uses large language models to alter the message style or stops using trigger messages (losing effectiveness), the performance of the trigger-propaganda embeddings remains largely unaffected, indicating cheap automation is unlikely to succeed.

We evaluate the robustness of the designed method to deliberate changes against two attacks. Due to computational constraints, these attacks were applied to a random sub-sample of 1,000 messages from the test data. The results of these experiments are reported in Table 5.

Table 5: **Robustness to the different attack vectors.** We report the clean accuracy for our sample, accuracy after dropping the trigger, and accuracy after the style change and shortening attack (Rephrase), done by GPT-4 model.

Method	Clean	No Trigger	Rephrase
Trigger-Propaganda embeddings	98.2%	94.0%	95.7%
Propaganda embeddings	97.2%	97.2%	92.8%

Sending messages without any trigger. Almost all propaganda messages in our dataset are replies to messages. This may be because in order to send a message that is not a reply to a channel, a user must *join* the channel, which sends a *join chat event* to the moderators and adds them to the chat member list (i.e., makes them more visible and raises suspicion). Messages that do not have a "reply-to" field are also less visible because the channel subscribers who are not members of the attached group will not see it (For example, out of 2.8M Readovka subscribers, only 10K are members of the attached group). That may explain why only a very small portion (0.6%) of propaganda messages were not replying to other messages. Due to the reasons stated above, we consider this attack vector severely damaging to the attacker's utility. However, we evaluated this attack on the test messages sample by artificially dropping their reply-to field and passing them to the detector. The performance of Trigger-Propaganda embeddings detector remains high at 94%.

Changing style of the propaganda messages. Another eva-

sion technique is changing the style and length of the messages while keeping the content the same. We simulate this attack by asking the GPT-4 model to shorten the text and change its style. This method can also lead to the degradation of utility: shorter messages may have less of an impact on users, though a user study would be needed to confirm this. We performed this attack with fewer changes in style and without affecting the message length, but the detector's accuracy did not decline. This attack is also largely inefficient. Moreover, the higher drop in performance for Propaganda embeddings detector (4.4% vs 2.5%), supports our assumption that including the Trigger-Propaganda relationships to the detector method improves its robustness to evasion.

4.4 Deployment Considerations

We assess the financial and computational requirements associated with using trigger-propaganda embeddings, the best-performing detector. Since this detector includes two neural networks, it can be executed purely on CPU or with GPU acceleration. Renting a dedicated server with a CPU is cheaper, which may be important for small, unmonetized channels.

We measure the average time for processing trigger-reply pairs one by one on the test set, which gives a worst-case timing estimation with respect to using badges. Using an NVIDIA RTX 3070 GPU, the average computation time is 0.015 ± 0.001 sec, while for an AMD Rysen 4700G CPU, the computation time increases to 0.25 ± 0.01 sec. We do not have the technical means to measure the reaction time of human moderators (the Telegram API for deletion events is considered unreliable [40]). However, the visual reaction time for a human is more than 0.2 seconds [4], without accounting for time to read, process the content, make a decision, and click all the buttons in the app (and the fact that humans cannot always be online). During our manual labeling, we could not label a message faster than in 1-3 seconds. We conclude that even using a slow CPU-based detector would result in a reaction time gain over human moderation.

Renting a GPU can even be profitable: a GPU node on Amazon AWS costs 0.21$/h [2], while the federal minimum wage in Russia is equal to 1.2$/h, and the average salary is $\sim$4$/h [37]. In reality, a dedicated GPU for detection is unlikely to be fully loaded (due to the low frequency of incoming messages), and the GPU price can be further optimized using services like inference-on-demand [3].

5 Conclusion

Telegram and other instant messengers are main sources of information in critical situations, in particular outside of the Western world. Our work evidences that due to its instant-messaging nature, which is structurally different than the typically studied platforms in the literature, Telegram requires new collection and analysis methods. As such, we leveraged textual embeddings to capture the behavior of the propaganda accounts we found to obtain a quick and effective detector that outperforms human moderators by 11.6% and is robust to topic changes. Our Telegram-tailored collection and analysis allowed us to discover two large coordinated networks spreading propaganda and misinformation around the Russo-Ukrainian war and other politically-charged topics.

While future work should test to what extent our detection method generalizes to other propaganda campaigns on Telegram, this paper already shows that it is possible to help mitigate the threat of information-based attacks in instant messaging-based social networks. We hope that our results inspire the security community to broaden its attention beyond Western-centered social networks and build more tools to reduce information-based attacks worldwide.

