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

Reviewing War: Unconventional User Reviews as a Side Channel to Circumvent Information Controls

José Miguel Moreno[1], Sergio Pastrana[1], Jens Helge Reelfs[2], Pelayo Vallina[1, 3], Savvas Zannettou[4], Andriy Panchenko[2], Georgios Smaragdakis[4], Oliver Hohlfeld[2, 5], Narseo Vallina-Rodriguez[3], Juan Tapiador[1]

[1] Universidad Carlos III de Madrid, Madrid, Spain
[2] Brandenburg University of Technology (BTU), Cottbus, Germany
[3] IMDEA Networks Institute, Madrid, Spain
[4] Delft University of Technology, Delft, Netherlands
[5] University of Kassel, Kassel, Germany
josemore@pa.uc3m.es, spastran@inf.uc3m.es, reelfs@b-tu.de, pelayo.vallina@imdea.org, s.zannettou@tudelft.nl, andriy.panchenko@b-tu.de, g.smaragdakis@tudelft.nl, oliver.hohlfeld@uni-kassel.de, narseo.vallina@imdea.org, jestevez@inf.uc3m.es

Abstract

During the first days of the 2022 Russian invasion of Ukraine, Russia's media regulator blocked access to many global social media platforms and news sites, including Twitter, Facebook, and the BBC. To bypass the information controls set by Russian authorities, pro-Ukrainian groups explored unconventional ways to reach out to the Russian population, such as posting war-related content in the user reviews of Russian businesses available on Google Maps or Tripadvisor. This paper provides a first analysis of this new phenomenon by analyzing the unconventional strategies used to avoid state censorship in the Russian Federation during the conflict. Specifically, we analyze reviews posted on these platforms from the beginning of the war to September 2022. We measure the channeling of war-related messages through user reviews on Tripadvisor and Google Maps. Our analysis of the content posted on these services reveals that users leveraged these platforms to seek and exchange humanitarian and travel advice, but also to disseminate disinformation and polarized messages. Finally, we analyze the response of platforms in terms of content moderation and their impact.

1 Introduction

With the beginning of the full-scale Russian invasion of Ukraine on February 24, 2022, Roskomnadzor—Russia's media regulator—implemented information control measures to block social media platforms like Facebook and Twitter, and news websites like the BBC (Meaker 2022b). These measures were seen not only as an attempt to stop the dissemination within Russia of any information not provided by official sources, but also as retaliation for the removal of Twitter and Facebook accounts allegedly belonging to two pro-Russian disinformation groups (Silverman and Kao 2022; Collins and Kent 2022) and the EU bans on Russia's news outlets Russia Today (RT) and Sputnik (Council of the EU 2022). The Open Observatory of Network Interference (OONI) confirmed the deployment of censorship mechanisms by Russian Internet Service Providers (ISPs) from the beginning of the 2022 Russian invasion in a report dated March 7, 2022 (Xynou and Filastò 2022).

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Information controls are frequent in times of war, and so are evasive maneuvers to bypass them. Russia's censorship efforts were answered with some inventive proposals. On February 28, 2022, an account presumably affiliated with the Anonymous movement suggested employing online user reviews in restaurants and other businesses located via Google Maps to deliver war-related information to the Russian population (@YourAnonNews 2022) to protest against the invasion. Tinder, Tripadvisor, and Telegram were also targeted as means to reaching out to the Russian population, thus bypassing the strict media control implemented by Roskomnadzor (Meaker 2022a). On March 4, 2022, the Squad303 group offered the possibility to target millions of Russian citizens with SMS via the *1920.in* site (Squad303 2022). This service was later extended to send emails, WhatsApp, and Viber messages. Some prominent online service providers responded to these campaigns by actively removing war-related content from their platforms. Google and Tripadvisor placed restrictions on reviews of Russian business, and Google Maps soon stopped accepting new reviews for places located in Russia, Ukraine, and Belarus. They argued that such reviews violate their platform policies (Hamilton 2022; Kaufer 2022; Deighton 2022).

The unconventional use of online services as side channels to bypass Russian information controls on the web was anecdotally echoed by the media (Squad303 2022). Yet, there is no quantitative or qualitative assessment of the user involvement, intentions, and intensity of these campaigns, nor the response by platform operators to moderate content. Motivated by this research gap, in this paper we attempt to shed light on how two popular platforms (namely, Google Maps and Tripadvisor) were used during the Russia-Ukraine war to disseminate war-related content. Particularly, we aim to provide answers to the following research questions:

- **RQ1.** How prevalent was the use of platforms such as Google Maps and Tripadvisor to disseminate war-related content and what topics were discussed?

- **RQ2.** What are the user intents and motives when sharing war-related content on online platforms?
- **RQ3.** How do online platforms react to the influx of war-related content on their site and how do they moderate such content?

To answer the above research questions, we performed a large-scale crawl of Google Maps and Tripadvisor, collecting a set of 2.2M posts shared between March 2022 and September 2022. We then perform a mixed-methods analysis to identify, measure, and characterize war-related content shared on Google Maps and Tripadvisor, assess user motives and intents when sharing war-related content, as well as quantify and characterize the platforms' reactions through the lens of moderation actions performed on war-related content.

Our key findings are:

- We study changes in the volume of reviews and moderated content in Google Maps and Tripadvisor (§5.1). Our analysis indicates an increment of nearly 100× in new posts for Tripadvisor during the early days of the war.
- We leverage text-based analysis techniques to label reviews as related to war or not. We find evidence that content posted on these platforms is used to deliver political messages related to the war (*i.e.*, to counter misinformation/propaganda) to Russian audiences. Topic analysis of posted messages (§5.3) confirms a noticeable change in user discourse in Tripadvisor and Google Maps. This ratifies the use of these platforms as channels to disseminate war-related content.
- We conduct a qualitative assessment of the content (§6) and find that messages can be grouped in four main categories: (dis)information campaigns, humanitarian help, travel advice, and polarized/hate speech. We leverage network analysis techniques to find evidence of organized campaigns to disseminate slogans (§6.2).
- Finally, we study the reaction of service operators to control or remove war-related information (§7). We find that Tripadvisor monitors and, if needed, removes war-related content not abiding by their Terms of Service (*e.g.*, hate speech) from their forums in less than two days on average. Also, we observe that the number of reviews in Google Maps was reduced up to an average of 8 reviews per day in conflict areas.

2 Background

State censorship during times of military conflicts and in oppressive regimes has been subject of analysis by academics for a long time (Price 1942; Morgans 2017; Pearce et al. 2017; Marczak et al. 2015; Niaki et al. 2020). The 2014 Russia-Ukraine conflict offered a valuable case study of Russia's information war and information control strategy. Russia started offensive cyber-operations against Ukraine not later than 2009 as a part of a broader war campaign against NATO and EU countries (Unwala and Ghori 2016). In 2014, information war operations intensified against Ukraine (Volkova and Bell 2016). While initially aiming at spreading misinformation and propaganda, with the start of a full-scale attack on Ukraine in 2022, Russian authorities boosted media control by blocking free access to the Internet. Table 1 provides a sample list of sites that were blocked soon after the invasion started, according to OONI's web connectivity public data (Open Observatory of Network Interference 2022). Sites such as *bbc.com* and *facebook.com* were blocked on March 4, 2022, while others such as *twitter.com* were censored just two days after the beginning of the armed conflict. Independent Russian news channels like *currenttime.tv* or *tvrain.ru*, and censorship evasion tools and VPN services were also blocked in late February 2022. Yet, popular Russian social networks like VKontakte (VK) and western ones such as Instagram and YouTube remained accessible to limit collateral damage. Some of them were gradually blocked in Russian ISPs as the conflict evolved (Troianovski 2022). This was done in fear of civil protests against the war, preceded by repression and mass arrests on March 4, 2022 (Shevchenko 2022). In fact, words such as *"war"* and *"invasion"* were officially banned in Russia's media (Troianovski 2022). The list of blocked websites extended to other news sites as the conflict developed, including Deutsche Welle, Radio Free Europe/Radio Liberty, and Voice of America. A February 2023 report by OONI provides a detailed list of censored topics, services, and websites that extend beyond media sites, including podcasts, gaming websites, LGBTQ+ related content, anime and cartoons, music and lyrics, and anti-censorship tools (OONI 2023).

