Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Partial Mobilization: Tracking Multilingual Information Flows amongst Russian Media Outlets and Telegram

Hans W. A. Hanley and Zakir Durumeric

Stanford University
hhanley@cs.stanford.edu, zakir@cs.stanford.edu

Abstract

In response to disinformation and propaganda from Russian online media following the invasion of Ukraine, Russian media outlets such as Russia Today and Sputnik News were banned throughout Europe. To maintain viewership, many of these Russian outlets began to heavily promote their content on messaging services like Telegram. In this work, we study how 16 Russian media outlets interacted with and utilized 732 Telegram channels throughout 2022. Leveraging the foundational model MPNet, DP-Means clustering, and Hawkes processes, we trace how narratives spread between news sites and Telegram channels. We show that news outlets not only propagate existing narratives through Telegram but that they source material from the messaging platform. For example, across the websites in our study, between 2.3% (ura.news) and 26.7% (ukraina.ru) of articles discussed content that originated/resulted from activity on Telegram. Finally, tracking the spread of individual topics, we measure the rate at which news outlets and Telegram channels disseminate content within the Russian media ecosystem, finding that websites like ura.news and Telegram channels such as @genshab are the most effective at disseminating their content.

1 Introduction

On February 24, 2022, the Russian Federation invaded Ukraine. In the buildup to and following the initial invasion, Russian state media conducted massive information campaigns justifying the Russian state's invasion as a "special military operation" to "liberate" the people of Ukraine (Wong 2022). In response, the EU and the UK among others banned or otherwise censored Russian news media. To circumvent this censorship, Russian outlets like RT and Sputnik News began to redirect users to and promote their content on the messaging app Telegram; ultimately causing Telegram to become one of the main platforms for Russian propaganda (Bergengruen 2022). When Telegram eventually succumbed to pressure to de-platform prominent Russian outlets (*e.g.*, rtnews), several Russian state media outlets created new Telegram channels (*e.g.*, swentr [rtnews backward])[1] and found other ways to circumvent the bans. However, while there has been extensive reporting on the Russian state media's usage of Telegram (Bergengruen 2022), there has been no systematic study of how information flows between Russian media and Telegram.

In this paper, we document the increased usage of Telegram by Russian outlets and present the first programmatic and multilingual study of the spread of news content amongst and between Russian news sites and Telegram. To do this, we crawl and gather content published between January 1 and September 25, 2022, from 16 Russia-based news sites (215K articles) and the 732 Telegram channels hyperlinked by these Russian news sites (2.48M Telegram messages). Leveraging a multilingual version of the large foundational model MPNet (Song et al. 2020), fine-tuned on semantic search, we perform semantic similarity analyses of the content spread between and amongst these news platforms and our corresponding set of English-language, Russian-language, and Ukrainian-language Telegram channels. Further, by improving upon an online and parallelizable non-parametric version of the K-Means algorithm, we cluster our dataset into fine-grained *topic clusters* to understand the topics discussed.

We find that much of the content shared between Telegram and Russian websites concerns the war in Ukraine and Western sanctions on Russia. By performing this same clustering on semantic content specific to only Russian news websites or only Telegram channels, we find an emphasis on the day-to-day machinations (*e.g.*, bombings of particular bridges) of the Russo-Ukrainian War on Telegram that contrasts with a focus on US politics specific on Russian news outlets. Next, we track the spread of topics amongst and between Telegram and Russian news media. We find that 33.2% of distinct topics discussed on our set of Telegram channels, making up 24.3% of all Telegram messages in our dataset, were first published in Russian news articles. In contrast, 13.9% of topics on Russian news websites, comprising 18.39% of all the text (all messages from Telegram and paragraphs from Russian state media) on our set of Russian websites, were posted first on Telegram. Telegram-originated messages comprise a particularly large amount of content on news websites like waronfakes.com (28.2% of text), ukraina.ru (27.9%), and ura.news (25.6%) that all maintain large Telegram presences. Applying time-series analysis with the Hawkes processes on our topic clusters, we

[1] https://web.archive.org/web/20220304200814/https://www.rt.com/russia/551256-how-access-rt-censorship-bypass/

then estimate the percentage of content on each platform that was influenced by activity on other platforms. While some websites like ura.news had relatively low amounts of content (2.4%) flowing from Telegram, others such as ukraina.ru had much larger (27.2%) percentages of their content possibly influenced by activity on Telegram. Finally, we analyze how quickly topics spread amongst our set of news websites and onto Telegram channels, finding that websites like ura.news, ukraina.ru, and news-front.info, and the Telegram channel @gensham's topics flow most quickly to other platforms.

Our work presents one of the first in-depth analyses of semantic content and semantic similarity among and between Russian state media websites and Telegram channels. We show that by leveraging multilingual models, we can programmatically track the spread of topics and ideas across platforms. As misinformation and propaganda increasingly spread on messaging platforms like Telegram and WhatsApp, we hope that our work can serve as the basis for future studies about the spread of misinformation online.

2 Background and Related Work

Telegram. Telegram, started in 2013, is a messaging platform (Baumgartner et al. 2020) with 700 million monthly users (Singh 2022) as of June 2022. Similar to WhatsApp and Facebook Messenger, users on Telegram can share messages, images, and videos. However, in addition to private conversations, Telegram users can create one-to-many public channels (where only channel creators and administrators can post content) to which other users can subscribe. Within this work, we focus on these administrator-run channels.

With a free-speech ethos, Telegram is a platform where extremist content, misinformation, and propaganda can thrive (Walther and McCoy 2021). Both Höhn, Mauw, and Asher (2022) and Baumgartner et al. (2020), for instance, develop public datasets for specifically analyzing the spread of misinformation on Telegram. Urman and Katz (2022) similarly examine far-right Telegram networks. Following the ban of many Russian news outlets throughout Europe, several Russian outlets have as a result turned to Telegram to spread their content, with Russia Today and Sputnik News even having pages dedicated to showing users how to download Telegram (Bovet and Grindrod 2022; Bergengruen 2022).

Types of Unreliable Information. Unreliable information can take the form of *misinformation*, *disinformation*, and *propaganda*, among other types (Jack 2017). *Misinformation* is any information that is false or inaccurate regardless of the intention of the author. *Disinformation*, in contrast, is false and inaccurate information spread with the express and deliberate purpose to mislead (Jack 2017). Similar to disinformation, *propaganda* refers to "deliberate, systematic information campaigns, usually conducted through mass media forms" regardless of whether the information is true or false (Jack 2017). Within this work, we do not make distinctions between different types of information flowing from Russian propaganda outlets, instead focusing on their overall use of Telegram and their relationships with one another.

Russian Propaganda. While we do not examine specific Russian propaganda stories, instead focusing on the different Russian platforms' use of Telegram, several other works have shown the widespread influence of Russian propaganda. For instance, following the 2016 US election, Badawy et al. found that Russian bots propagated US pro-conservative and divisive messages, especially towards users in the US South (Badawy, Ferrara, and Lerman 2018). A similar study from the RAND corporation identified two communities of over 40,000 users on Twitter that promoted anti- and pro-Russian stories throughout Eastern Europe (Helmus et al. 2018). Finally, Hanley et al. (Hanley, Kumar, and Durumeric 2023b,a) examined the spread of Russian propaganda to US and Chinese social media.

Platform	Art.	Emb.	Platform	Art.	Emb.
Telegram	–	2.48M	rbc.ru	16.0K	179K
geopolitca.ru	544	4.22K	rt.com	12.7K	138K
global-research.ca	5,904	158K	southfront.org	7.79K	79.0K
govorit-moskva.ru	57.9K	223.9K	sputnik-news.com	18.2K	183K
journal-neo.org	2.51K	15.4K	strategic-culture.org	2.17K	25.1K
katehon.com	4.39K	45.8K	tass.com	9.11K	44.9K
lug-info.com	1.22K	8.92K	ukraina.ru	17.4K	193.6K
news-front.info	29.0K	261.9K	ura.news	7.92K	125.6K
			waronfakes.com	2.5K	10.9K

Table 1: The number of articles and embeddings extracted from Telegram and our set of 16 Russian media websites. Articles are embedded on a paragraph level while each Telegram individual message is embedded.