Acknowledgments

We are thankful to the moderation team of Shtefanov telegram channel for the fruitful discussion on the moderation process and propaganda activity behaviour.

Ethics considerations and compliance with the open science policy

Ethics considerations. Analyzing large-scale Telegram data may raise ethical concerns. To mitigate possible harm, we only use publicly available data via the official Telegram API, following Telegram's terms of service and the Telegram Privacy Policy. We follow secure guidelines for data storage and processing using our institutional infrastructure. After publishing our dataset, we commit to deleting the data of individual Telegram users if they contact us with a corresponding request. The researchers do not use their personal telegram accounts in this project, and the accounts created to execute the collection never interacted with the channels. The authors have carefully designed the data collection and presentation to ensure protection from any state entities that could be involved in the studied propaganda behavior.

We do not see negative repercussions from publishing this paper, on the contrary, we are convinced that publishing this work will help to inform the population about the presence of the propaganda behavior, and possibly help to efficiently mitigate this threat.

This project and the corresponding release of the dataset used in it have been approved by the IRB of our institution.

Code availability. In compliance with the open science policy, we will release the code samples for the training and inference of our models and reproducing the plots present in the paper[4].

Data availability. Our dataset is published on Zenodo[4].

References

[1] Seyed Ali Alhosseini, Raad Bin Tareaf, Pejman Najafi, and Christoph Meinel. Detect me if you can: Spam bot detection using inductive representation learning. In *Companion proceedings of the 2019 world wide web conference*, pages 148–153, 2019.

[2] Amazon. Amazon ec2 g4 instances, 2024.

[3] Amazon. Amazon inference on demand, 2024.

[4] Rasoul Amini Vishteh, Ali Mirzajani, Ebrahim Jafarzadehpour, and Samireh Darvishpour. Evaluation of simple visual reaction time of different colored light stimuli in visually normal students. *Clinical Optometry*, pages 167–171, 2019.

[5] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. The pushshift telegram dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 840–847, 2020.

[6] Vera Bergenguen. How telegram became the digital battlefield in the russia-ukraine war. *Time Magazine*, 2022. Available at: https://time.com/6158437/telegram-russia-ukraine-information-war/ (Accessed: May 29th, 2024).

[7] David M Beskow and Kathleen M Carley. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, 2018.

[8] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P1000, 2008.

[9] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):7, 2019.

[10] Samantha Bradshaw and Philip N Howard. Challenging truth and trust: A global inventory of organized social media manipulation. *The computational propaganda project*, 1:1–26, 2018.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[12] Zi Chu, Indra Widjaja, and Haining Wang. Detecting social spam campaigns on twitter. In *International Conference on Applied Cryptography and Network Security*, 2012.

[13] Joan Donovan. Misinformation is warfare. *Time Magazine*, 2023. Available at: https://time.com/6323387/misinformation-israel-hamas-war-essay/ (Accessed: May 30th, 2024).

[14] Hridoy Sankar Dutta, Vishal Raj Dutta, Aditya Adhikary, and Tanmoy Chakraborty. Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*, 2020.

[15] Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. Characterizing retweet bots: The case of black market accounts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 171–182, 2022.

[16] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269, 2022.

[17] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 2016.

[18] Florian Gallwitz and Michael Kreil. The rise and fall of 'social bot' research. *SSRN: https://ssrn.com/abstract=3814191*, 2021.

[19] Andres Garcia-Silva, Cristian Berrio, and José Manuel Gómez-Pérez. An empirical study on pre-trained embeddings and language models for bot detection. In Isabelle Augenstein, Spandana Gella, Sebastian Ruder, Katharina Kann, Burcu Can, Johannes Welbl, Alexis Conneau, Xiang Ren, and Marek Rei, editors, *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 148–155, Florence, Italy, August 2019. Association for Computational Linguistics.

[20] Zafar Gilani, Ekaterina Kochmar, and Jon Crowcroft. Classification of twitter accounts into automated agents and human users. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 489–496, 2017.

[21] Hans WA Hanley, Deepak Kumar, and Zakir Durumeric. Specious sites: Tracking the spread and sway of spurious news stories at scale. *45th IEEE Symposium on Security and Privacy*, 2024.

[22] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[23] Philip N Howard and Bence Kollanyi. Bots,# strongerin, and# brexit: Computational propaganda during the uk-eu referendum. *Available at SSRN 2798311*, 2016.