The intensification of information controls by Russian authorities motivated the use of creative and novel non-blocked side-channels. A Twitter campaign led by the group Anonymous provided information and called for action in Google Maps (@YourAnonNews 2022), with a tweet providing possible message templates to spread war-related information over Russian places. By October 2022, this tweet had been reposted more than 27k times, with more than 76k likes. As a result, these websites soon became a niche for spreading information about the war. In fact, on March 2, 2022, Tripadvisor's CEO recommended the use of Ukraine and Russia forums to *"enable users to share information"* about the situation in the country (Kaufer 2022). However, Tripadvisor staff soon posted messages for certain Russian places indicating that reviews were disabled due to a high volume of war-related content, and that users should use the forums to inform about available travel options within Ukraine instead (Deighton 2022). Since then, travel forums for Russia and Ukraine turned into a platform for discussion about the war situation.

3 Related Work

Prior work extensively analyzed Internet censorship and content moderation on the Web (Nourin et al. 2023; Sundara Raman et al. 2020; Ensafi et al. 2015; Marczak et al. 2015; Aguerri, Santisteban, and Miró-Llinares 2022), as well as anti-censorship evasion approaches(Bock et al. 2019; Clayton, Murdoch, and Watson 2006; Tschantz et al. 2016; Tran, Bock, and Levin 2023). Howard, Agarwal, and Hussain (2011) studied 566 incidents from 1995 to 2011 where different countries shut off social media at politically sensi-

Category	Website	Start of censorship
Social media	twitter.com	2022-02-26
	facebook.com	2022-03-04
	vk.com	not blocked
	youtube.com	not blocked
International media outlets	dw.com	2022-03-04
	bbc.com	2022-03-04
	nbc.com	not blocked
	nytimes.com	not blocked
	theguardian.com	not blocked
Independent media outlets	interfax.ru	2022-02-26
	currenttime.tv	2022-02-28
	tvrain.ru	2022-03-02
Ukranian media outlets	glavcom.ua	before 2022
	glavnoe.ua	before 2022
	maidan.org.ua	before 2022
	qha.com.ua	before 2022
	hromadske.ua	2022-02-08
	24tv.ua	2022-03-02
	atr.ua	2022-03-05
	1plus1.ua	2022-03-09
	5.ua	2022-03-18
	nr2.com.ua	2022-07-13

Table 1: Website censorship in Russia during the earliest weeks after the invasion, as detected by OONI.

tive times, such as military coups, mass demonstrations, and elections. Similarly, Dainotti et al. (2011) studied how Internet access was disrupted in Egypt and Libya during the first months of the 2011 Arab Spring.

Recent efforts studied the effects of the 2022 Russia-Ukraine war on the Internet from different angles. Xue et al. (2022) developed a novel approach to measure state-censorship on RuNet at different network levels. Jonker et al. (2022) focused on the infrastructure powering websites under Russian ccTLDs (*e.g.*, ".ru") and how the conflict has triggered a *repatriation* of such infrastructure towards national hosting providers. Ortwein, Bock, and Levin (2023) analyzed how Russian ISP censorship policies caused collateral damages on global Internet traffic transiting Russian ASes, mostly in central Asian ISPs. Finally, Xue et al. (2021) explored the Russian government's unprecedented use of throttling methods to censor Twitter in March 2021 using a centrally coordinated censorship model.

Yet, most prior work mainly focused on understanding and analyzing content moderation and the dissemination of misinformation on traditional social networking sites like Twitter and Facebook. Both Chen and Ferrara (2023); Pohl et al. (2023) compiled and released large-scale datasets of war-related Twitter activity, including misinformation dissemination, to foster further research and historical reasons. Hanley et al. studied the influence of Russian propaganda websites on other platforms during the conflict, *i.e.*, political subreddits (Hanley, Kumar, and Durumeric 2022) and Telegram channels (Hanley and Durumeric 2023). Aguerri, Santisteban, and Miró-Llinares (2022) studied the application of warning labels on Russian state-sponsored accounts, finding that on Twitter, the application of these warning labels had a significant reduction in the reach of content. Pierri et al. (2023a) studied account creation and suspensions during the Russia-Ukraine war on Twitter; finding that most accounts with suspicious activity got suspended within a few days of their creation. Geissler et al. (2023) show the importance of moderating content from bot accounts on Twitter; they showed that bots played a substantial role in disseminating pro-Russia content on Twitter. Pierri et al. (2023b) study the spread of propaganda and misinformation on Twitter and Facebook; they found that during the first few months of the Russia-Ukraine war, only between 8% and 15% of Russian propaganda was removed/moderated.

Our study complements prior research by shedding light on how other non-social-network-related services, like Tripadvisor and Google Maps were used to evade state censorship and how these platforms reacted to and moderated an influx of war-related content within their platforms. In fact, our study helps to draw a richer picture of how online services were used during the war to circumvent state-level information controls.

4 Methodology

Due to anecdotal evidence suggesting that both Google Maps and Tripadvisor were being used to disseminate war-related information (Deighton 2022; Kaufer 2022), we focus our study on these platforms. We note that both services were available to citizens accessing them from Russian ISPs (Open Observatory of Network Interference 2022). This section describes our crawling efforts, the obtained datasets, and the data labeling pipeline used to distinguish between war-related and non-war-related content.

4.1 Data Collection

Tripadvisor. We crawl Ukraine and Russian travel forums in Tripadvisor for seven months (from March 12, 2022, to September 30, 2022), split into two periods. Travel forums are a section of the Tripadvisor website, separate from reviews, where users can ask other travellers questions about particular destinations and trip planning. As mentioned in §2, reviews in Russian places were disabled by Tripadvisor staff, and these forums were therefore used for discussion about war. First, we conduct regular crawls for a period of two months (from March 12, 2022, to May 12, 2022) harvesting all the posts made since the beginning of the war. As some posts were being removed by Tripadvisor due to infringement of its Terms of Service, the crawler checked for new posts every hour and collected any new content, thus allowing us to flag and measure removals. This crawling period allowed us to conduct online monitoring on the platform. We also conducted one single crawl to obtain pre-war posts since May 12, 2021. Due to technical limitations and a decrease in post volume over time, we decided to stop the data collection and resume it in September 2022 with a lower crawling frequency. Overall, the dataset contains 7,330 posts made in 1,319 different threads by 1,229 different users.

Dataset	Crawling period	Data period	Size
Tripadvisor	Mar 12, 2022 - Sep 30, 2022	May 12, 2021 - Sep 20, 2022	7,330 posts
Google Maps	Mar 4, 2022 - Jun 30, 2022	Jan 1, 2020 - Jun 30, 2022	2,200,368 reviews

Table 2: Datasets with their respective volume, crawling and data periods.

Google Maps. Using a purpose-built crawler, we harvested 2,200,368 reviews obtained from 122,826 locations in Russia. We started crawling these reviews on March 4, 2022. We fetch new reviews every 2 hours and update the list of places daily. Our Chromium-based instrumentation makes use of the "Nearby" search functionality offered by Google Maps to lists any places (*e.g.*, hotels, restaurants, museums) found in a given location or its vicinity. We feed the crawler with a seed formed by 321 predefined Russian towns (Wikipedia contributors 2022) from where to discover places. In the end, combining these two methods we covered 8,660 different towns. These reviews were posted by 1,164,002 unique users. We stopped crawling on June 30, 2022, because the activity on this platform had stopped, as we mention in §1.

4.2 Data Labeling

To label war-related content, we apply a combined approach using both quantitative and qualitative methods. We start with a manual analysis of a subset of messages (70% threads of conversations from Tripadvisor and 80% of reviews from Google Maps) posted during the first 2 months of the war. This allows us to get an overview of the most common topics being discussed, which we describe in detail in §7.