Multilingual Analyses of News and Misinformation. As machine learning and natural language processing tools have improved, various works have taken multi-modal and multilingual approaches to detect and track news, disinformation, and propaganda. Using a multilingual BERT model, Panda and Levitan (2021) analyze the spread of COVID-19 misinformation in Bulgarian, Arabic, and English. We note that multilingual models have been shown to suffer from reduced performance, especially for rarer languages (Joshi et al. 2020). However, as shown by Verma et al. (2022), multimodal approaches and investment in the training of more robust models can help ameliorate many of these issues.

3 Datasets

We use several datasets to understand the interaction amongst and between Russian news media sites and Telegram. We provide an overview of these datasets here.

Russian Propaganda and State Media. Our study examines 11 Russian propaganda and state media websites previously analyzed by the US State Department and prior work (Rus 2020; Hanley, Kumar, and Durumeric 2023b,a) (Table 1). In addition, we extend our set of websites to include five additional Russian-language news websites that have been documented to spread Russian propaganda (Park et al. 2022; Aleksejeva 2022; von Twickel 2017): lug-info.com, govorit-moskva.ru, ura.news, rbc.ru, and ukraina.ru.

For each website, between August 20 and September 25 2022, utilizing the Go library colly and the Python library Selenium, we collect the articles published between

January 1 and September 25, 2022. Specifically, for each website, we scrape 10 hops from the root page, collecting the article content published on each page (*i.e.*, we collect all URLs linked from the homepage [1st hop], then all URLs linked from those pages [2nd hop], and so forth). To extract article content and associated metadata, we use the Python libraries newspaper3k and htmldate. For our Russian-language websites, newspaper3k was largely unable to collect article content. As a result, for each of these websites, we built a custom Python script to parse out article text from the scraped HTML. Altogether, we collect the content for 215,359 articles across the 16 websites (Table 1).

Telegram Channels. To understand Russian platforms' use of Telegram, we curated the set of Telegram channels hyperlinked by our set of Russian news media websites throughout 2022. Specifically, from the pages of our Russian websites, we identified 802 Telegram channels. Removing private Telegram channels and those that were censored by Telegram or deleted (*i.e.*, @rtnews), we were left with a total 732 Telegram channels. Then, as in prior work (Hoseini et al. 2023), we scraped the public content of these channels from www.t.me/channel_name/s/ URLs. For each channel, we scrape all messages that were published between January 1 and September 25, 2022. Across the 732 channels, we collect a total of 2,592,772 Telegram messages (Table 1).

4 Methodology

Having described our dataset, in this section, we outline several key algorithms that we utilize to track information flows across websites. Specifically, after detailing how we preprocess our data, we then describe how we determine the similarity of content between different websites/telegram channels after embedding *text paragraphs* using a multilingual and fine-tuned version of the MPNet (Song et al. 2020) large foundation model. We finally specify the fine-grained clustering mechanism that we use to isolate and track specific topics across our set of websites.

Preprocessing

For each news article and Telegram message, before performing any analysis, we first remove all URLs, emojis, and HTML tags. We further discard any Telegram message that does not contain text or that has fewer than four words (Hanley, Kumar, and Durumeric 2023b). Altogether, we removed 115,208 messages that did not meet our criteria, leaving us with 2.48M Telegram messages for our study (Table 1). For our news articles, we segment each article into its constituent paragraphs based on where newline (`\n`) or tab (`\t`) characters appear within the text. We note that this segmenting approach was successful for every website except sputniknews.com and govoritmoskva.ru. For both of these websites, we built custom scripts to segment their articles based on the text block elements specified in their HTML pages.

Embedding Articles and Messages

For every article and message, we embed its constituent paragraphs using our multilingual MPNet model. The specific model version we use is fine-tuned to the *semantic search task* and trained to handle over 50 languages including English, Russian, and Ukrainian (the primary languages within our dataset).[2] This version of MPNet was trained so that texts with similar semantic content have higher cosine similarity. We utilize the intuition, as in Hanley, Kumar, and Durumeric (2023b), that when embedding text and performing subsequent analyses such as topic analysis, each embedding vector represents only one topic or idea. Any given article can contain multiple topics or ideas; thus we embed each paragraph of each article. Embedding paragraphs rather than sentences as in Hanley, Kumar, and Durumeric (2023b) enables us to obtain additional context while also obtaining an embedding for the (often) one topic/idea present within the paragraph. Altogether we embed 1.62M paragraphs and 2.48M Telegram messages (57 minutes utilizing an Nvidia RTX A6000 GPU).

Many of our articles and messages are not in English. Using the Python library langdetect, we find that 73.2% of article paragraphs are in Russian, 22.3% in English, and 2.1% in Ukrainian. The remaining belong to an assortment of other languages. Similarly, 81.3% of our Telegram dataset is in Russian, 6.0% in English, and 5.3% in Ukrainian.

Comparing Semantic Content

We compare the semantic content of our embedded messages and paragraphs utilizing cosine similarity (Song et al. 2020). As found in previous works (Hanley, Kumar, and Durumeric 2023b; Song et al. 2020; Bernard et al. 2022), a cosine similarity threshold between 0.60 and 0.80 can be utilized to determine whether two pieces of text are about the same topic. For instance, with the same model, Phan et al. (2022), found that a threshold near 0.715 achieved the best results. However, in order to further verify these past results, we perform our own evaluation to determine a threshold at which two messages/paragraphs can reliably be said to be about the same topic.

We take two approaches to determine the appropriate threshold for considering two paragraphs to be about the same topic. First, we take 250 random paragraph pairs with similarities at various thresholds (*i.e.*, for 0.60, messages/-paragraphs with similarities between 0.59 and 0.61) and have an expert determine whether they are about the same topic as outlined in Hanley, Kumar, and Durumeric (2023b); Soper et al. (2021). We perform this evaluation in monolingual settings for English and Russian (the two most prominent languages in our dataset) and a multilingual setting with English and Russian. Given that our expert could not speak Russian, we utilized Google Translate to translate our set of Russian paragraphs into English. As seen in Table 2, our selected pre-trained model achieves a near 85.2% topic-similarity precision at a threshold of 0.7 only for English. For Russian, and in a multilingual setting of English and Russian, our model only achieves this precision at a threshold of 0.8. This is in line with prior work (Verma et al. 2022) that has found that multilingual models often over-perform in English while underperforming in rarer languages.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Threshold	English	Russian	English & Russian
0.6	48.0%	24.4%	28.8%
0.7	85.2%	49.2%	62.4%
0.8	99.2%	89.2%	89.6%
0.9	∼100.0%	∼100.0%	∼100.0%

Table 2: Precision evaluation of whether embedded paragraphs/messages have the same topic at various thresholds and across different languages.

Second, to confirm our manual evaluation, we utilize Chen et al. (2022)'s dataset of 10K multilingual article text paragraph pairs across 18 languages (including Russian and English) annotated for whether they were about the same news story. Analyzed across 7 criteria dimensions, this dataset labeled each paragraph pair on a scale of 1 (not about the same news story) to 4 (about the same news story). Embedding each pair with our multilingual model, we find that 93.6% of paragraph pairs with similarity above that at the threshold of 0.80 had very high similarity (a labeled score of 3.0 or higher). Similarly, 83.6% of pairs below this threshold had lower similarity (lower than 3.0). Benchmarking our approach with this dataset, we confirm that the cosine similarity threshold of 0.80 is a strong indication that two paragraphs are about the same story. Throughout the rest of this work, when comparing two embeddings, we utilize a threshold of 0.8. This new evaluation matches previous works' evaluation of this particular model (Huertas-García et al. 2021; Phan et al. 2022). When two messages/paragraphs reach this threshold, we consider them to be *similar* or to *correspond* with one another.