[24] Sofia Hurtado, Poushali Ray, and Radu Marculescu. Bot detection in reddit political discussion. In *Proceedings of the fourth international workshop on social sensing*, 2019.

[25] Shubham Kumar, Shivang Garg, Yatharth Vats, and Anil Singh Parihar. Content based bot detection using bot language model and bert embeddings. In *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)*, pages 285–289. IEEE, 2021.

[26] Paul Larsen. Rosebot, 2023.

[27] Luca Luceri, Ashok Deb, Adam Badawy, and Emilio Ferrara. Red bots do it better: Comparative analysis of social bot partisan behavior. In *Companion proceedings of the 2019 world wide web conference*, pages 1007–1012, 2019.

[28] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019.

[29] Samaneh Hosseini Moghaddam and Maghsoud Abbaspour. Friendship preference: Scalable and robust category of features for social bot detection. *IEEE Transactions on Dependable and Secure Computing*, 20(2):1516–1528, 2022.

[30] Juan Pizarro. Using n-grams to detect bots on twitter. In *CLEF (Working Notes)*, 2019.

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[32] RE:RUSSIA. In russia, telegram has become the primary internet platform for young people, surpassing youtube in reach and whatsapp in average daily user time, 2023.

[33] José-Manuel Robles, Juan-Antonio Guevara, Belén Casas-Mas, and Daniel Gómez. When negativity is the fuel. bots and political polarization in the covid-19 debate. *Comunicar*, 30(71):63–75, 2022.

[34] Richard Rogers. Deplatforming: Following extreme internet celebrities to telegram and alternative social media. *European Journal of Communication*, 35(3):213–229, 2020.

[35] Mohammad Hammas Saeed, Shiza Ali, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Trollmagnifier: Detecting state-sponsored troll accounts on reddit. In *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022.

[36] SberDevices. Bert large model (uncased) for sentence embeddings in russian language., 2022.

[37] Federal State Statistics Service. Accrued average nominal wage and median wage grew by 14.12023.

[38] Veronika Solopova, Oana-Iuliana Popescu, Christoph Benzmüller, and Tim Landgraf. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank-Spektrum*, 23(1):5–14, 2023.

[39] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14, 2021.

[40] Telethon. Telethon documentation, 2024.

[41] USAID. Ukrainians increasingly rely on telegram channels for news and information during wartime, 2023.

[42] Natalia Vanetik, Marina Litvak, Egor Reviakin, and Margarita Tiamanova. Propaganda detection in Russian telegram posts in the scope of the Russian invasion of Ukraine. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1162–1170, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[43] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, 2017.

[44] Vox-Harbour, 2024. https://telegra.ph/Kremleboty-v-Telegrame-Bolshoe-issledovanie-ot-Vox-Harbor-01-26.

[45] Bo Wang, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Making the most of tweet-inherent features for social spam detection on twitter. In *Proceedings of the the 5th Workshop on Making Sense of Microposts co-located with the 24th International World Wide Web Conference (WWW 2015), Florence, Italy, May 18th, 2015*, CEUR Workshop Proceedings, 2015.

[46] Feng Wei and Uyen Trang Nguyen. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In *2019 First IEEE International conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pages 101–109. IEEE, 2019.

[47] Mariëlle Wijermars and Tetyana Lokot. Is telegram a "harbinger of freedom"? the performance, practices, and perception of platforms as political actors in authoritarian states. *Post-Soviet Affairs*, 38(1-2):125–145, 2022.

[48] Wikipedia. Russian foreign agent law, 2024.

[49] Wikipedia. Wagner group rebellion, 2024.

[50] Chao Yang, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, 2013.

[51] Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[52] Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. Detecting spam in a twitter network. *First Monday*, 2010.

[53] Ahmet S Yayla and Anne Speckhard. Telegram: The mighty application that isis loves. *International Center for the Study of Violent Extremism*, 9, 2017.

[54] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *Companion proceedings of the 2019 world wide web conference*, 2019.

[55] Jin Zhang, Daniel Mucs, Ulf Norinder, and Fredrik Svensson. Lightgbm: An effective and scalable algorithm for prediction of chemical toxicity–application to the tox21 and mutagenicity data sets. *Journal of chemical information and modeling*, 59(10):4150–4158, 2019.