Though informative, such a time-consuming manual labeling process does not scale. To overcome this challenge, we first employed unsupervised topic mining techniques (Grootendorst 2022)—namely masked LM embedding, dimension reduction, and clustering—, which indeed identified topics related to war. However, the results were very specific and produced many false positives (FPs). Instead, we opted for an automated labeling approach based on a set of keywords obtained from our qualitative evaluation. In this process, we aim to reduce the number of FPs. The methodology we use is as follows:

1. We take a random sample of 3,200 messages (660 from Tripadvisor and 2,540 from Google Maps).
2. Three annotators, including one Russian native speaker, manually label them as "war-related" or "non-war-related." We consider a message war-related if it clearly and explicitly discusses aspects of the war, *e.g.*, giving advice to refugees, positioning themselves pro or against the invasion, or trying to bypass Russian censorship (we provide examples of war-related reviews and posts in §6.2).
3. We normalize all messages and we extract the most common keywords present in war-related messages. Our normalization pipeline consists on removing URLs, punctuation symbols, and any character that is not used in either the English, Cyrillic, or Polish alphabets.
4. We analyze all obtained keywords and classify them into war-like keywords, violence-like keywords, and other recurrent keywords in war-related posts (*e.g.*, regime or nuclear). We obtain an independent inter-agreement coincidence score of 92% in this stage, then discussing and classifying the words that were not consistently labeled into the same category.
5. We score each keyword with (i) 3 points up to a maximum of 6 points if it is war-like; (ii) 2p up to 4p if violence-like; and (iii) 1p up to 2p if, while frequently occurring, does not belong to any of the former.

The reason why we limit the maximum score per group of keywords is to prevent long messages, which are more likely to have more keywords, from gaining an unfair advantage. We also slightly increase the final score of messages with very few keywords (*e.g.*, *"Glory to Ukraine"* or *"Peace please!! Stop Putin"*), since otherwise these would have been misclassified for being short. Additionally, we use corrective keywords with negative weights to minimize the number of FPs. To allow for validation and reproducibility of our method, we provide an online artifact containing the list of keywords and weights (*i.e.*, scores) given to each of them.[1] This list was validated by eight English and four Russian/Ukrainian native speakers to discard false positives. We get a score for each message based on the number of keywords found and their scores. We empirically determined a classification threshold of 5, *i.e.*, we classify a message as war-related if it has a score ≥ 5, and non-war related otherwise. This threshold was fine-tuned after several iterations conducting manual validation on the posts.

We measure classification quality via a set of lightweight crowdsourced campaigns. Due to the predominant presence of the Russian language on Google Maps reviews, we employ (i) non-native speaking coders using machine translation, and (ii) verify the results with a second labeling pass from native-speaking expert coders. The campaigns were set up for each dataset by carefully sampling sets of 25 posts identified as non-/war related for the time before and after the beginning of the war. We also focus on non-deletions, resulting in 750 labels of which 94.7% were consistent between non-expert coders.

Accuracy Evaluation. Table 3 shows the evaluation results of the classifier. Our keyword-matching approach overall works surprisingly well at a precision of 0.97, with an F_1 score of 0.85. It can be observed that our method prioritizes precision, *i.e.*, more than 97% messages classified as war-related are indeed war-related. Also, we observe increased figures of false negatives (*i.e.*, lower recall), as expected. Thus, during our analysis, while we discuss non-war-related posts, we note that (some of) these might be war-related.

4.3 Limitations

Despite our best-effort data collection approach, we faced technical challenges while crawling. Google Maps soon

[1] Available at https://zenodo.org/records/10848304.

Dataset	Accuracy	Precision	Recall	F1
Tripadvisor	0.8523	0.9565	0.6735	0.7904
Google Maps	0.9231	0.9910	0.8594	0.9205
Total	0.8752	0.9773	0.7588	0.8543

Table 3: Accuracy metrics for the resulting content classifier.

disallowed reviews in Russian places at the beginning of March 2022 due to the high amount of off-topic messages (Deighton 2022). This caused a drop in the volume of new posts harvested by our crawler. Therefore, we cannot reliably determine which fraction of new messages would be focused on war-related content should the moderation policy had not been implemented, and we cannot determine the cause for the removals (see §7). In Tripadvisor's case, we collected data every hour for the first two months. At the beginning of May 2022, we experienced temporary bans on our crawler, which we did not attempt to circumvent due to ethical reasons (§8). We increased the frequency of our crawls to one per day. Afterwards, we observed a decrease in the number of war-related posts (see §5) and decided to stop the crawling and resume it in September 2022 to gain a longitudinal view of this phenomenon. Despite these interferences in the crawling period, we collect historical data written by users, and thus, this does not affect most of our analyses. However, we cannot precisely measure removals as discussed in §7. Finally, the use of manual data labeling could bias the correctness of the ground truth. Our process involves three annotators, one of them is a Russian native. We got a 92% of inter-agreement during the keyword classification. Additionally, we provide the list of keywords as an online artifact for reproducibility and validity checking.

5 RQ1: Prevalence of War-related Content

In this section we present our analysis on the quantification of war-related content on Tripadvisor and Google Maps. We focus on (i) the observed traffic shifts that correlate with the beginning of the war; (ii) the analysis of how much new content is related to the war; (iii) the change of topics being discussed; and (iv) the Russian places that are targeted in Google Maps with war-related content.

5.1 Changes in Volume

Figure 1 shows the number of posts or reviews that were published in each studied platform over time. For both Tripadvisor and Google Maps, we see a clear change in volume right after the beginning of the war on February 24, 2022 (denoted by a vertical black line), this change manifests differently across platforms in terms of intensity. In the case of Google Maps, there is a clear drop in the daily amount of published reviews, suggesting some kind of content moderation, which is consistent with Google's policy of not accepting new reviews for places located in Russia, Ukraine, and Belarus. For Tripadvisor, instead, we see a slight increase in the number of posts due to a larger volume of war-related posts in these travel forums, which was indeed allowed by forum administrators (see §5.2).

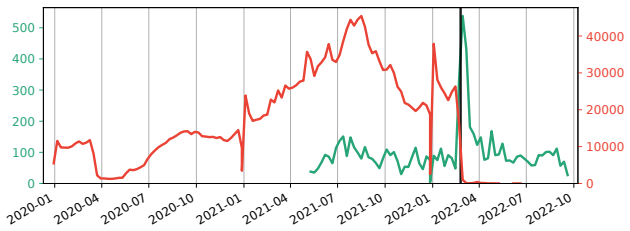


Figure 1: Volume of Tripadvisor posts (green, left axis) and Google Maps reviews (red, right axis). The vertical black line represents the beginning of the invasion. Volume dips in Google Maps are related to data collection issues (see §4.3).

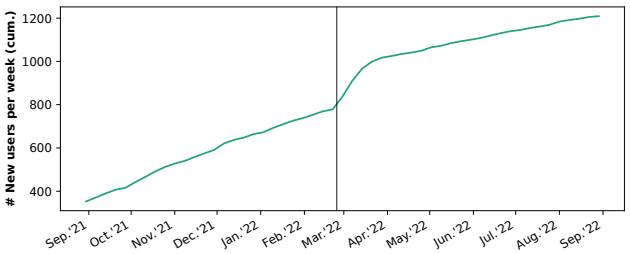
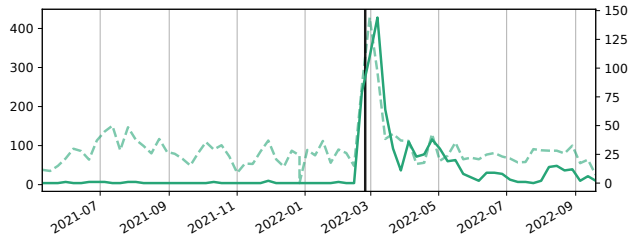


Figure 2: Cumulative number of new users actively participating in Tripadvisor forums. The vertical black line represents the beginning of the invasion.