Computing Similarity Scores Between Platforms. Utilizing our threshold of 0.8, we compute the percentage of different platforms' messages/paragraphs that convey the same topic. This is such that we calculate the percentage of messages/-paragraphs from one website whose topic/idea appears on another website (*i.e.*, what percentage of rt.com's paragraphs also appear on sputniknews.com and conversely what percentage of sputniknews.com's paragraphs appear on rt.com). To consolidate the two similarity values into one average to approximate platform similarity, we take the geometric average of two returned percentages (We take the geometric average rather than arithmetic as the two numbers are largely non-independent (Huntington 1927)).

Isolating Individual Topics

In addition to identifying the similarity between individual messages/paragraphs and amongst websites, we further seek to identify individual topics/units of information within the Russian news ecosystem and track their spread. As in Hanley, Kumar, and Durumeric (2023b), we utilize clustering to identify *topics* present within our dataset. However, finding that the density-clustering approach utilized by Hanley, Kumar, and Durumeric (2023b) could not scale to our dataset, we utilize DPMeans to identify topic clusters.

DP-Means (Dinari and Freifeld 2022) is a non-parametric extension of the K-means algorithm that does not require the specification of the number of clusters *a priori*. Within DP-Means, when a given datapoint is a chosen parameter λ away from the closest cluster, a new cluster is formed, and that datapoint is assigned to it. As such, this enables us to specify how similar individual items must be to one another to be part of the same cluster. Similarly, because DP-Means is non-parametric, it does not require the specification of *a priori* how many topics are present.

We note that we make three key changes to the released version of parallelizable DP-Means:[3] (1) We cluster embeddings based on their cosine similarity with one another rather their Euclidean distances (also altering the cost function to consider cosine similarities); (2) We set new clusters to be formed whenever the message/paragraph is less than $\lambda = 0.8$ similar to its nearest cluster; (3) We remove the random reinitialization of clusters in the released algorithm (Dinari and Freifeld 2022); we find that this step often led to over-clustering given that many website paragraphs are slight variations of each other. These changes enable us to form highly fine-grained and semantically specific topic clusters both in monolingual and multilingual settings.

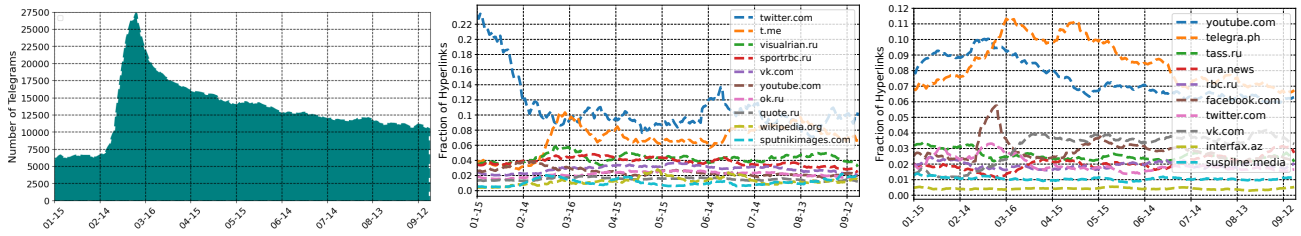
Estimating Platform Influence

We utilize Hawkes processes to estimate the influence of the content on one platform on the content published on another. Hawkes processes are statistical models of event frequencies that account for the effect of other processes (Linderman and Adams 2015). This is such that a Hawkes process of one set of event frequencies can model the influence of another set of frequencies of its own frequency (*e.g.*, rt.com mentioning a story may affect when and how much sputniknews.com reports on that same story). In this work, we fit the time series frequencies of particular events utilizing Gibbs sampling with settings as specified in past works (Linderman and Adams 2015).

Upon fitting our Hawkes processes, we note that the following weights are returned: (1) background rates from each process (also captures the influence of other processes not modeled [*i.e.*, websites not included in our dataset]), (2) influence weights for each process to each other, and finally (3) shapes of the impulses of one process on another process *and* itself. The background rate returned models the expected rate at which events will occur without influence from past events or from other processes (*i.e.*, the rate at which a given platform writes about a given story or topic without influences from other websites or its previous set of articles on the given topic). The influence weights model how one event on one process (*i.e.*, one mention of a given topic) influences the frequency of events on another process. For example, a weight of two from rt.com to sputniknews.com for a given topic means that each mention of that topic results in an *expected* two more mentions by sputniknews.com. Finally, the shape of impulses from one process to another process models the probability that the *expected* events caused by the first process will occur at a given point of time on the second process (*i.e.*, one day after, two days after, *etc...*)

Utilizing these values and the steps laid out in past work (Zannettou et al. 2018), we calculate the influence of

[3] https://github.com/BGU-CS-VIL/pdc-dp-means



(a) Number of Telegrams (b) Hyperlinks from Websites (c) Hyperlinks from Telegram Channels

Figure 1: Following, the Russian invasion of Ukraine, the number of Telegram posts in our set of 732 channels spiked from an average of 6,500 to over 27,500. Similarly, our set of Russian websites began to utilize Western media sites like Twitter less often while increasing their use of Telegram (Telegram links increased from 4% of these websites' external hyperlinks to nearly 10%). Our set of Telegram channels themselves began to steadily use platforms like YouTube less often.

one platform on another (the percentage of a platform's messages or paragraphs that may have been caused by another platform) as well as the efficiency of different platforms in getting their content posted elsewhere (the amount of influence each platform has on another relative to the number of messages it itself posted).

Ethical Considerations We utilize only public data and follow ethical guidelines as outlined by others for scraping data (Hanley, Kumar, and Durumeric 2022). We recognize that the Russo-Ukrainian War is an ongoing conflict and that sensitivity is paramount. We hope that our work provides objective insight into the behavior of Russian media outlets throughout the conflict.

5 Behavior of Russian State Media and Associated Telegram Channels

In this section, we analyze how Russian state media has changed its utilization of Telegram since the beginning of the Russo-Ukrainian War, as well as how the content promoted on the largely uncensored Telegram platform has differed from the content on news media websites. To do so, we first determine whether Russian state media websites promoted Telegram more often than before the onset of the war. Finally, we utilize our topic clustering methodology to determine the topics that were (1) shared, (2) specific to Russian media websites, and (3) specific to Telegram.

The Increased Use of Telegram and the Decline of Western Platforms

As seen in Figure 1a, following the Russian invasion of Ukraine, activity within our set of 732 Telegram channels spiked heavily, growing from an average 6,500 messages per day to nearly 27,500. While by September 2022 posting decreased, the average remained near 11,000 messages/day, illustrating a dramatic increase in activity compared to early 2022. Plotting the top 10 external domains linked to by our set of Russian websites' articles in Figure 1b between January and September 2022, we observe an increase in the promotion of Telegram (t.me domain) by these outlets. While at the beginning of 2022, only 4.0% of links to external domains were to Telegram, this soon spiked to nearly 10.5%, remaining above 6.0% throughout the period measured. This confirms that across Russian media sites, there **has been a dramatic increase in the use and promotion of Telegram since the beginning of the conflict.**

Top YouTube Channels	# Telegram Channels
Россия 24	61
Анатолий Шарий	50
Max Chronicler	46
Комсомольская Правда	39
Информационное агентство БелТА	39
RT на русском	34
Metametrica	34
Fox News	32
ИЗОЛЕНТА live	31
Телеканал Рада	31
SHAMAN	29

Table 3: Top YouTube channels mentioned by our set of 732 Telegram channels.