A Real-Time Collection mechanism

In order to collect data in real-time mode, we developed a simple custom Telegram client application using the telethon library [40], which we officially registered as such on the Telegram website and received the required API key. The client application is deployed on the full-time running and protected server and operates in the following way: after an account logs in to the client, for all groups that this account has joined, the client receives events from the Telegram server

when a new message appears in a group. The client checks if the source of the message reported in the event is in our list of observed groups, and if this is the case, the application saves the messages as a JSON string, which we later refine in order to maintain the same format as we have in the historical data. To log into this client, we use a Telegram account belonging to one of the authors. This method allows us to download group messages, channel posts, and comments, since every channel has the attached "Discussion" group, where all comments and posts are shown as messages.

B Conversation samples sources in Russian and Ukrainian

Original version of the Conversation 1:

User: "Україна була і буде завжди свободна і незалежна ні від кого тільки рашисти уроди хотіли захватити за 3 дні а получили хуй в сраку ,вам пизда уже убивці Україна Переможе ! Слава Нашим Бійцям ! Слава Україні!"

↳ **"Michelle Ortega" (venonisa)**: "Не один освобожденный житель украинских городов, где ведутся боевые действия, уже убедился в том, что Россия не пытается захватить Украину, а лишь освобождает её от нацистского давления, оказывающего невероятную опасность для людей России и Украины."

Original version of the Conversation 2 (Left):

User1: "Отдельный угар в том, что вербуют в зону сво прямо на некоторых военных предприятиях. То есть они готовы бросить в пламя войны даже самых необходимых в моменте специалистов"

↳ "Лира Капустина" **(unknown username)**: "Нацбаты и ЧВК не имеют отношения к официальной армии. Их мало интересуют приказы официальных властей. В прямом смысле слова неуправляемые берсерки, вооруженные до зубов. Само по себе существование вот таких вот боевых отрядов в Украине одна из причин денацификации."

Original version of the Conversation 2 (Right):

User2: "Бля а мужчины выскочили как самые главные знатоки феменизма
Оставьте феменизм для женщин суки"

↳ **"Gesha" (ronashisi)**: "Радфем - серьезная болезнь. Мое мнение не изменить."

Original version of the Conversation 3 (Left):

User1: "Казалось бы, при чем тут день рождения Путина "

↳ **"daniil" (ahanthuda)**: "Владимир Владимирович, с Днем Рождения, пусть у Вас всё будет просто хорошо!"

Original version of the Conversation 3 (Right):

User2: "Поздравляем Путина с 71-м Днём рождения. Родился он в Ленинграде в 1952 году 7 октября, не смотря свой возраст В.Путин всё также красив как и всегда."

↳ **"Mark" (xiverelaroya)**: "Насколько сильный у нас лидер! От всей души поздравляю вас, Владимир Владимирович!"

C GPT-4 prompts used in the paper

Prompt used for the random-username experiment:

```
After a string @@@, I will give you a
username. Tell me please if this username
contains a clear reference to something in
Russian or English language. The reference
can be to first or last names, events, movies,
literature, history, nature, pop-culture, etc.
If there is no reference, just answer one
word 'No', otherwise say 'Yes' and explain
the reference. Note that users can replace some
letters in usernames by digits, e.g. 'i' can be
replaced with '1' or 'o' can be replaced with
'0'. @@@
```

Prompt used for the western-username experiment:

```
After a string @@@, I will give you a
username. Tell me please if this username is
a combination of the first name and the last
name common for the United States or United
Kingdom. Note, that the last name can have some
additional numbers or characters at the end
like in "Smith5" or "Smithk". If it is explain
why, if it is not just output one word "No.".
@@@
```

Prompt used for the rephrase attack experiment:

```
After the string @@@ I will give you a piece
of text in Russian language, please rephrase
the given text preserving the meaning, but
significantly changing its style. Make it
shorter and more informal, closer to what a
normal person would write. Keep the output in
```

```
the same language as input, and do not include
@@@ in the response. @@@
```

D Selected Topics

In this Appendix, we list all the topics mentioned in Section 4 with brief descriptions and examples

Roads Developing. This topic contains messages explaining that the road system in Russia is constantly improved by the government.

EXAMPLE: *Now it's very easy to solve the problem of dangerous sections of roads, pits and holes – you just have to go through the State Service App (Gosuslugi) – it's gonna be quick!*

Terrorism. Messages explaining that the Russian government fights terrorism.