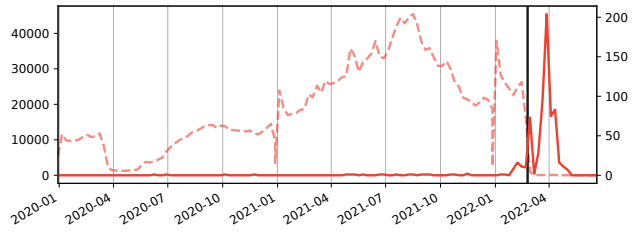
Figure 2 shows the evolution of users interacting for the first time within the Tripadvisor forums. In September'21, there were 353 users that had mad at least one post in the previous 4 months (our dataset spans from May'21). Since then, it can be observed that there is a constant increase of new members engaging in the forums, at a rate that fluctuates between 10 and 25 per week, with a peak of 30 new users in a week right before Christmas. However, in the 3 weeks following the start of the conflict, we note an anomalous number of users started for the first time to post in the forums (*i.e.*, 59 new users in the last week of February, 73 new users in the first week of March, and 57 in the second week of March). In total, 450 users in our dataset ($\sim$35%) started their interaction in the forum after the conflict. These users highly engaged in war-related discussions, mostly those posting in the earlier days after the conflict. Concretely, we observe that 20% of the new users posted more war-related than not-war related posts. This percentage increases to 32% in posts made in the three first weeks of March'22. We further analyse the difference between war-related and not-war related discussions in the following section.

5.2 War-related Content

Using the methodology from §4.2, we analyze the content of the posted messages to determine how much of the observed traffic volume increase can be attributed to war-related content. Figure 3 shows the weekly rate of war and non-war-related messages in both Tripadvisor and Google Maps. In the previous subsection, we noted an increased amount of

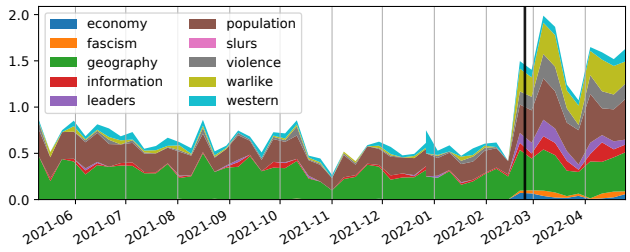


(a) Tripadvisor

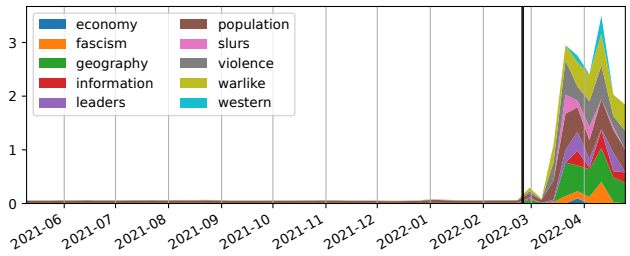


(b) Google Maps

Figure 3: Volume of weekly published posts and reviews per dataset. Dashed lines show non-war messages (left axis).



(a) Tripadvisor



(b) Google Maps

Figure 4: Evolution of the ratio of topics found in posts and reviews per week. The bold vertical line indicates the beginning of the Russian invasion.

activity on Tripadvisor. Now, we observe that this increase was mostly due to war-related content. We note that, due to the conversational nature of Tripadvisor's forum threads, some messages, even if not classified as war-related, were replies to previous war-related messages. Also, during the first two months right after the war began, we see a similar pattern for both war and non-war-related content, whereas from early May 2022, the number of weekly war-related content decreased, observing weeks with less than 10 messages related to war. In the case of Google Maps, our volume analysis showed a steep drop in the number of messages, which is attributed to the active removal (first) and complete blocking (afterwards) by Google of reviews in places from Russia. An interesting pattern is that, even with such a drop, we observe that most of the messages posted after February 24, 2022, in Google Maps were war-related, with a peak of nearly 200 messages in the first week of March 2022. This coincides with the beginning of our data collection. Our results confirm that Tripadvisor and Google Maps were used as effective side channels to avoid state information controls and to communicate war-related content, possibly because of the blocking of other platforms.

5.3 Topics

We perform an analysis of the topics mentioned in Tripadvisor posts and Google Maps reviews to understand the difference in discourse before and after the war. We determine the topics of a message based on the keywords it contains. We manually group all the war-related keywords into different classes to generate a list of 10 topics (see artifact listing keywords for the mappings). This process was conducted by two expert coders.

Figure 4 shows the change over time in the ratio of topics for the mentioned platforms. The number of messages is normalized to account for variations in volume. In the case of Tripadvisor, we observe that topics like *"warlike"*, *"violence"*, or *"fascism"* gained popularity immediately after the beginning of the war. Similarly, Google Maps reviews posted before the war do not tend to mention any of the identified topics and focus almost exclusively on user recommendations. We observe a radical change in users' discourse shortly after March 2022, which correlates with the addition of war-related reviews to the platform.

5.4 Targeted Places in Google Maps

We studied which cities and regions tend to be the objective of war-related reviews. Unfortunately, this analysis is only possible on Google Maps, as it is the only data source associated with locations in Russia. We observe that 108 towns out of a total of 8,660 found on Google Maps have at least one war-related review. These towns cover significant areas of Russia, including cities like Krasnoyarsk in Siberia, Vladikavkaz in the Caucasus, and Moscow in the central district (Figure 5). We observe that half of the total war-related reviews concentrate on Moscow and St. Petersburg, with 37% and 15% respectively. However, the ratio between war and non-war reviews on cities close to the Ukrainian border or around Moscow is higher than in other regions (red dots in Figure 5). Some examples are Belgorod (frontier city) with 0.6% of all the reviews being war-related ones or Ryazan (southeast of Moscow) with 0.25%, while in cities such as St. Petersburg drops to 0.06%.

Figure 5: Ratio of war-related and non war-related reviews per Russian town (in red), non-war-related reviews (in green) and both (in black).

Takeaway. We observe substantial changes in the number of daily posts and reviews since the beginning of the war. These correlate with the blocking of major social platforms and news sites in Russia, and with a call by activist groups to reach out to the Russian population. Our automatic content labeling reveals that the traffic increase is mostly due to war-related content, and topic analysis confirms a noticeable change in user discourse. These findings ratify that Tripadvisor and Google Maps were used as side channels to disseminate war-related information to Russian citizens, possibly because of state-level blocking of other platforms.

6 RQ2: Intent and Purpose of Content

Section 5 offered a high-level overview of the main war-related topics discussed using quantitative analysis. Based on the extracted topics, in this section, we perform an in-depth qualitative analysis of the four identified main topics. Also, we present evidence of orchestrated content dissemination in the platforms.

6.1 Topics

We next describe in further detail the four main topics identified. Table 4 contains illustrative examples of war-related posts and reviews we found across the studied platforms. We provide a general discussion on the topics, showing some of the sample messages verbatim.

(Dis)information and Censorship Bypass. Our analysis suggests that the two analyzed platforms have been used as a channel to fight against Russia's information war because they were not blocked by Russian authorities—though, as shown in §7, their content was moderated by administrators. Thus, most of the reviews found in Google Maps and the posts from Tripadvisor seek to inform Russian citizens about the war (Table 4).

These messages are often orchestrated as organized campaigns, as discussed in §6.2. For example, GM02 (Table 4) appears in 52 different reviews in Google Maps, written both in English and Russian.