Consistent with both the banning of many Russian media firms by Western social media sites and the corresponding banning of Western social media platforms by the Russian government (Chee 2022; Bergengruen 2022), in Figure 1b, we observe a slight decrease in hyperlinks to Twitter from Russian state media sites. Extracting the external hyperlinks from our set of 732 Telegram channels, we similarly observe a steady decrease in the use of YouTube. Many of the most hyperlinked YouTube channels (Table 3) were pro-Russian channels that were later restricted (Wong 2022). We note the presence of several conservative news websites within the top-linked YouTube channels. As noted elsewhere (Stone 2022), US-based conservatives have sometimes adopted Russian state narratives surrounding the war. This is largely reflected in the number of Telegram channels linking to YouTube channels like Fox News and Metamerica (Table 3). We thus see **a decline in the use of Twitter and YouTube by Russian state media organizations following the invasion of Ukraine**.

The Shared Ecosystem

Having observed a noted increase in the usage of Telegram, we now determine the shared and differing content within these different platforms. To understand the degree to which different Russian websites in our dataset conformed to the messages and topics present within Telegram,

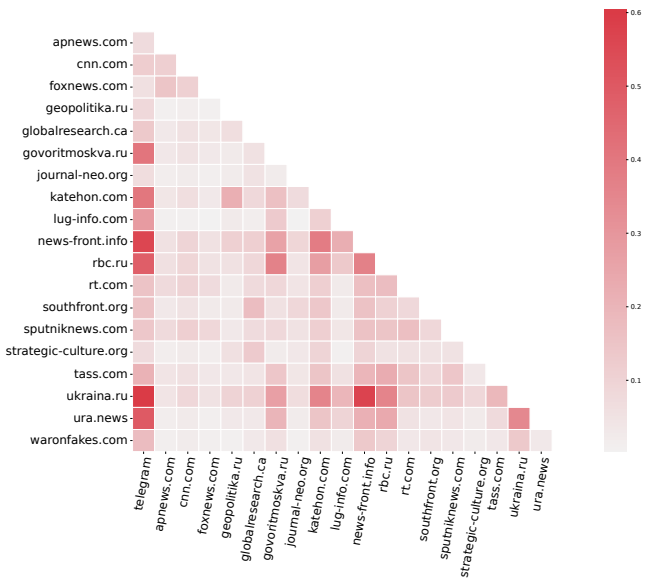


Figure 2: Similarity matrix between our considered websites. Ukraina.ru and newfront.info have the highest similarity with one another. All Russian websites have a high content similarity with the collective messages posted to our set of 732 Telegrams. The three US-based websites have high similarities amongst themselves.

we determine the semantic similarity between our set of state media websites and linked Telegram channels. Altogether 1,812,648 out of the 2,477,564 (73.2%) Telegram messages had a corresponding paragraph on a Russian site. Conversely, 899,589 of the 1,616,946 paragraphs (55.6%) from our set of Russian websites had a corresponding Telegram message. We thus observe a high degree of but not complete similarity in terms of topics on Russian media articles compared to Telegram.

Similarity to Telegram. We now determine the similarity scores between our given platforms (Section 4) based on the geometric mean of their percentages of shared messages. We note that for this analysis, we combine our set of Telegram messages (as if all came from one website), rather than examining each of the Telegram channels individually. To provide a reference point, as well as to validate our metric, we scrape and compare the similarities for three English-language news websites: cnn.com, foxnews.com, and apnews.com. Using the methodology outlined in Section 3, we gather an additional 41,452, 78,494, and 104,206 articles from cnn.com, foxnews.com, and apnews.com respectively. We follow the same methodology outlined for segmenting and embedding the articles (Section 4).

Calculating the similarities between each website and Telegram, as seen in Figure 2, the US-based websites have the highest semantic similarity with each other, and the second highest semantic similarities with the Russian websites sputniknews.com, rt.com, and tass.com. We note that all of the Russian websites included in our dataset have relatively high semantic similarity with our set of Telegram messages compared with cnn.com, foxnews.com, and apnews.com with ura.news having the highest similarity and journal-neo.org having the least.

Website	Sim. Telegram	Simalarities
geopolitika.ru	rossiyaneevropa,	0.116,
globalresearch.ca	gr-crg	0.109
govoritmoskva.ru	radiogovoritmsk	0.340
journal-neo.org	rossiyaneevropa	0.095
katehon.com	rossiyaneevropa	0.364
lug-info.com	lic-lpr	0.516
news-front.info	ukraina_ru	0.473
rbc.ru	rbc_news	0.486
rt.com	vzglyad-ru	0.135
southfront.org	southfronteng	0.150
sputniknews.com	vzglyad_ru	0.125
strategic-culture.org	strategic_culture	0.087
tass.com	tassagency_en	0.273
ukraina.ru	ukraina_ru	0.526
ura.news	vzglyad_ru	0.354
waronfakes.com	warfakes	0.299

Table 4: Most similar Telegram chanels to each website.

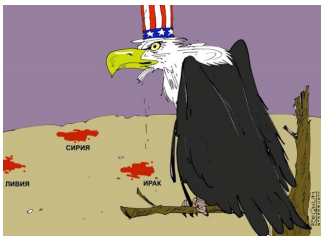


Figure 3: Propaganda image from @rossiyaneevropa. @rossiyaneevropa argues NATO cannot criticize Russia's activities in Ukraine given the West's war crimes in Libya.

Most Similar Telegram Channels to Russian Websites. To further examine the correspondence between Telegram and our set of Russian media outlets, we determine the most similar sets of Telegram channels to each website. As seen in Table 4, several Telegram channels individually post many of the same topics present on Russian websites. Unsurprisingly, across our set of websites, the most similar Telegram channels to each of our websites are the official Telegram channels utilized by these online newspapers.

Besides the official channels, one of the the channels most similar to our Russian websites is @rossiyaneevropa (14.6K subscribers), a pro-Russian channel, ostensibly run by "Alexander Burenkov, director of the Institute of Russian-Slavonic Studies." As pictured in Figure 3, this channel often amplifies Russian propaganda. We further observe that several other similar Telegram channels are run by Russian state-controlled media (@vzglyad_ru, 88.1K subscribers) or by Russian-backed Ukrainian-separatists (@nm_dnr, 85.5K subscribers; @lic_lpr 31.3K subscribers; @gtrklnr_lugansk24 7.4K; @dnr_sckk 17.2K). In addition, to these channels, we further find two of the other most similar Telegrams (not shown in Table 4) @rentv_news (49.7K subscribers) and @killnet_reservs (93.9K subscribers) are pro-Russian channels that often repeat or rebroadcast videos, official documents from the Russian government, disinformation, and anti-American and anti-Ukrainian messages. For example, on October 10, 2022, @killnet_reservs posted a message advocating for their subscribers to disrupt

	Telegram Specific Topics	Telegrams	Russian Site Specific Topics	Articles	Shared Telegram and Russian Site Topics	Telegrams +Articles
1	Lukashenka on the possibility of achieving peace in Donbas (Russian → English)	5,295	According to him, the problem is very acute, especially since now it is very difficult to engage in unauthorized extraction of gas going to Europe from the Ukrainian pipeline because this will be monitored not only in Russia, which is losing money but also in Europe. (Russian → English)	635	Regarding the events in Shchastya, Luhansk region, where militants hit the building of the fire and rescue department (Ukrainian → English)	19,467
2	Footage of live firing of Grad multiple launch rocket systems of a self-propelled artillery regiment of the Central Military District in the zone of a special military operation. (Russian → English)	2,326	Mass riots in Iran get out of control of the security forces Amid the riots in Iran, caused, allegedly, by direct incitement from the United States (Russian → English)	436	The Russian Defense Ministry showed the advancement of airborne units during a special military operation in Ukraine. (Russian → English)	9,939
3	Krajina security forces will not be given a corridor to exit Mariupol, Basurin said. (Russian → English)	2,182	Cessation of hostilities and political dialogue, negotiations, mediation, and other peaceful means aimed at achieving a lasting peace. (Russian → English)	197	Today, the Russian military attacked the Khmelnytsky region from the air. According to the head of the Khmelnytsky OVA Serhiy Gamaly, there were no casualties. (Ukrainian → English)	5,217

Table 5: The top three topics—by the number of distinct articles or distinct Telegram messages—within the Telegram-specific, the Russian website-specific, and the shared Telegram and Russian website ecosystems.