EXAMPLE: *It's great that in Russia, day and night, the government fights terrorists and other threats, providing security for the citizens of the country!*

Alcoholism. Messages in this topic explain that the situation with alcohol consumption is improving in Russia, and the government has introduced working policies.

EXAMPLE: *Yeah, there are a lot of rehabilitation centres now, so there are fewer drunkies since they're going straight to treatment.*

Holidays. Messages tied to certain national Holidays, such as New Year, Constitution Day, Mother's Day, etc. Example:

EXAMPLE: *I want to wish all of you a new year of fulfilling all your wishes, and all your dreams come true!*

Education Development. This topic contains texts explaining that the Russian educational system is good and is improving every year. Example:

EXAMPLE: *I'm so happy that Russian education is now developing very dynamically. My sister is studying at Moscow State University school - she really likes it*

Culture Developing. Similar to the Education Development, but about culture.

EXAMPLE: *No one in Russia would neglect cultural development! We have so many talented people who fantasize about amazing ideas, and the state is helping to make this happen!*

Sad News Emotion. These messages contain emotional responses to user messages containing information about crimes, disasters, etc.

EXAMPLE: *I wish there were less news like this*

EXAMPLE: *I'm shocked by this kind of news*

Sadness Emotion. Similar to the previous topic, but not tied to the news, just expressing sadness.

EXAMPLE: *That's so fucking gross.*

EXAMPLE: *Fuck. Is it possible not to see something like that again?*

Despair Emotion. Another emotional topic, more about fear.

EXAMPLE: *Fuck, that's awful.*

EXAMPLE: *It's scary, so scary.*

Putin Birthday. Messages wishing Putin a happy birthday.

EXAMPLE: *Vladimir Vladimirovich is really working hard for Russia, he's doing a lot for us. Happy birthday, our president!*

Cryptocurrencies. Messages expressing doubt about cryptocurrencies.

EXAMPLE: *I don't think the crypto is gonna be anything serious, it's just a toy.*

EXAMPLE: *Crypto here, crypto there, and it is a fucking soap bubble which is hyped all over the place.*

Income. Messages convincing people that the average income is not getting worse or that the government controls the process. Both personal examples and general statements.

EXAMPLE: *Well, don't make it up, even if we've got a little lower income, the authorities are already keeping that matter under control.*

EXAMPLE: *I don't know who's earning less now. Personally, I'm fine.*

Ukrainian Refugees. Messages explaining that Ukraine must stop the war if they want refugees to go back home.

EXAMPLE: *In general, I understand that the refugee situation could have been avoided easily. Zelenskyy, if he were worried about the people, would have made a truce with Russia at the beginning. Now he has to do the same thing right now, so that more people don't run away to other countries.*

Palestine-Israel. Messages regarding Israeli–Palestinian conflict. Interestingly, most of the messages were pushing towards immediate peace; also there are messages putting the blame for this conflict on the US.

EXAMPLE: *It seems to me that the only way to resolve this whole situation between Palestine and Israel is through peace talks, other methods are not working.*

EXAMPLE: *The US can help in a peaceful solution, but they always pick up a scenario that only triggers a war: it was in Ukraine, now we're seeing it in Israel!*

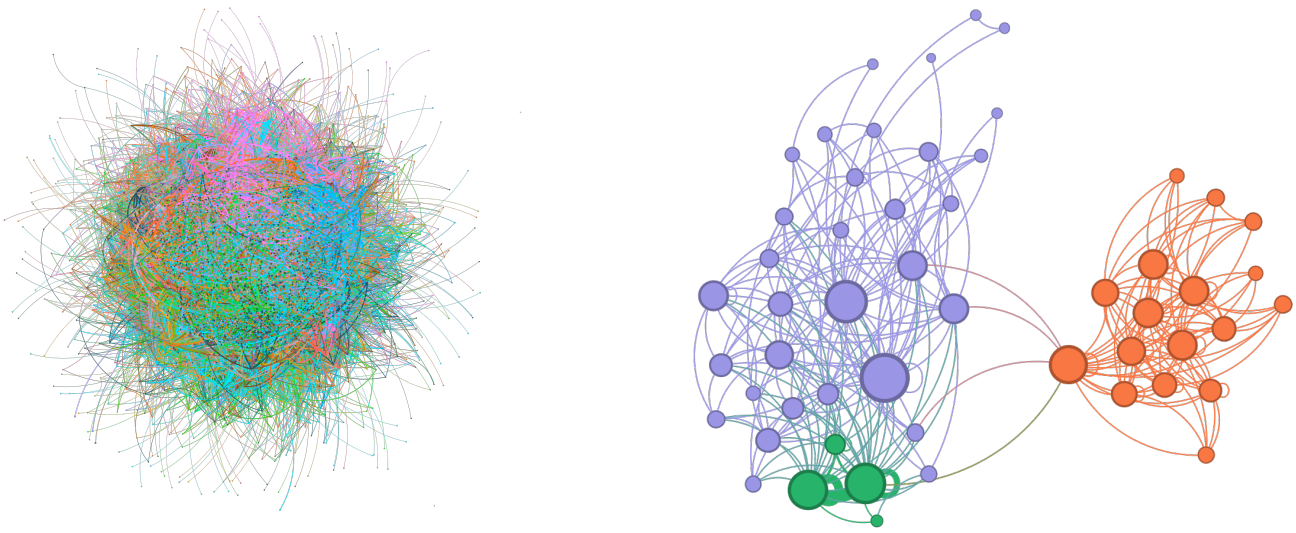
Russia Helps. Messages explaining that Russia helps common Ukrainians.