As part of this information war, one Tripadvisor user claimed on the earliest days that *"[Tripadvisor] has been the target of a team of Kremlin trolls who are paid to try to control the narrative on here around the current state of Russia."* In this post, the user asked the community for *"comments correcting falsehoods"* and pledge for *"cooperation between internet users and observers who are able to expose and compromise trolls,"* also asking for platform moderation due to the aggressive speech *"extensively employed by trolls."* This discourse is often repeated in the platforms, and the intent is to inform Russian citizens about the actual events happening:

> Much Fake news is spreading in the official Russian Press Media. [...] Be careful with the Russian agents that will be paid by the Putin government, some of them are trolls of the KGB. They are rats sewer who help Putin spread false information, be careful.

We also observe messages assisting Russians to bypass censorship. For example, a message with instructions on *"how to get around restrictions on BBC services in Russia"* was posted by the same user in 8 threads simultaneously. We identified several pro-Russian users who also engaged in these efforts. For instance, one user in Tripadvisor claiming that *"any message that does not fit into your picture of the world is automatically branded with the phrases 'your garbage', 'nonsense', 'propaganda',"* or a review in Google Maps, which was repeated in 22 different places, that claims that *"Moscow already knows that the dill [a Russian language ethnic slur to refer to Ukrainians] will hit"* and *"we are left here like a live shield and do not give information, they want bombs for the picture to fall on us"*, finally claiming that this is *"information from a very reliable person in the army"*.

Overall, propaganda and disinformation messages were one of the main topics of discussion observed.

Humanitarian Help. Another frequent topic of discussion is humanitarian help and advice. Due to these platforms being accessible worldwide, people decided to use them to inform and raise awareness about the humanitarian needs of Ukrainian refugees. Users both asked for humanitarian help and also provided advice to Ukrainian refugees in various aspects. These include cheap or free options to get out of the country (*"Free travel on [redacted] flights for Ukrainians"*) or free accommodation in hotels and apartments abroad (*"An independent platform connecting Ukrainian refugees with hosts who can offer free housing"*). Also, various users promote Non-Governmental Organizations that are helping refugees, such as Razom or Redcross (*"Please ask your rugby club to donate to Razom for Ukraine."*), and informing on charities helping kids (*"A friend is at present involved in getting orphaned kids out of Dnipro [...] The charity, Edinburgh based Dnipro Kids' is hoping to get many more children to safety"*).

Polarization. Besides the main topics discussed, we observe one common pattern common to both studied platforms: there is a high level of polarization. There is a clear po-

ID	Content
TA01	PUTIN is killing is killing innocent Ukrainians; men women, children and their pets. Please help.
TA02	Here's a reminder of how to get around restrictions on BBC services in Russia: Download the Psiphon app from the AppStore or Google Play Store - Look for the dedicated BBC site on the Tor Browser which can be found using this URL. [...]
TA03	[NICKNAME], I don't like to speak about political opinions, Kills thousand of people in Ukrania and thousand of Russian young soldiers are not acceptable in this century, I can not speak about tourism and forget the atrocities in this crazy war. [...]
TA04	Since you want to turn this thread into politics you might want to check and find out that Ukraine killed 14000+ innocent civilians -Ethnic Russians in Donbass, Eastern Ukraine in 2014-2022 and it keeps shelling this region to cause them to die or leave. Why everyone thinks this was/is OK? And noone talks about it.
GM01	*Your children, relatives are being mobilized. Do you want to find them on our lists? We don't. How many of your soldiers died? - THOUSAND?! Don't believe it! - Already over 17k dead Russian soldiers in Ukraine!!! - All Actual Information about the Fallen Russian Armed Forces in Ukraine is here: - https://t.me/rf200_now - https://youtube.com/c/VolodymyrZolkin - We are a humanitarian project to inform the relatives of those killed about their fate WATCH, LISTEN, THINK, ANALYZE - IT IS IMPOSSIBLE. YOUR AUTHORITIES, THE MEDIA LIE TO YOU...*
GM02	BE AWARE!!!! MURDERERS! THEY (RUSSIANS) ARE KILLING KIDS, SENIORS AND WOMEN IN MARIUPOL AND MANY OTHER CITIES IN UKRAINE!!!!! SHAME!!! DON'T BUY THEIR ROTTEN FOOD!!! VNIMANIE!!! UBIJCY! ONI (RAShISTY) UBIVAJuT DETEJ, STARIKOV I ZhENShhIN V MARIUPOLE I MNOGIH DRUGIH GORODAH UKRAINY!!!!! NE POKUPAJTE IH GNILUJu EDU, IH RUKI V KROVI!!!
GM03	I am an American and I am writing about Ukraine. If the only news you receive in Russia is from your state media, you are being told lies. It is critical that you know the truth about the Russian invasion of Ukraine. The Russian pretext of restoring peace to Ukraine and removing Nazis from the government is a lie. [...]
GM04	*Ukrops have already been given many offensive weapons concealed!!!! A strike on Belgorod is being prepared. So far, there is no exact date when it will happen. Moscow already knows that Ukrop will strike, they are already digging trenches in Shebekino!!!!* [...]

Table 4: Example posts and reviews. Translated texts appear in *italic*.

sitioning of pro-war and anti-war (predominantly the second). This leads to various instances of hate speech that we will not reproduce in this study. As discussed in §7, Tripadvisor actively banned such content and removed (at least) 191 posts, including entire threads, when these turned into somehow aggressive discussions between members. Also, as we discussed in §5.3, we observe a large increase in messages that include violent and offensive words (*e.g.*, *"killer"*, *"troll"*, *"monster"*, or *"nazi"*) since the beginning of the war. Therefore, we observe how platforms that are initially intended for licit purposes (*e.g.*, providing travel advice or offering customer reviews) have been radicalized to an unprecedented extent.

Travel advice. The final common topic which we have observed, mostly in Tripadvisor, is about assistance for travelling in and out of Russia and Ukraine. Even if this is expected—this was the original purpose of enabling war-related content in the forums—we observe that sometimes the conversations are easily polarized towards one side. For example, one user informed about the potential danger in traveling between Moldova and Odessa (*"I recommend refraining from traveling in this direction, Zelensky's soldiers will be happy to use you for another provocation"*). However, these are isolated cases, and most of the advice is done for refugees willing to leave the country (*"Tomorrow will be a bus (for free) in Lviv. Bus will be waiting for children and women next to main train station in Lwiw"*).

6.2 Orchestrated Content Dissemination

We identified instances of the same slogans being posted by different users within and across platforms. Given the public calls made by activist groups to inform Russian citizens about ongoing war events, we analyzed if the posted messages contain organized campaigns (*i.e.*, spamming the same text) in addition to personal or individual statements. We identify clusters of identical posts on Tripadvisor and Google Maps with patriotic statements or informing Russian citizens about the war, pledging for an end. For Google Maps, we identify 8 posts, each being replayed more than 10 times, totaling 188 instances. The same holds for Tripadvisor, though the amount in this case is smaller: 7 posts posted 17 times. As discussed before, our qualitative look into the contents provides a wide spectrum of pro-Russian and pro-Ukrainian messages, whereas others primarily pledge to stop the war.

Figure 6 shows the top largest campaigns found in Google Maps by volume of messages, alongside the users that took part in them. We provide examples in Table 4 for the campaign nodes that have a label. Campaigns are almost exclusively posted by a single user that distributes their messages across different places. Overall, we do not observe any patterns of viral spread of content that is more prevalent on traditional social media platforms like Twitter (Lerman and Ghosh 2010) and Facebook (Friggeri et al. 2014), which is likely because of the differences in the platform affordances of Google Maps and Tripadvisor compared to social media platforms.

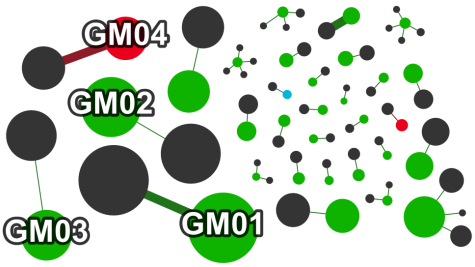


Figure 6: Network graph of the largest campaigns observed in Google Maps. Nodes in red show pro-war campaigns; in green against-war campaigns; and in blue neutral or inconclusive. Gray nodes represent users.