US infrastructure:

> *We invite everyone to commit DDOS on the civilian network infrastructure of the United States of America! Subject to DDOS attacks are:- All Airports - Marine terminals and logistics facilities - Monitoring weather centers - Health care system - Metro (buying tickets, registering the route) - Exchanges and online trading systems.*[4]

We note that while we utilize this methodology to identify the messages that bear the closest resemblance to specific Russian websites, and thus to content sponsored by Russian=-state media, among our set of 732 Telegram channels, this methodology can be easily extended to identify suspect Telegram channels that repeat or mention propaganda on a much wider scale, which we leave to future work.

Most Prominent Topics in Shared Ecosystems. Next, to understand the most prominent topics present among all shared texts between Telegram and Russian websites, we utilize our clustering algorithm outlined in Section 4. Specifically, we cluster only the set of paragraphs from Russian websites that had a corresponding Telegram message together with all Telegram messages that had a corresponding paragraph on a Russian website; altogether this consists of 2.7M messages/-paragraphs (66.24% of all texts [from both Telegram and Russian articles]). We note, as previously discussed, that due to the intrinsic guarantees of our clustering algorithms, each embedding has a high cosine similarity with the center. After clustering with $\lambda = 0.8$, our embeddings had an average of 0.882 cosine similarity with their respective cluster centers. Each cluster had an average cosine similarity of 0.0292 with the remaining cluster centers. This illustrates the degree of semantic cohesiveness within each cluster and that each cluster is distinct. For each cluster, to build an intuition for the topic, we extract the most representative paragraph/message from the cluster by getting the paragraph with the highest cosine similarity to the cluster center.

Determining the top topics by the number of articles, the most shared topic within our datasets concerned the destruction of buildings following Russia's invasion of Ukraine, specifically, the operations in the Luhansk region (Table 5). This was largely anticipated given how pivotal the Donbas region became to the war and propaganda surrounding it (Laryš and Souleimanov 2022); 19,467 separate articles (note, not paragraphs) and Telegram messages mention the topic. The second most popular topic with 9,939 messages/articles was about the actual invasion of Ukraine by Russian units. Within Vladimir Putin's declaration indicating Russia's intention to invade Ukraine, he called for a "special military operation" to "denazify" and "demilitarize" Ukraine. We see this particular language repeated by both Russian propaganda outlets and Telegram channels. The third topic (again a specific aspect of the war) concerns the Russian bombing of the Khmelnytsky region of Ukraine on April 28, 2022.

Topics Specific to Russian News Websites

Having explored the shared content amongst Russian news and Telegram channels, we now analyze the content specific to our set of Russian news websites. We cluster only the set of paragraphs from Russian sites that did not have a corresponding Telegram message in our dataset (727,357 paragraphs from 120,198 articles). Across this clustering, each paragraph had an average 0.926 cosine similarity with their respective cluster center. Each cluster has an average similarity of 0.0674 with other clusters in the dataset.

As seen in Table 5, the most common topic specific to Russian news websites concerned the extraction of gas from Ukraine to Europe; an estimated 635 articles (note, not paragraphs) across our 16 news websites were associated with this topic. As noted by others, following the Russian inva-

[4]https://web.archive.org/web/20221011223806/https://t.me/s/killnet_reservs

sion, the sanctions on Russian gas and oil to Europe have caused enormous major economic fallout (Soldatkin, Donovan, and Faulconbridge 2022) and we see mentions of this topic repeated across our Russian media ecosystem. Besides content concerning the shipment of gas from Ukraine to Europe, we further see content about protests in Iran that followed the death of Mahsa Amini on September 16, 2022 (436 articles) and calls for peace talks between Russia and Ukraine (197 articles).

Topics Specific to Telegram

We now extract content specific to Telegram. To do so, we cluster only the set of messages from our Telegram channels that did not have a corresponding text from our Russian media outlets' articles. Altogether, we cluster the 664,916 distinct Telegram messages. Across this clustering, each Telegram message had an average 0.900 cosine similarity with their respective cluster center. Each cluster center has an average cosine similarity of -0.0708 with the other clusters.

As seen in Table 5, many of the most prominent messages concern specific day-to-day updates on the war in Ukraine (Topic 2 and 3). For example, the topic cluster with the second most Telegram messages concerns updates on the rocket launches of a Ukrainian city (Holder, Hernandez, and Huang 2022). The Telegram topic with the most corresponding messages however concerns an interview with Belarusian President Alexander Lukashenko on the possibility of peace in the Donbas region of Ukraine. We thus see within many of these Telegram-specific content clusters highly specific updates about news and individual interviews rather than larger news stories.

6 The Spread of Topics

Having analyzed the static behavior of the topics shared among our Telegram and Russian news ecosystems, in this section, we examine the speed at which topics spread amongst and between Russian websites and Telegram channels. To estimate the spread of topics amongst and between different Russian websites and Telegram channels, we first cluster all 4,094,510 texts (Telegram messages and paragraphs from articles) as specified in Section 4. Across this new clustering, with 29,875 unique centers, each message/-paragraph had an average cosine similarity of 0.883, with its respective cluster center. Each cluster further had an average cosine similarity of 0.0321 with the remaining clusters. We note that when performing our analysis using the timestamps of our articles and Telegram messages given that we only managed to determine the publish date and not the *exact* timing information of our articles' publication (*i.e.*, the hour and second of publication), we perform our analysis on the scale of days. This is such that we consider a Telegram message to precede an article only if it was published at least a day before an article (*e.g.*, 04-05-2022 before 04-06-2022).

Content First Published on Telegram and Russian Websites. To understand the interchange of topics between our Russian website and Telegram datasets, we first determine the percentage of each website's topics that began on Telegram and *vice versa*. We note that for this analysis, we remove the set of Telegram channels that officially operate in coordination with each of our news websites. This enables us to determine how much each website's content began on ostensibly "independent" (*i.e.*, not controlled by same Rusian-state media entities) Telegram channels.[5] Determining the amount of a website's content that each cluster contains (*i.e.*, the number of paragraphs), we thus determine the percentage of each platform's content that was first posted on Telegram (Table 6).

Website	% First on Telegram	Website	% First on Telegram
geopolitika.ru	8.10%	rt.com	6.70%
globalresearch.ca	4.60%	southfront.org	8.55%
govoritmoskva.ru	11.6%	sputniknews.com	5.06%
journal-neo.org	7.83%	strategic-culture.org	6.06%
katehon.com	15.0%	tass.com	11.6%
lug-info.com	19.7%	ukraina.ru	27.9%
news-front.info	21.2%	ura.news	25.6%
rbc.ru	16.1%	waronfakes.com	**28.2%**

Table 6: Percentage of each website's topics that were posted first on Telegram (after removal of official channels). We bold the largest percentage.

Website	Telegram	% Content First Posted
geopolitika.ru	rossiyaneevropa	0.47%
globalresearch.ca	karaulny	0.32%
govoritmoskva.ru	karaulny	1.5%
journal-neo.org	parstodayrussian	0.50%
katehon.com	karaulny	1.0%
lug-info.com	lic_lpr	2.7%
news-front.info	karaulny	1.2%
rbc.ru	karaulny	1.6%
rt.com	karaulny	0.43%
southfront.org	new_militarycolumnist	0.57%
sputniknews.com	karaulny	0.38%
strategic-culture.org	karaulny	0.37%
tass.com	karaulny	0.61%
midrule ukraina.ru	karaulny	1.7%
ura.news	karaulny	2.6%
waronfakes.com	senkevichonline	5.5%

Table 7: Top telegrams — by percentage of content first posted — that post content prior to our set of websites.