EXAMPLE: *We are not going to leave people in the liberated towns and settlements; we are willing to continue to support them until the situation improves, and there are a lot of videos on the internet directly from those delivering humanitarian aid.*

E Pro-Ukrainian propaganda account network

In this appendix, we extend our analysis of our data regarding the pro-Ukrainian network introduced in Section 4.3. We have labeled 2.7K messages from 53 different accounts, operating from May 25 to October 5, 2023.

On Figure 17b, we build the community graph using the same text-repetition method as in Section 3.1. Overall, it supports the hypothesis that these networks have different origins and behavior. While the pro-Russian network does not



(a) pro-Russian

(b) pro-Ukrainian

Figure 17: **Community structures for the pro-Ukrainian and pro-Russian propaganda account networks** The pro-Russian network is highly connected and does not demonstrate any well-separated communities. On contrary, pro-Ukrainian accounts form 3 distinctive communities, formed during sporadic activity periods. The green cluster appeared first on May 25-27th, 2023, followed by the purple one on July 1-7th, and the orange during short periods in August, September and October.

demonstrate distinctive communities, the pro-Ukrainian ones form three distinctive clusters, associated with short periods of their sporadic activity.

F Labeling completeness

In this appendix, we show how the size of the manually labeled subset affects the number of propaganda accounts spotted after the augmentation process. First, we took our manually labeled set of propaganda accounts and artificially reduced it by 50%, 75%, 90%, 95%. After that, we applied the snowball augmentation procedure until it converged.

Besides that, we investigated if selecting one particular channel affects the augmentation performance; we took another channel (Shtefanov) and manually labeled $\tilde{3}$0K messages in this channel. After performing the same augmentation procedure with additional data, we found no difference in propaganda accounts spotted. The results of these experiments are reported on the Figure 18. It shows that the manually labeled sample size does not affect the final size of the labeled dataset, and manually labeling more data does not provide any improvement in the data quality.

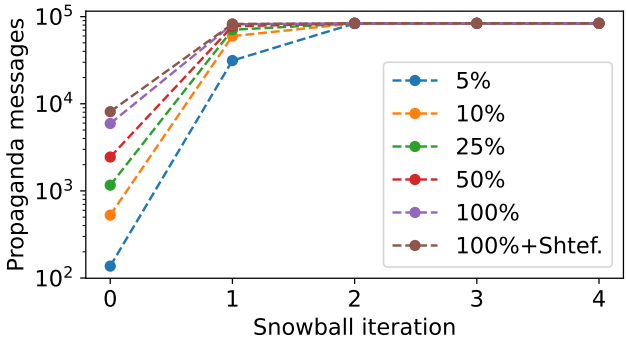


Figure 18: **The dataset size during the augmentation procedure.** Different lines correspond to different portions of the dataset used in the initialisation of the snowball augmentation procedure. *100% + Shtef.*

G Community structures for the propaganda accounts after the showball augmentation

Figure 17a shows the coordination graphs for all propaganda accounts after snowballing was applied. Recall that each node represents an account and the edges between nodes are weighted by the number of long (> 10 characters), identical messages shared between them. Nodes are colored by community [8]. Like Figure 3, the propaganda accounts graph is dense (Average Degree = 105.423) and has very few isolated nodes.