Takeaway. Tripadvisor and Google Maps were used to disseminate messages about: (i) (dis)information targeting Russian citizens, (ii) humanitarian aid, (iii) hate speech, and (iv) travel advice. We found evidence of information campaigns supported by multiple users who contributed to disseminating both pro-Ukraine and pro-Russia messages and slogans.

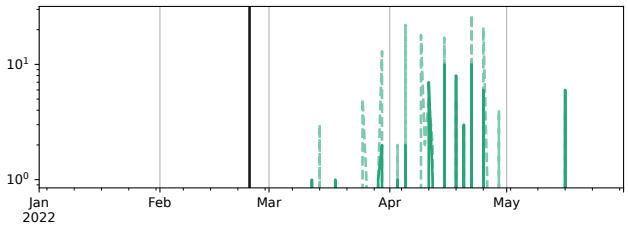
7 RQ3: Platform Moderation

Platform operators reacted differently to the use of their services as side channels for disseminating war-related information. We next discuss content moderation—or lack thereof—in both platforms analyzed in this study. For reference, we first analyze the removal of posts by administrators, as observed during our daily crawls.

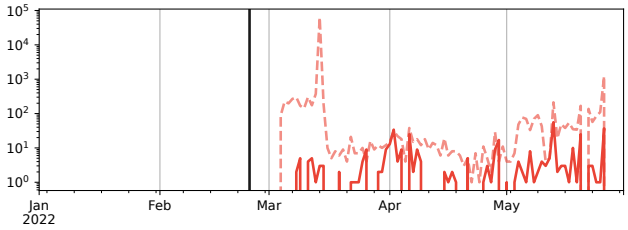
We restrict this analysis to the period when we conducted hourly crawls, *i.e.*, during the first two months of the war. For Google Maps, we estimate removals based on the latest time our crawler recorded a review. This yields good results as we daily crawl all reviews (new and previously existing ones) for all monitored places. In the case of Tripadvisor, a removed post is replaced by a placeholder message from the admins indicating the reason for removal. However, the timestamp of this replacement is not provided. For Google Maps, we estimate the removal date as the first timestamp where we observe that a message was replaced by a placeholder message. We conduct hourly crawls for the analyzed period so that we can measure platform moderation with one-hour precision.

Figure 7 shows the number of war and non-war deleted posts for both Google Maps and Tripadvisor. We observe that the volume and the frequency of removals are higher for Google Maps than Tripadvisor, as opposed to the new entries where we observe similar patterns. This suggests that both sites implemented different content moderation policies during the war time. We next discuss each platform's moderation and reasons for removals when available.

Tripadvisor. During our analysis, we noted that various posts and entire threads were removed by forum administrators. Not all of the messages in the removed threads were



(a) Tripadvisor



(b) Google Maps

Figure 7: Number of daily removed posts labeled as war-related (solid line) and non-war-related (dashed line).

about the war, which explains the larger volume of non-war posts compared to war-related posts. The reasons for these removals are analyzed in Table 6, which shows the number of posts and the reason provided in the placeholder message left. We narrow down our analysis on messages removed after the war started. Note that up to 121 posts were removed at the author's request according to the metadata offered by the platform. Hate speech and harassment are the two more prevalent categories when the platform removes posts. For the 34 threads (215 posts) that have been removed completely, we ignore the reasons for such removals. Out of these, 13 ($\sim$38%) have a unique post, and 4 ($\sim$12%) only contain one reply. Meanwhile, 7 threads ($\sim$20%) have more than 13 replies. We confirm through manual inspection that the reasons for thread removal typically fall into two categories: (i) the conversation in the threads completely deviates from its original purpose (*e.g.*, provide objective information about the war) towards political discussions or even hate speech; or (ii), the thread is initiated with the sole purpose of propaganda or another advertisement, and it is removed quickly, sometimes even before it gets any reply. Table 5 shows the lifespan of the 69 items (*i.e.*, a post or the entire thread) removed by administrators in the period when we conducted hourly crawls (between March 12, 2022, and May 12, 2022). Posts, in general, are removed faster than entire threads, but we also observe that those threads without replies (*i.e.*, containing only the OP message) are removed as quickly as regular posts. This confirms that active platform moderation is in place and that much of the content removal is linked to the war.

Google Maps. We find evidence suggesting that Google Maps performs platform moderation as war-related reviews have a much shorter lifespan than the rest, typically lasting less than 50 days as opposed to those until the end of

	Threads	Posts	All
Total	31	38	69
Mean	7d 23h 15m	2d 16h 16m	5d 1h 19m
Median	2d 16h 44m	1d 14h 32m	2d 8h 28m
Stdev	12d 8h 42m 40s	2d 14h 0m	8d 20h 5m

Table 5: Lifespan of content removed in Tripadvisor.

Reason	Posts
Entire thread was removed	215
Off-topic chat	147
Removed by author	121
Harassment to other users	102
Hate speech or inappropriate language	89
Self-promotional advertising	31
Not written in English	14
Copyright infringement	6
Multi-account detected	2
Total	727

Table 6: Reasons for content removals in Tripadvisor.

our crawling. Shortly after the beginning of the war, Google Maps temporarily suspended posting new reviews on Russian places to prevent the generation of content that violates company policies (Deighton 2022). This led to a drastic reduction in daily published reviews, as shown in Figure 1. It also led to a massive removal of non-war-related reviews on March 15 (Figure 7b). However, we did not find any apparent correlation between these reviews and the war, *i.e.*, based on their contents, geolocation or time of publication. While war-related content is not explicitly prohibited in Google Maps, Google alleged that these reviews were considered "off-topic," a category prohibited in Google Maps, justifying their temporal suspension (Google Help 2022). According to our data, this temporary banning was still active by May 2022—just 8 posted daily reviews. Nevertheless, we find 18 war-related reviews that bypassed Google Maps' moderation.

Takeaway. There is evidence of both platform operators actively removing messages that contain war-related content. The most common reasons for content removal include off-topic message, harassment and hate speech, or more generally "content that violates ToS." The lifespan of moderated content ranges from a few hours to several weeks, with some messages escaping moderation.

8 Discussion and Conclusion

This paper studies empirically how online platforms such as Tripadvisor and Google Maps were used to bypass Russian state-level censorship during the 2022 Ukrainian war and the platform efforts to moderate such content. Russian censorship did not block these two online platforms, presumably because they are not social networks or news sites. Using a dataset collected during the first weeks of the war, we observe the shifts in the pattern of daily user post volume and removals, as well as duplicated content suggesting intentional and organized campaigns to disseminate such information. Our content analysis using both quantitative and qualitative methods confirms that there is indeed a peak in war-related narratives in the reviews posted by users on these platforms.

The unconventional use of place and business reviews as side channels to evade censorship forced platforms to apply content moderation policies. In the case of Tripadvisor, our analysis suggests that administrators perform intensive content moderation. However, they allow (and indeed encourage users) to discuss and inform about the war, mostly to provide information about safe traveling in and out of Russia and Ukraine. In the case of Google Maps, there is also evidence of bulk removals leading to a temporary suspension of the reviewing activity that extends up to the time of this writing.

Overall, our study reveals how two unblocked online platforms were leveraged to circumvent state-wide information controls. Our findings provide new insights into human behavior displacement in times of crisis and raises the question of the role of the Internet in these periods and the effectiveness of Internet censorship.

Despite the limitations faced during the data collection and analysis (see §4.3), we believe that our work has important implications for various stakeholders, including researchers, policymakers, and operators of Web services, who are interested in understanding the Web and its potential impact on society. For the research community, our study provides important insights into how two seemingly innocuous and non-war-related Web services were used to disseminate war-related content and essentially circumvent censorship during a crucial real-world event. Our work assists in raising awareness about these social phenomena and in further understanding the diverse Web ecosystem. At the same time, our work highlights the need to perform analyses involving multiple platforms as it provides a more comprehensive view of such social phenomena through the lens of the Web.