As seen in Table 6, even after removing official Russian Telegram channels, most of our websites had a noticeable portion of their content that was first posted on Telegram. This is particularly true of websites like waronfakes.com (28.2%), ura.news (25.6%), and ukraina.ru (27.9%). Across all websites, 13.9% of their topics began on Telegram, accounting for 18.4% of all paragraphs in our dataset. We further see that as a whole 33.2% of Telegram's topics began on Russian websites, accounting for 24.3% of all Telegram content. Thus while several of our news websites like waronfakes.com, ura.news, ukraina.ru, and rbc.ru, indeed publish content after it first appears (at least a day later) on Telegram, Telegram channels themselves also utilize topics from Russian news sources for nearly a third of their topics.

[5]We remove the following Telegram channels: tass_agency, tassagency_en, uranews, radiogovoritmsk, ukraina_ru, rbc_news, southfronteng, strategic_culture, waronfakesen, warfakeres, warfakes, warfakebelgorod, warfakeskrm, warfakeszo, and warfakers.

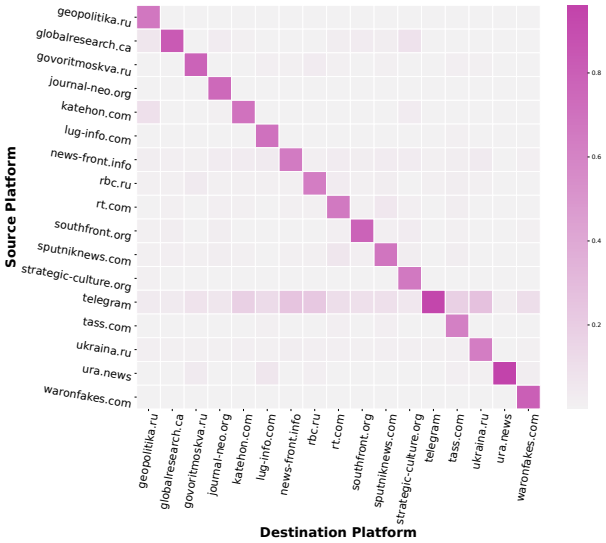


Figure 4: We report the percentage of each platform's content that was estimated (using the methodology outlined in Section 4) to have been influenced by another platform.

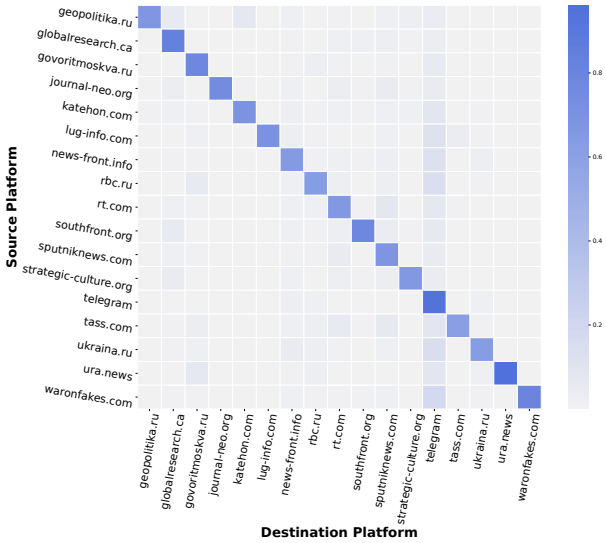


Figure 5: We report the estimated efficiency (using the methodology outlined in Section 4) of each platform in getting their content onto different platforms.

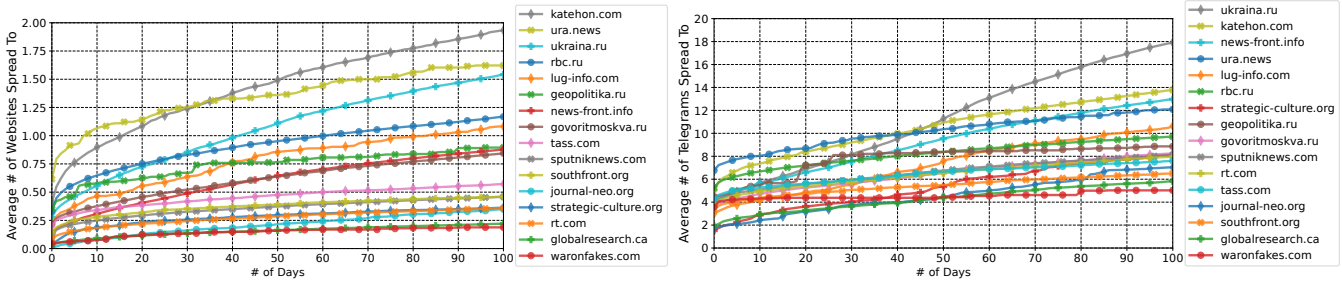
Examining the set of Telegram channels that most often first posted each website's content in Table 7, we see the Telegram channels @kalaruny (159K subscribers), @new_militarycolumnist (213K), and @bbbreaking (1.38M; not shown in table) often posted content first across nearly all of our websites. As reported elsewhere, @karaulny (screen name Karaulny-Z) is a pro-Russian government channel that abides by a "stop list" of prohibited topics and that was purchased by supporters of the Kremlin in 2017 (Rothrock 2018). @new_militarycolumnist, another pro-Russian channel, gives continuous updates on the Russian military (Sabah 2020). Finally, @bbbreaking, another pro-Russian Russian-language channel, with the motto *"Earlier than others. Almost"* appears to also in several cases give updates on the news before many of our Russian news websites.

Influence Estimation through Hawkes Processes. As just seen, a substantial portion of the content on websites like waronfakes.com, ura.news, and rbc.ru first appears on Telegram. Despite this and even given the close relationship that many of these websites have with our set of Telegram channels, it is largely infeasible to *know* definitively if these topics' appearance on Telegram *caused* their later appearance on our set of websites. However, in this section, we utilize Hawkes processes, to estimate the probability that a message first appearing on a given Telegram influenced a given website to write about similar content. Utilizing this approach, we give the estimated percentage of content that appears on one platform that may have been influenced by another platform as well as the efficiency of this influence.

As previously specified, to estimate the influence of one platform on another (Section 2), we fit 17 Hawkes interacting processes (one for each platform) using Gibbs sampling for our 29.9K topic clusters and utilizing the daily frequencies of each website reporting on each of these topics (Linderman and Adams 2015; Zannettou et al. 2018). We report the estimated percentage of each platform's paragraph/messages that may have been caused by the other platforms in our dataset (Figure 4). We note that because we limited our study to our set of 16 websites and 732 Telegram channels, other platforms' influence would be included within each website's estimated self-influence as its background rate.

Plotting the estimated influence of each website on each other and Telegram, we see in Figure 4, that Telegram typically has a moderate estimated influence on the content published on each of the websites. Most prominently, Telegram has a possible influence on 18.4% of content on katehon.com, 25.2% on news-front.info, 23.0% on rbc.ru, and 26.7% on ukraina.ru. Conversely, it has the weakest influence on the website ura.news (despite often writing about its topics prior to ura.news [Table 6]). Conversely, with the smallest amount of articles, we see that geopolitika.ru also had the smallest estimated influence on Telegram. Further as largely expected, we see in the estimated topic spread efficiency in Figure 5 that many of our websites are somewhat efficient at seeing their content migrate to Telegram, with waronfakes.com being the best at 18.8% efficiency. This indicates across our websites, many may have to post a few articles before that article's topic appears on Telegram.

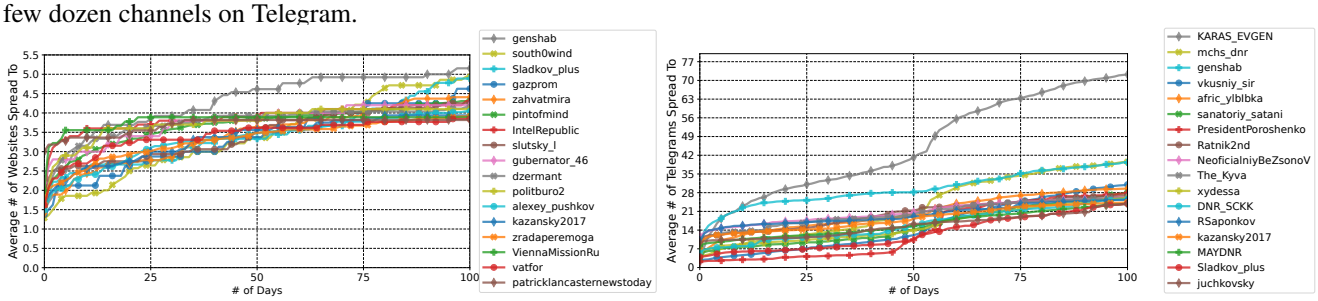
In addition, to estimating the influence of each website on our set of Telegram channels, we further estimate the relationships that each website has with one another. As largely expected prominent websites like rt.com and sputniknews.com have the largest estimated possible influence on the content of one another relative to other websites (around 6% [Figure 4]). Looking at the efficiency of these influence relationships, we see again a stronger relationship between many of the large English-language Russian news sites (tass.com, sputniknews.com, and rt.com). Similarly, we see again that news-front.info, which has been



(a) Russian Websites to Russian Websites

(b) Russian Websites to Telegram

Figure 6: Katehon.com is the most effective at getting its original content reposted by other Russian news websites. Despite many Russian websites utilizing Telegram, it often takes months on average before a topic is widely addressed by more than a few dozen channels on Telegram.



(a) Telegram to Russian Websites

(b) Telegram to Telegram

Figure 7: Despite their smaller size several Telegram channel post topics and content that are later repeated by Russian news outlets. The pro-Russian channel @genshab is particularly effective at getting its content echoed within this ecosystem.

previously documented as heavily influencing conversations within Russian propaganda ecosystems (Hanley, Kumar, and Durumeric 2023b) has a marked estimated influence on other websites, compared to the rest of the websites, playing a possible role in nearly 3% of the content on each of other websites (Figure 4).

Speed of Spread of Topics From Russian Websites. Having estimated the influence of each website, we now model and determine how quickly, on average, *original content* (*i.e.*, content whose topic was first posted on a particular website or a Telegram channel) from each of our websites and Telegram channels travel to *other* websites and Telegram channels. As in Section 5, we perform this analysis on the scale of days. As seen in Figure 6a, on average, katehon.com and ura.news's topics tend to spread the fastest. Across all topics, we further observe that after initially spiking in popularity on the first day it is published, topics often take a longer period to spread throughout the entire ecosystem. Examining each topic individually, we found one of the fastest topics to spread was a story that originated on sputniknews.com, rt.com, and tass.com. This story, which reached every website within six days, was about a potential United Nations-brokered meeting between Russian Federation President Vladimir Putin and Ukrainian President Volodymyr Zelenskyy.[6]

Examining the spread of each website's content to our set of 732 Telegram channels in Figure 6b, we see that on average, stories did not typically spread to more than approximately 20 Telegram channels within the first 100 days. Furthermore, we see that the websites whose content spreads the fastest katehon.com and ura.news, are again the websites whose content spreads the quickest on Telegram. Indeed, we see among our websites, katehon.com, ura.news, ukraina.ru, and news-front.info, are the best progenitors of content across both other websites as well as Telegram. This adheres to results in prior work that show newsfront.info and katehon.com as two of the key creators of Russian propaganda on the Internet (Hanley, Kumar, and Durumeric 2023b; Rus 2020). Examining one of the most prolific stories, one originating from katehon.com published the military results of the first day of the Russian invasion of Ukraine and spread to 613 different Telegram channels.[7]

Spread of Topics From Telegram Channels. We now examine the rate at which topics spread from some Telegram channels amongst themselves *and* to Russian websites. We display the set of 16 Telegram channels whose content spreads the furthest (*i.e.*, the channels that originated topics that then spread prolifically and widely) in Figure 7. As seen in Figure 7, one of the Telegram channel most effective

[6]https://web.archive.org/web/20220919021146/https://sputniknews.com/20220919/putin-zelensky-meeting-far-from-possible-but-un-ready-to-help-facilitate-guterres-says-1100939551.html

[7]https://web.archive.org/web/20220510150650/ttps://katehon.com//ru/news/igor-strelkov-itogi-pervogo-dnya-boevyh-deystviy-i-ih-kratkiy-analiz

at originating content/topics that are then reposted elsewhere *both* on Telegram and within the Russian website ecosystem is @gensab (86.9K subscribers). A pro-Russian propaganda channel, the channel continuously comments on the Russian invasion of Ukraine, writing on August 29th:

> *"Kyiv launched a widely publicized "offensive" on Kherson. Due to the clumsy, on the verge of debility fake propaganda, the offensive is developing so far only in the minds of propagandists and those who believe them."*[8]

Finally, again examining the individual topics from Telegram channels that spread the furthest (the most amount of Telegram channels), we see a message from @uranews about the Donbas regions of Ukraine ostensibly wanting to declare independence and join Russia that spread to 557 other Telegram channels.[9]

> *Everyone to defend the Motherland! Donbass is our land!" The DPR authorities, with the help of billboards, are calling on the residents of Donetsk to take arms in their hands to defend the republic from Ukrainian aggression.*[10]

We further see a message from the pro-Russian Telegram channel @optimisticus007 (35.5K subscribers) that spread to all our websites echoing the desire for Russia to win in Ukraine.

> *Zhirinovsky's last speech in the State Duma: We must win. This is our last decisive battle - the spring of 22. Let it be in April, May, but this year it can be done. We will win!*

7 Discussion and Conclusion

In this work, we explored the usage of Telegram by Russian news websites. To do this, we introduced a new, scalable methodology for tracking news narratives across multiple languages and platforms. Our approach, unlike past work, does not depend on dimensionality reduction across all articles at once nor a complete pair-wise similarity calculation, which allows us to scale to several orders of magnitude more articles and to track topics across 215K news articles and 2.48M Telegram messages. We find that, unexpectedly, Russian media organizations have not only dramatically increased their usage of Telegram but also are regularly sourcing news topics from the messaging platform. We discuss several implications and future directions of this work here.

Identifying Propaganda Channels. As briefly discussed in Section 5, our approach can also be used to identify new Telegram channels that align with Russian state-promoted narratives. As was seen in Table 4, our method identified Telegram channels that supported anti-Ukrainian and pro-Russian-separatist sentiments in an automated fashion (*e.g.*, @lic_lpr, @gtrklnr_lugansk24, @dnr_sckk) as well as several channels that largely repeated Russian state-media topics (*e.g.,* @rossiyaneevropa, @pintofmind, @russtrat) (Afroz and Sehgal 2022). We note that while this method requires lists of predefined propaganda or disinformation platforms/websites, our method can further be utilized beyond Russian and Ukrainian-focused topics to identify websites that are similar to disinformation, hyperpartisan, or other types of malicious websites.

Content Spreading From Different Platforms. By utilizing MPNet, DP-Means clustering, and Hawkes processes, we estimated the role of Telegram and identified the most influential Russian websites within our ecosystem. Unlike past work that has relied on the presence of hyperlinks (Hanley, Kumar, and Durumeric 2022), our approach approximates influence by directly measuring topics themselves. This approach can be extended to understand how larger platforms like Facebook, Reddit, and Twitter interact with Russian propaganda. While we limited ourselves to analyzing the semantic correspondence of Russian websites, our approach can be used to estimate the influence of news websites more generally (*e.g.*, cnn.com, foxnews.com).

Responding to Propaganda. As seen in this work, for websites like waronfakes.com, nearly 30% of their content appears on Telegram at least a day before appearing on the website. By monitoring and understanding the Telegram ecosystem, researchers and fact-checkers can more quickly respond to topics that start on Telegram and make their way to other news sites. We argue that by ignoring Telegram, the research community has discounted a large and influential aspect of Russian propaganda pathways.