Also, our work can assist operators of Web services in understanding how their services can be the recipients of a large volume of war-related or otherwise irrelevant/disrupting to their service content. Particularly, our analysis of content moderation highlights the need for effective and timely content moderation of such content (given that we find polarized and potentially harmful content). At the same time, it emphasizes the need for content moderators who are experts in the subject (in this case, the Russia-Ukraine war) to effectively perform content moderation. Taken altogether, operators of Web services can benefit from our work, as it can potentially assist them in improving/refining their platform's governance (*e.g.*, improving their terms of service and policies of what is allowed/disallowed), as well as improving their content moderation procedures to account for social phenomena that are not expecting (like the influx of war-related content due to

the Russia-Ukraine war).

Finally, our work is of great interest to policymakers and provides evidence about potential harms that may arise from the use of seemingly innocuous services like Tripadvisor and Google Maps. For instance, our study sheds light on the war-related content disseminated within these platforms, finding polarized and offensive content that can adversely affect end-users exposed to such content. Given that policymakers have recently released the Digital Services Act (European Commission 2023), which, among others, states that Web services will be held accountable for potential harms that may arise from the use of their services, we believe that our work provides some evidence of how such online harms may appear in services that are generally general-purpose without serious concerns about online harms.

Ethics and Broader Perspective

In this section, we discuss our ethical considerations when conducting our data collection, when analyzing and presenting the results, and the broader perspective of this work.

Data Collection. Data was collected by automatic means (crawlers) using standard ethical guidelines (Kenneally and Dittrich 2012; Fiesler, Beard, and Keegan 2020). We used sequential crawlers and did not produce more traffic to the servers than a human user would do. The dataset used in this study might contain sensitive data, including Personally Identifiable Information (PII) like usernames. Unfortunately, it is impossible to obtain informed consent from all users. According to the British Society of Criminology Statement on Ethics, we do not require informed consent from the participants because the dataset (i) is publicly accessible; and (ii) will be used for research on collective behavior without aiming to identify particular members. However, as this study involves analyzing content generated by human subjects, it does require ethical review. We applied and obtained approval from our IRB using the following research protocol:

1. The data is stored in our secured private servers with strict access control mechanisms.
2. We do not store users' personal identifiers, which are replaced by unique hashes obtained from their usernames. No efforts to deanonymize the data are carried out.
3. We do not store or visualize any images or other multimedia material posted on the crawled platforms.
4. The data will be used only for the purpose of this research.
5. Because of the sensitive nature of the dataset, it will not be made publicly available. The dataset might be shared with other researchers using a controlled sharing policy, *i.e.*, a policy that restricts uses and complies with the present research protocol.

Our data collection methodology might not be in full accordance with the ToS of the two platforms analyzed, which restrict content scrapping either totally or for specific purposes. Nevertheless, we consider that undertaking this study presents a reasonable risk-benefit trade-off, as the advantages derived from understanding the phenomena under investigation outweigh the potential harms. Our approach to data collection, storage and sharing is particularly sensible to this consideration, and we implemented precautions to mitigate risks and potential harm.

Data Analysis and Presentation. When conducting our analysis and visualizing the results we follow ethical standards and recommendations (Rivers and Lewis 2014). Particularly, we do not attempt to track users across websites, we protect the anonymity of the users, we respect the context that the content was shared, and we report most of the results on aggregate. To better articulate the main points of the paper, in some cases, we report examples of posts; we ensured that the quoted text can not lead to the original posts in Google Maps or Tripadvisor, hence linking it to the user that posted the content (*e.g.*, when someone tries to use search engines to find the original post from the quoted text).

Broader Perspective. Our work benefits the research community by providing insights and raising awareness about innovative ways of leveraging online platforms like Google Maps and Tripadvisor to bypass informational censorship during important real-world events like the Russia-Ukraine war. We believe that our study will positively inform the research community about this topic and improve our current understanding of how the Web ecosystem works. Our work emphasizes the need to look at and analyze the Web ecosystem through the lens of multiple online platforms since this study shows that people can leverage online platforms, not necessarily related to the content they aim to share, to disseminate information and bypass censorship controls. To conclude, we do not foresee any potential harm arising from our study. We believe that our study and the presented results assist in raising awareness and quantifying these online phenomena and it is highly unlikely that our study will inspire adversaries who aim to bypass censorship.

Acknowledgments

This research was partially supported by the Spanish AEI grants CYCAD (PID2022-140126OB-I00) and CIARRES (TED2021-132170A-I00), the EU's Horizon 2020 R&I Programme grant TRUST aWARE (101021377), and by the European Research Council (ERC) under Starting Grant ResolutioNet (ERC-StG-679158). José Miguel Moreno is supported by the Spanish Ministry of Science and Innovation with a FPI Predoctoral Grant (PRE2020-094224). Narseo Vallina-Rodriguez has been appointed as a 2020 Ramon y Cajal Fellow (RYC2020-030316-I).

The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of any of the funding agencies.

References

Aguerri, J.; Santisteban, M.; and Miró-Llinares, F. 2022. The fight against disinformation and its consequences: Measuring the impact of "Russia state-affiliated media" on Twitter. https://doi.org/10.31235/osf.io/b4qxt.

Bock, K.; Hughey, G.; Qiang, X.; and Levin, D. 2019. Geneva: Evolving censorship evasion strategies. In *CCS*, 2199–2214.

Chen, E.; and Ferrara, E. 2023. Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia. In *ICWSM*, volume 17, 1006–1013.

Clayton, R.; Murdoch, S. J.; and Watson, R. N. 2006. Ignoring the great firewall of china. In *International Workshop on Privacy Enhancing Technologies*, 20–35. Springer.

Collins, B.; and Kent, J. L. 2022. Facebook, Twitter remove disinformation accounts targeting Ukrainians. https://www.nbcnews.com/tech/internet/facebook-twitter-remove-disinformation-accounts-targeting-ukrainians-rcna17880. Accessed: 2022-05-04.

Council of the EU. 2022. EU imposes sanctions on state-owned outlets RT/Russia Today and Sputnik's broadcasting in the EU. https://europa.eu/!fGpJKJ. Accessed: 2022-05-16.

Dainotti, A.; Squarcella, C.; Aben, E.; Claffy, K. C.; Chiesa, M.; Russo, M.; and Pescapé, A. 2011. Analysis of Country-Wide Internet Outages Caused by Censorship. In *IMC*, 1–18.

Deighton, K. 2022. Tripadvisor, Google Maps Suspend Reviews of Some Russian Listings. https://www.wsj.com/livecoverage/russia-ukraine-latest-news-2022-03-02/card/vM2no1PgGDmMkL2TSvPZ. Accessed: 2022-10-13.

Ensafi, R.; Winter, P.; Mueen, A.; and Crandall, J. R. 2015. Analyzing the great firewall of china over space and time. *PETS*.

European Commission. 2023. The Digital Services Act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Accessed: 2023-09-15.

Fiesler, C.; Beard, N.; and Keegan, B. C. 2020. No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 187–196.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2023-09-15.

Friggeri, A.; Adamic, L.; Eckles, D.; and Cheng, J. 2014. Rumor cascades. *ICWSM*, 8(1): 101–110.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Geissler, D.; Bär, D.; Pröllochs, N.; and Feuerriegel, S. 2023. Russian propaganda on social media during the 2022 invasion of Ukraine. *EPJ Data Science*, 12(1): 35.

Google Help. 2022. Maps user-generated content policy. https://support.google.com/contributionpolicy/answer/7422880. Accessed: 2022-05-19.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hamilton, I. A. 2022. Google and TripAdvisor disable restaurant reviews in Russia after they were flooded with protests against the Ukraine invasion. https://www.businessinsider.com/google-tripadvisor-disable-reviews-russia-ukraine-2022-3. Accessed: 2022-05-04.

Hanley, H. W.; and Durumeric, Z. 2023. Partial Mobilization: Tracking Multilingual Information Flows Amongst Russian Media Outlets and Telegram. *arXiv preprint arXiv:2301.10856*.

Hanley, H. W. A.; Kumar, D.; and Durumeric, Z. 2022. Happenstance: Utilizing Semantic Search to Track Russian State Media Narratives about the Russo-Ukrainian War On Reddit. In *ICWSM*.

Howard, P. N.; Agarwal, S. D.; and Hussain, M. 2011. When Do States Disconnect Their Digital Networks? Regime Responses to the Political Uses of Social Media. *SSRN*.

Jonker, M.; Akiwate, G.; Affinito, A.; kc Claffy; Botta, A.; Voelker, G. M.; van Rijswijk-Deij, R.; and Savage, S. 2022. Where.ru? Assessing the Impact of Conflicton Russian Domain Infrastructure. In *IMC*.

Kaufer, S. 2022. An open letter on Ukraine from Tripadvisor's Steve Kaufer. https://www.tripadvisor.com/Articles-l1poX3FsJ3oo-Statement_from_steve_kaufer.html. Accessed: 2022-05-04.

Kenneally, E.; and Dittrich, D. 2012. The Menlo Report: Ethical principles guiding information and communication technology research. *Available at SSRN 2445102*.

Lerman, K.; and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 4(1): 90–97.

Marczak, B.; Weaver, N.; Dalek, J.; Ensafi, R.; Fifield, D.; McKune, S.; Rey, A.; Scott-Railton, J.; Deibert, R.; and Paxson, V. 2015. An Analysis of China' "Great Cannon". In *FOCI*.

Meaker, M. 2022a. Activists Are Reaching Russians Behind Putin's Propaganda Wall. https://www.wired.com/story/russia-propaganda-wall/. Accessed: 2022-05-04.

Meaker, M. 2022b. Russia Blocks Facebook and Twitter in a Propaganda Standoff. https://www.wired.com/story/russia-ukraine-social-media/. Accessed: 2022-05-04.

Morgans, M. J. 2017. Freedom of Speech, the War on Terror, and What's YouTube Got to Do with It: American Censorship during Times of Military Conflict. *Fed. Comm. LJ*, 69: 145.

Niaki, A. A.; Cho, S.; Weinberg, Z.; Hoang, N. P.; Razaghpanah, A.; Christin, N.; and Gill, P. 2020. ICLab: A global, longitudinal internet censorship measurement platform. In *SP*, 135–151. IEEE.

Nourin, S.; Tran, V.; Jiang, X.; Bock, K.; Feamster, N.; Hoang, N. P.; and Levin, D. 2023. Measuring and Evading Turkmenistan's Internet Censorship: A Case Study in Large-Scale Measurements of a Low-Penetration Country. In *Web Conference*, 1969–1979.

OONI. 2023. How Internet censorship changed in Russia during the 1st year of military conflict in Ukraine. https://ooni.org/post/2023-russia-a-year-after-the-conflict/. Accessed: 2023-09-15.

Open Observatory of Network Interference. 2022. Data — OONI. https://ooni.org/data/. Accessed: 2022-05-18.

Ortwein, A.; Bock, K.; and Levin, D. 2023. Towards a Comprehensive Understanding of Russian Transit Censorship. In *FOCI*.

Pearce, P.; Jones, B.; Li, F.; Ensafi, R.; Feamster, N.; Weaver, N.; and Paxson, V. 2017. Global measurement of DNS manipulation. In *USENIX Security*, 307–323.

Pierri, F.; Luceri, L.; Chen, E.; and Ferrara, E. 2023a. How does Twitter account moderation work? Dynamics of account creation and suspension on Twitter during major geopolitical events. *EPJ Data Science*, 12(1): 43.

Pierri, F.; Luceri, L.; Jindal, N.; and Ferrara, E. 2023b. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. In *WebSci*, 65–74.

Pohl, J. S.; Markmann, S.; Assenmacher, D.; and Grimme, C. 2023. Invasion@Ukraine: Providing and Describing a Twitter Streaming Dataset That Captures the Outbreak of War Between Russia and Ukraine in 2022. In *ICWSM*, 1093–1101.

Price, B. 1942. Governmental censorship in war-time. *American Political Science Review*, 36(5): 837–849.

Rivers, C. M.; and Lewis, B. L. 2014. Ethical research standards in a world of big data. *F1000Research*, 3: 38.

Shevchenko, V. 2022. Ukraine war: Protester exposes cracks in Kremlin's war message. https://www.bbc.com/news/world-europe-60749064. Accessed: 2022-05-11.

Silverman, C.; and Kao, J. 2022. Infamous Russian Troll Farm Appears to Be Source of Anti-Ukraine Propaganda. https://www.propublica.org/article/infamous-russian-troll-farm-appears-to-be-source-of-anti-ukraine-propaganda. Accessed: 2022-05-04.

Squad303. 2022. Media coverage. https://1920.in/media.html. Accessed: 2023-09-15.

Sundara Raman, R.; Shenoy, P.; Kohls, K.; and Ensafi, R. 2020. Censored planet: An internet-wide, longitudinal censorship observatory. In *CCS*, 49–66.

Tran, N.; Bock, K.; and Levin, D. 2023. Crowdsourcing the Discovery of Server-side Censorship Evasion Strategies. In *FOCI*.

Troianovski, A. 2022. Russia Takes Censorship to New Extremes, Stifling War Coverage. https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html. Accessed: 2022-05-11.

Tschantz, M. C.; Afroz, S.; Paxson, V.; et al. 2016. Sok: Towards grounding censorship circumvention in empiricism. In *SP*, 914–933. IEEE.

Unwala, A.; and Ghori, S. 2016. Brandishing the cybered bear: Information war and the Russia-Ukraine conflict. *Military Cyber Affairs*, 1(1).

Volkova, S.; and Bell, E. 2016. Account deletion prediction on RuNet: A case study of suspicious Twitter accounts active during the Russian-Ukrainian crisis. In *Workshop on Computational Approaches to Deception Detection*, 1–6.

Wikipedia contributors. 2022. List of cities and towns in Russia by population. https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Russia_by_population. Accessed: 2022-05-11.

Xue, D.; Mixon-Baca, B.; ValdikSS; Ablove, A.; Kujath, B.; Crandal, J. R.; and Ensafi, R. 2022. TSPU: Russia's Decentralized Censorship System. In *IMC*.

Xue, D.; Ramesh, R.; Evdokimov, L.; Viktorov, A.; Jain, A.; Wustrow, E.; Basso, S.; and Ensafi, R. 2021. Throttling Twitter: an emerging censorship technique in Russia. In *IMC*, 435–443.

Xynou, M.; and Filastò, A. 2022. New blocks emerge in Russia amid war in Ukraine: An OONI network measurement analysis. https://ooni.org/post/2022-russia-blocks-amid-ru-ua-conflict. Accessed: 2022-05-16.

@YourAnonNews. 2022. Go to Google Maps. Go to Russia. Find a restaurant or business and write a review. When you write the review explain what is happening in Ukraine. Idea via @Konrad03249040. https://twitter.com/YourAnonNews/status/1498337491056836610. Accessed: 2022-05-04.

Ethics 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
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see contributions from §1
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, see §4
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? N/A
   (e) Did you describe the limitations of your work? Yes, see §4.3
   (f) Did you discuss any potential negative societal impacts of your work? N/A, as there are no negative societal impacts derived from our findings
   (g) Did you discuss any potential misuse of your work? N/A, as there are no potential misuse derived from our work
   (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? Yes, see *Ethics and Broader Perspective* section
   (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? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
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)? N/A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (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)? N/A

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? N/A

(b) Did you mention the license of the assets? N/A

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we provide the list of keywords used for labeling war-related content

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see *Ethics and Broader Perspective* section

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see *Ethics and Broader Perspective* section

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? N/A

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see *Ethics and Broader Perspective* section

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

(d) Did you discuss how data is stored, shared, and deidentified? Yes, see *Ethics and Broader Perspective* section