[8]https://web.archive.org/web/20220829120625/https://t.me/genshab/846

[9]https://web.archive.org/web/20220515075654/https://t.me/uranews/41595

[10]https://web.archive.org/web/20220515083452/https://t.me/optimisticus007/155

References

2020. GEC Special Report: Russia's Pillars of Disinformation and Propaganda - United States Department of State.

Afroz, S.; and Sehgal, V. 2022. Russian Disinformation Spreading Across the Globe — Avast.

Aleksejeva, N. 2022. Russian War Report: Kremlin-controlled outlet rehashes narrative that Poland plans to annex western Ukraine - Atlantic Council.

Badawy, A.; Ferrara, E.; and Lerman, K. 2018. Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign. In *IEEE Conf. on advances in social networks analysis and mining (ASONAM)*.

Baumgartner, J.; Zannettou, S.; Squire, M.; and Blackburn, J. 2020. The pushshift telegram dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 840–847.

Bergengruen, V. 2022. Telegram Becomes a Digital Battlefield in Russia-Ukraine War — Time.

Bernard, G.; Suire, C.; Faucher, C.; Doucet, A.; and Rosso, P. 2022. Tracking news stories in short messages in the era of infodemic. In *Intl. Conference of the Cross-Language Evaluation Forum for European Languages*.

Bovet, A.; and Grindrod, P. 2022. Organization and evolution of the UK far-right network on Telegram. *Applied Network Science*, 7(1): 1–27.

Chee, F. Y. 2022. EU bans RT, Sputnik over Ukraine disinformation — Reuters.

Chen, X.; Zeynali, A.; Camargo, C.; Flöck, F.; Gaffney, D.; Grabowicz, P.; Hale, S.; Jurgens, D.; and Samory, M. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *16th International Workshop on Semantic Evaluation*.

Dinari, O.; and Freifeld, O. 2022. Revisiting DP-Means: Fast Scalable Algorithms via Parallelism and Delayed Cluster Creation. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Hanley, H. W.; Kumar, D.; and Durumeric, Z. 2022. No Calm in The Storm: Investigating QAnon Website Relationships. In *Intl. AAAI Conference on Web and Social Media*.

Hanley, H. W.; Kumar, D.; and Durumeric, Z. 2023a. "A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems' Coverage of the Russo-Ukrainian War. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 339–350.

Hanley, H. W.; Kumar, D.; and Durumeric, Z. 2023b. Happenstance: utilizing semantic search to track Russian state media narratives about the Russo-Ukrainian war on Reddit. In *Proceedings of the international AAAI conference on web and social media*, volume 17, 327–338.

Helmus, T. C.; Bodine-Baron, E.; Radin, A.; Magnuson, M.; Mendelsohn, J.; Marcellino, W.; Bega, A.; and Winkelman, Z. 2018. *Russian social media influence: Understanding Russian propaganda in Eastern Europe*. Rand Corporation.

Höhn, S.; Mauw, S.; and Asher, N. 2022. BelElect: A New Dataset for Bias Research from a "Dark" Platform. In *Intl. AAAI Conference on Web and Social Media*.

Holder, J.; Hernandez, M.; and Huang, J. 2022. Russia's Shrinking War - The New York Times.

Hoseini, M.; Melo, P.; Benevenuto, F.; Feldmann, A.; and Zannettou, S. 2023. On the globalization of the QAnon conspiracy theory through Telegram. In *Proceedings of the 15th ACM Web Science Conference 2023*, 75–85.

Huertas-García, Á.; Huertas-Tato, J.; Martín, A.; and Camacho, D. 2021. Countering misinformation through semantic-aware multilingual models. In *International conference on intelligent data engineering and automated learning*.

Huntington, E. V. 1927. Sets of independent postulates for the arithmetic mean, the geometric mean, the harmonic mean, and the root-mean-square. *Transactions of the American Mathematical Society*, 29(1): 1–22.

Jack, C. 2017. Lexicon of lies: Terms for problematic information. *Data & Society*, 3(22): 1094–1096.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *58th Annual Meeting of the Association for Computational Linguistics*.

Laryš, M.; and Souleimanov, E. A. 2022. Delegated Rebellions as an Unwanted Byproduct of Subnational Elites' Miscalculation: A Case Study of the Donbas. *Problems of Post-Communism*, 69(2): 155–165.

Linderman, S. W.; and Adams, R. P. 2015. Scalable bayesian inference for excitatory point process networks. *arXiv preprint arXiv:1507.03228*.

Panda, S.; and Levitan, S. I. 2021. Detecting Multilingual COVID-19 Misinformation on Social Media via Contextualized Embeddings. In *Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*. Online: Association for Computational Linguistics.

Park, C. Y.; Mendelsohn, J.; Field, A.; and Tsvetkov, Y. 2022. Challenges and Opportunities in Information Manipulation Detection: An Examination of Wartime Russian Media. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5209–5235. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Phan, Q. L.; Doan, T. H. P.; Le, N. H.; Tran, N. B. D.; and Huynh, T. N. 2022. Vietnamese Sentence Paraphrase Identification Using Sentence-BERT and PhoBERT. In *International Conference on Intelligence of Things*.

Rothrock, K. 2018. New investigative report explains how the Kremlin conquered Russia's Telegram channels — Meduza.

Sabah, D. 2020. US forces block another Russian convoy in Syria as tensions rise. Accessed: 2023-01-15.

Singh, M. 2022. Telegram tops 700 million users, launches premium tier — TechCrunch. Accessed: 2022-08-25.

Soldatkin, V.; Donovan, K.; and Faulconbridge, G. 2022. Russian pipeline gas exports to Europe collapse to a post-Soviet low — Reuters.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Adv. in Neural Information Processing Systems*.

Soper, E.; Hosier, J.; Bales, D.; and Gurbani, V. K. 2021. Semantic Search Pipeline: From Query Expansion to Concept Forging. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2309–2314. IEEE.

Stone, P. 2022. Top US conservatives pushing Russia's spin on Ukraine war, experts say — The Guardian.

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

Verma, G.; Mujumdar, R.; Wang, Z. J.; De Choudhury, M.; and Kumar, S. 2022. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. In *International AAAI Conference on Web and Social Media*.

von Twickel, N. 2017. Annual Report on the Events in the "People's Republics" of Eastern Ukraine 2016. *Berlin: Germany-Russian Exchange (DRA-Deutsch-Russischer Austausch)*.

Walther, S.; and McCoy, A. 2021. US extremism on Telegram. *Perspectives on Terrorism*, 15(2): 100–124.

Wong, Q. 2022. Facebook, YouTube to Restrict Some Russian State-Controlled Media Across Europe — CNET. https://www.cnet.com/news/politics/facebook-youtube-to-restrict-some-russian-state-controlled-media-across-europe/.

Zannettou, S.; Caulfield, T.; Blackburn, J.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the origins of memes by means of fringe web communities. In *ACM Internet Measurement Conference*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our work analyzes news articles in an international context without commenting on any specific culture.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. Our abstract is largely reflective of our work.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. We outline our methodology Section 4 .
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in Section 3, we detail the various aspects of our datasets.
   (e) Did you describe the limitations of your work? Yes, and we detail the implications and some limitations of our work in the Discussion.
   (f) Did you discuss any potential negative societal impacts of your work? No, we do not believe that there are any immediate negative social repercussions of our work.
   (g) Did you discuss any potential misuse of your work? No; our work is largely an analysis and we see no immediate misuse or negative social repercussions.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No, given the largely analytical nature of our work, we do not see immediate negative social consequences and we have documented the models and methods that we utilize.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we have included a GitHub link to the original code for DP-Means clustering from Dinari et al (Dinari and Freifeld 2022).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, we have outlined these details in Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? We performed clustering so there is not an immediate error bar analysis. However, we did provide an estimated precision of similar content in each cluster.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we have outlined these details in Section 3.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, we have outlined these details in Section 4.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes, we have included a GitHub link to the original code for DP-Means clustering from Dinari et al (Dinari and Freifeld 2022).
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? NA
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA