Wartime Media Monitor (WarMM-2022): A Study of Information Manipulation on Russian Social Media during the Russia-Ukraine War

Maxim Alyukov
King's College London
maxim.alyukov@kcl.ac.uk

Maria Kunilovskaya
University of Saarland
maria.kunilovskaya@uni-saarland.de

Andrei Semenov
Higher School of Economics
andrey.semenov@hse.ru

Abstract

This study relies on natural language processing to explore the nature of online communication in Russia during the war on Ukraine in 2022. The analysis of a large corpus of publications in traditional media and on social media identifies massive state interventions aimed at manipulating public opinion. The study relies on expertise in media studies and political science to trace the major themes and strategies of propagandist narratives on three major Russian social media platforms over several months as well as their perception by the users. Distributions of several keyworded pro-war and anti-war topics are examined to reveal the cross-platform specificity of social media audiences. We release WarMM-2022, a 1.7M posts corpus. This corpus includes publications related to the Russia-Ukraine war, which appeared in Russian mass media (February to September 2022) and on social networks (July to September 2022). The corpus can be useful for the development of NLP approaches to propaganda detection and subsequent studies of propaganda campaigns in social sciences in addition to traditional methods, such as content analysis, focus groups, surveys, and experiments.

1 Introduction

Contemporary autocracies rely on media manipulation more than violent dictatorships of the past (Guriev and Treisman, 2020). As citizens might recognise manipulative intent (Roberts, 2018), authoritarian governments attempt to veneer the propaganda messages via state-sponsored social networks. These online "astroturfing" campaigns (Zerback and Töpfl, 2022) appear as a genuine grassroots support for the regime and artificially inflate the visibility of pro-regime messages. In this paper, we document the presence of such a campaign in Russia in 2022 and explore its key characteristics, using a large corpus of online messages from Russian social media about the Russian-Ukrainian war (WarMM-2022).

Our data-driven approach can provide a more realistic picture of audience response to political information in the context of war than traditional methods of communication research, such as surveys. An important outcome of this media monitoring project is a corpus of online publications on the Russian-Ukrainian war which appeared the websites of Russian newspapers and TV channels between February and September 2022 and on a number of social media platforms between July and September 2022. The corpus includes temporal, spatial, and some socio-demographic metadata, which can be used to develop NLP approaches to the detection of various forms of propaganda.

To the best of our knowledge, there is only one published dataset, *VoynaSlov* (Park et al., 2022), which is specifically designed to capture media coverage of, and public reaction to, content related to the Russia-Ukraine war. *VoynaSlov* includes posts from a limited pre-defined number of news outlets (42 in total) published on either *VKontakte* or *Twitter*. The data from these platforms have been sampled following dissimilar approaches: the *Twitter* subset includes posts with war-related hashtags while there is no such filter for *VKontakte*. In terms of the amount of textual data, *VoynaSlov* includes 597K documents published on *VKontakte* and 219K on *Twitter*. Our dataset has a much broader coverage with regard to the sources of information and includes only the publications about the war, although our time frame is more limited. Our dataset presents a realistic snapshot of the online information environment experienced by Russian internet users in real-time during the war, and we hope that this resource can be useful not only for the NLP community but also for communication scholars and political scientists.

2 Corpus Description

The WarMM-2022 corpus represents online political discourse produced in Russia for and by domestic audiences. The corpus is composed of two parts: a subcorpus of publications by traditional mass media (press and TV) and public posts from social networks. The full list of sources includes 415 websites of media outlets, 25 websites of TV channels, and 85 social media platforms. The distribution of publications is very uneven across the sources of each type. Most active media websites are by *gazeta.ru, ura.news, ren.tv, vz.ru, russian.rt.com, iz.ru*. The content of TV programmes related the Russian-Ukrainian war is captured by the respective transcripts published by TV channels on their websites. In our collection, the transcripts most often come from *Channel One, REN TV, Channel 5*, and *Russia 24*. Importantly, WarMM-2022 includes the regional affiliations of information sources allowing one to trace the specifics of Ukraine-related news coverage across Russia. By the number of collected documents, the most represented social media platforms are *VKontakte, Odnoklassniki, Telegram, Twitter, Facebook, LiveJournal, YouTube* (in decreasing order). The full list of mass-media, TV and social-media sources is released with the corpus. The textual data and associated metadata, including public reactions, such as the number of views, likes, re-posts and comments, were obtained with the technical support of *Scan Interfax* and *Brand Analytics* media monitoring systems. The parameters of data collection were configured to meet the requirements of the current project. Media sources were limited by their availability to Russian audiences, i.e. the crawled webpages were directly accessible in Russia at the time of collection. Social media sources were limited to posts from accounts that were registered in Russia and published in Russian, where possible.

We aim to produce a realistic snapshot of the online information environment regarding the war and during the war. The data collection began in July 2022 in a monitoring mode. We were aggregating publications that appeared on mass media websites and on social media daily until the end of September 2022. A separate subcorpus of publications on mass media and TV websites for the preceding months (February to June) was collected in a one-time retrospective crawling effort in mid-July 2022. The corpus was built using a list of eight general context keywords to filter in publications related to the war in Ukraine. These eight terms were *war, special operation, military operation, SVO (special military operation), special operation, military operation, denazification, and demilitarization* (in Russian). This list was developed as a result of iterative filter-setting experiments and manual analysis of daily crawls in the first two weeks preceding the start of the data collection.

The basic statistics for the WarMM-2022 corpus are presented in Table 1.

period	Press+TV	Social Media
February	12.7 K	–
March	27.3 K	–
April	19.9 K	–
May	21.4 K	–
June	15.9 K	–
July	18.2 K	602 K
August	28.7 K	546 K
September	38.3 K	558 K
Total	182 K	1,706 K

Table 1: Number of posts by month and media type

Table 1 shows that traditional media (Press+TV) and social media subcorpora are not well balanced by the number of included documents. Only about 10% of texts come from traditional media websites. The disbalance between subcorpora persists in terms of the overall word counts (not shown in Table 1): the overall size of the press subcorpus is 24.4 M tokens, TV transcripts - 1.7 M tokens, and social media subcorpus includes 268.4 M tokens. The analysis presented in this paper is based on the data from three months (July to September) - the period present in both press+TV and social media subcorpora.

Section 5 reports a cross-platform study focusing on the three most popular social media platforms in Russia (at the backdrop of traditional media content): *Odnoklassniki* (OK), *Telegram* (TG) and *VKontakte* (VK). Table 2 displays the parameters of the underlying subcorpus.

After Facebook and Instagram were banned in March 2022, OK, TG and VK became the dominant platforms in Russia alongside WhatsApp and YouTube. According to April 2022 data, 62% of Russians used VK, 55% used TG, and 42% used OK[1]. OK is often considered a space of Putin's electorate. Its audience is much older than the audience

[1] WCIOM (2022) Rossiyskaya Auditoriya Socialnih Setey: Izmeneniya Na Fone Specoperatsii. Avaliable at: https://wciom.ru/analytical-reviews/analiticheskii-obzor/rossiiskaja-auditorija-socialnykh-setei-i-messendzherov-izmenenija-na-fone-specoperacii

period	network	docs	words
July	ok.ru	153.8 K	26.6 M
	telegram.org	18.2K	2.7 M
	vk.com	334.4 K	87.1 M
August	ok.ru	169.5 K	33.7 M
	telegram.org	14.4 K	2.3 M
	vk.com	278.9 K	69.8 M
September	ok.ru	250.8 K	51.3 M
	telegram.org	15.0 K	2.4 M
	vk.com	309.9 K	76.2 M
Total		1,545 K	352.0 M

Table 2: Social media subcorpus size by network in tokens (the counts are given after pre-processing and annotation)

of other platforms. According to 2021 data, 7.4% of OK users were under 24, 25.2% were between 34-44, and the dominant 49.5% were older than 45 [2]. Public groups on OK are often anti-Western and pro-Kremlin and constitute the regime's 'Virtual Russian World' not only in Russia but in other countries with significant Russian-speaking populations (Teperik et al., 2018). VK has a much younger audience than OK: according to 2021 data, a dominant 31.3% of VK users were under 24 and only 18% were older than 45. Finally, the audience of a relative newcomer, TG, is slightly older than the audience of VK. The 2021 data reveals that 29.6% of TG users were under 24, dominant 30.6% were 24-34, and 18.5% were older than 45 [3].

In what follows, we describe the social and political context of the study, introduce theoretical concepts from media and communications research necessary to interpret our data (Section 3), present methodology (Section 4), and report analytical results and their interpretation (Section 5). We conclude with a summary (Section 7) and reflections on the limitations and ethical aspects of our project.

3 Background

3.1 Russia's Networked Authoritarianism

The rapid development of digital media in the early 2000s led some analysts to praise it as a "liberation technology" (Diamond and Plattner, 2012). Responding to this threat, authoritarian governments have been investing significant resources in creating various forms of "networked authoritarianism" (MacKinnon, 2011). Different models of control over online media emerged over time from largely hands-off policies to nurturing sophisticated digital environments conducive to authoritarian messaging (Greitens, 2013). More recently, *online astroturfing* – the strategy of giving online conversations seemingly genuine pro-governmental spin – emerged as "a novel form of disinformation that relies on the imitation of citizen voices to create the false impression that a particular view or idea has widespread support in society" (Zerback and Töpfl, 2022).

In Russia, the initial Kremlin's position not to disrupt online communications changed after the 2011-12 post-electoral protest (Sanovich et al., 2018). As a part of the "third generation controls" (Deibert et al., 2010), in the past ten years, Putin's government has been actively using automated bots and trolls (paid humans who rely on scripts to produce content) to shape online discussions (Sanovich et al., 2018). The Kremlin has been extensively using bots to create information noise, to promote pro-governmental messages in search engine results and in news aggregators, and to manufacture popularity of autocratic agents (Stukal et al., 2017, 2022). Research also shows that the Kremlin-linked agencies have conducted multiple information campaigns attempting to influence public opinion abroad (Linvill and Warren, 2020; Elshehawy et al., 2021).

After the full-scale invasion of Ukraine, the Kremlin shut down many remaining independent media and introduced repressive laws effectively imposing wartime censorship, hammering any public expression of discontent with the war. In addition, it started to use paid commentators and *"voenkors"* (military reporters on the battlefront) to shape citizens' perceptions of the invasion[4]. In our study, we attempt to document and explore the astroturfing campaign related to war using the WarMM-2022 corpus.

[2] Statista (2021). Gender and age distribution of Odnoklassniki. Available at: https://www.statista.com/statistics/1065018/russia-odnoklassniki-users-share-by-age/

[3] Brand Analytics (2021). Socialniye Seti v Rossii: Trendi i Tsifri, Osen' 2021. Available at: https://br-analytics.ru/blog/social-media-russia-2021/

[4] Fontanka (2022). «Vy ved' ne verite, chto jeto nastojashhie otzyvy?» Kak «Fontanka» zagljanula na peredovuju informacionnyh frontov Z. Available at: https://www.fontanka.ru/2022/03/21/70522490/

3.2 NLP for Communication Research and Political Science

The NLP community developed multiple methods for the analysis of mass media communications - in particular, for detecting various forms of information manipulation online. Most NLP research on propaganda uses supervised methods that require manual annotation, sometimes very fine-grained (see, for example, Da et al., 2020). These projects are focused primarily on solving NLP tasks rather than obtaining results requested by social sciences with regard to unfolding events. Park et al. (2022) pursued a task similar to what we face: they explored the proportion of topics in social media publications to reveal such information manipulation strategies as agenda-setting, framing, and priming. They employ state-of-the-art *unsupervised* methods for topic modeling (a structured topic model and a contextualized neural topic model) and frame analysis (using a zero-shot learning scenario and ignoring the differences in language, style, and cultural context between the available training data and intended end-use domain). However, they admitted that the results were contradictory, obscure, and difficult to interpret in both cases. Interestingly, they fall back on word statistics as a more reliable yardstick to evaluate their models. Elshehawy et al. (2021) relied on constructed lexicons to provide evidence that the Kremlin promoted refugee stories in the German media sphere in an attempt to influence the outcome of the elections. Following the principles of transparency of analysis and interpretability of its outcomes, we opted for the keyword frequency analysis approach as our main method for this preliminary study.

4 Methodology

The analysis of news topics includes statistical and unsupervised methods. The findings are interpreted in the context of external unfolding events. In this project, we constructed expert-curated lists of topical keywords, and their normalised frequencies were used to compare messaging across social media platforms (OK, VK, TG) in a time series fashion.

General frequency analysis setup. We focused on frequency of individual keywords and aggregated frequencies of pre-defined terms that marked a particular topic. In total, we explored 20 thematic aspects of the publications and extracted the frequencies of over 250 words and phrases.

The full list of search items in their non-lemmatised version in Russian for each topic is available in the corpus documentation. It is designed to include topics that are typical for pro-war and anti-war discourses as well as shared between the two. Appendix A lists the topics and subtopics. For example, we traced "dehumanisation", the topic marked by the use of derogatory names for the Ukrainians (e.g. *ukrop*) and numerous derivatives with *ukro-* and *nazi-* prefixes (e.g. *ukronazist, ukrofascist, banderovetz, nazbat*) as well as loaded ideological terms used to describe Ukraine (e.g. *Kyiv regime, sneaky, guileful, hypocrite*). The frequency of each item was based on a lemmatised version of the corpus to account for possible grammatical forms, which is important for morphologically-rich languages like Russian. The raw texts went through minimum preprocessing before lemmatisation, including symbol unification, discarding .png/.jpg and url to reduce noise. The lemmatised version of the corpus was obtained from morpho-syntactic annotation produced using UDPipe (v1, Straka and Straková, 2017), a parser within the Universal Dependencies framework. All frequencies were normalised to the size of the respective subcorpora within a given time series and subcorpus, with the normalisation base of 100,000 words. This made possible the comparison of frequencies across subcorpora of various sizes directly, including using them in graphs based on the same scale.

Time series. As we were interested in fluctuations of topical content, we constructed time series using three-day intervals as our default setting, i.e. most results in this study reflect the frequencies of search items in the documents published within successive 3-day periods. Whenever we wanted to explore a specific timespan in more detail or have a more aerial perspective, we analysed daily or monthly frequencies, respectively.

Unique publications vs repetitive content. Taking into account the anticipated repetitiveness of publications, we compared the frequencies of selected keywords before and after deleting duplicate posts. Duplicate posts were identified by matching the first 20 words in the raw text. The ratio of repeated texts (excluding the first occurrence) amounts to 47.98% on social media, with about 23.4% being exact unmodified copies of the original publication, often repeated many times.

User attitude studies. Several analytical approaches were employed in an attempt to reveal the users' attitudes to the topics discussed online. It is a challenging task as Russia's information environment is heavily censored and populated with bots and trolls masquerading as real citizens; risks of legal prosecution make it difficult for users to state their positions publicly. In an attempt to overcome these limitations, we analysed (i) publications by users with different levels of publication activity and (ii) publications with the highest engagement scores.

User groups by publication activity. This analysis was based on social media subcorpus only. The total number of unique users in this subcorpus is 263,665. Users produced 1,544,918 posts in three months. In particular, we distinguish between professional users (more likely accounts of established information agencies) who publish more than 20 posts about Ukraine a week (over 260 in three months) and the general public, i.e. users with one or fewer posts a week across 13 weeks. Other users include an intermediate group of active users with over 13 but less than 260 posts per week. The parameters of each group and their contribution to the production of content on the analysed social networks can be found in Table 3.

user group	users	% users	% posts
professional	710	0.27	22.62
active users	17,630	6.69	44.59
general public	245,325	93.04	32.78

Table 3: Activity of media outlets and ordinary users on social media

Activity patterns varied across the groups, with the professional users capable of generating surplus texts during some periods of time and being less active during others. These groups also demonstrated different patterns in the use of keywords from a range of analysed topics (see Section 5).

Engagement scores. The posts on social media were analysed from the point of view of public reactions they generated. We calculated the engagement index as a sum of likes, re-posts, and comments to each post and built a subcorpus of the most popular posts, which included 5% of all posts sorted by the involvement score. This subcorpus included 85,317 posts (out of 1,706,343 in the entire social media corpus across three months). The engagement scores in this subcorpus ranged from 110,528 to 39, with a mean of 444.1. Furthermore, to estimate the level of support for the pro- and anti-war messages and assess their visibility and influence in public discussion, we selected the top-1000 posts with the highest engagement score reactions and identified their sources. Using external knowledge, we classified these 215 authors as pro-war, anti-war, or neutral. Finally, we counted the number of posts by these authors and their overall engagement scores.

5 Results: Cross-platform Analysis

This section applies the described methodology to explore users' participation in online discussions, their attitudes and reactions to media content.

To capture the astroturfing campaign, we use a range of selected topics (pro-war and anti-war), which also help us to reveal political attitudes prevailing on the selected social media platforms. Generally, we noticed that social media had deviated from the press and TV in the frequency patterns of pro-war topics. The usage of keywords was very volatile with several regular peaks. We identified these peaks as massive infusions of almost identical messages. In Figure 1 these peaks are smoothed out by removing duplicate publications. In combination with observations for other topics, this can be a sign of information manipulation.

As discussed above, the major Russian social media platforms (Odnolassniki, VKontakte, Telegram) vary according to socio-demographic parameters. To disentangle the effect of repetitive publications observed in the entire social media subcorpus, shown in Figure 1, we looked at the frequencies of each topic by network.

First, the frequencies of *denazification* and *demilitarisation* – the key terms justifying Russian invasion – on TG and VK are both steady and low in comparison to abnormal spiky patterns registered in OK publications (the extreme frequencies ranging between 45 and 60 per 100K words vs over 140 on OK). Note that *demilitarisation*, which disappeared from state-controlled media accounts towards the end of summer 2022, fell into disuse on OK, too. Removing identical posts levels off the spikes in the use of justifications on OK. With duplicates removed, the three platforms demonstrate similar frequencies.

The same pattern across platforms is observed for *dehumanising language*. There is a significant amount of anti-Ukrainian derogatory terms on all

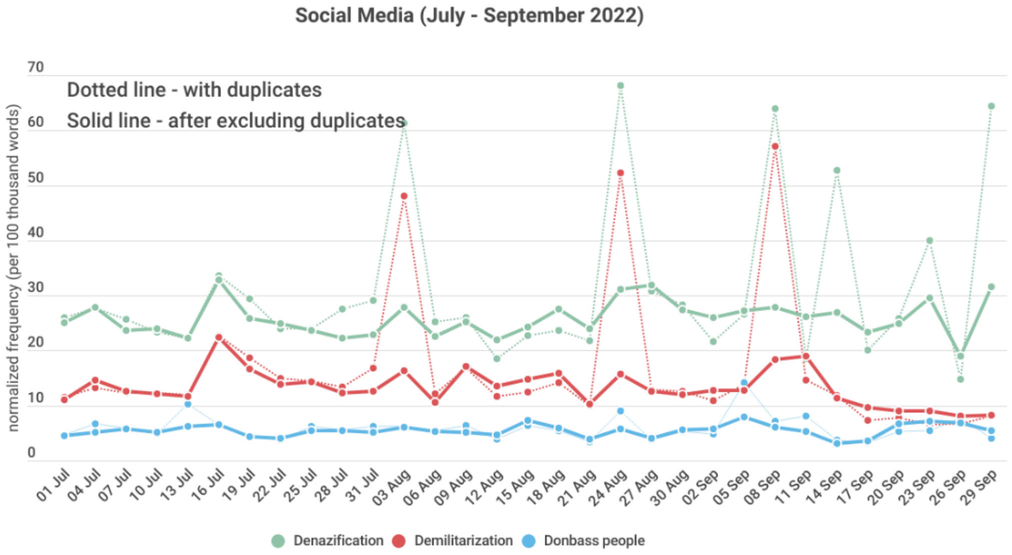

Figure 1: Impact of repeated publications: key concepts used to justify Russian actions in Ukraine (3-day aggregates)

three platforms. Markers of this language are more frequent on OK, where every month there are significant spikes that exceed the volume of VK and TG by the factors from two (late July and late August) to four (early September), and six (early August). The patterns on VK and TG are stable and do not suggest any artificial inflation. Nevertheless, all social media platforms remain plagued with hate speech toward Ukrainians.

To double-check that the observed peaks are artificial, aggregated frequencies based on the entire corpus after removing duplicates were compared. The amount of dehumanising vocabulary decreased manifold. Nonetheless, even without identical messages, OK remains the most pro-war platform followed by VK and TG. These observations suggest that the Kremlin disproportionately targets OK with pro-war online astroturfing.

The function of state-controlled trolls cannot be reduced to producing identical content. Paid users are often instructed to improvise and can produce original messages different from each other. Hence, identical messages alone do not represent the scale of online astroturfing accurately. However, as identical messages are unlikely to be attributed to anything else but artificial content, removing it can make us underestimate the scale of online astroturfing, not to overestimate it. In other words, identical messages are a conservative estimate of the scale of Kremlin-related astroturfing.

One might argue that the messages we identified as astroturf were viral and resonated with the online public. While we cannot exclude such a possibility, the short life-span of these messages (they disappeared completely from communications in 1-2 days) and a poor quality of its content (we found nothing sensational or novel in terms of production for several manually selected entries) indicate that the traction with the general public was limited at best.

To further investigate differences in ideological spin and the scale of online astroturfing, we focus on anti-war vocabulary. Figure 2 (on the left) shows the aggregated frequencies of keywords typical for Kremlin opponents, such as *Russian aggression, annexation, occupation of Ukrainian territories, Russian invasion, occupation of Donbas/Crimea, Russian occupants*, etc.

The graphs reflect the fluctuations in the use of anti-war language across platforms from July to September. It shows that TG is the most anti-war platform among the three. Despite the presence of many pro-Kremlin channels, the independent media flourish, too. Anti-war vocabulary is the least present on OK. Removing duplicates (the right panel in Figure 2) does not change the observed pattern significantly. If anything, the absence of repetitive content makes anti-war stance on VK more visible (notice the upward shift of the orange line in the right-hand panel). This might suggest that spamming the information space with duplicates makes sense as it creates additional noise and makes it more difficult for users to hear other voices. With identical messages removed, TG remains the most anti-war platform. Patterns on VK and OK resemble each other implying that they are

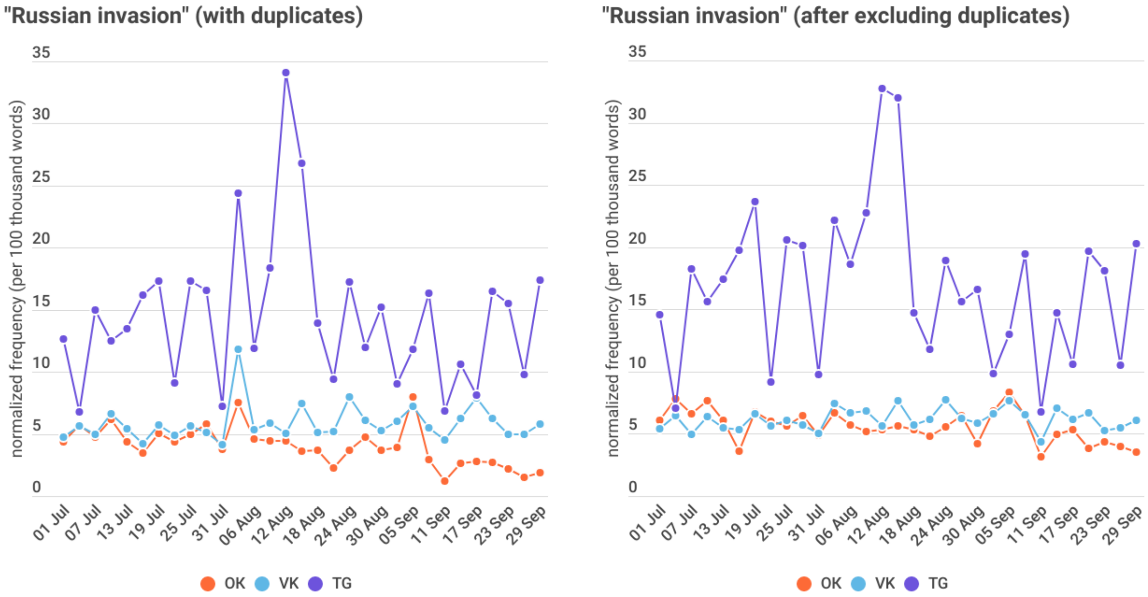

Figure 2: Anti-war language across platforms (weekly aggregates)

more similar in terms of ideological spin.

Additionally, we explored topics related to time-specific external events, such as the Ukraine's advances on the battlefield in late August - early September (e.g. *advances of Ukrainian army, Ukraine's military success, retreat of Russian troops, successful counterattack of Ukrainian forces, Russian defeat*, etc). Despite censorship on official press and TV, the news about the failures of the Russian military percolated to social media, especially TG. The frequencies capturing this topic were very low but genuine: removing duplicates did not affect the counts.

We conclude that OK was disproportionately targeted by the regime. The presence of astroturfing can also be confirmed by looking at the phrases from the *temniki* – the guidelines issued by the Kremlin to cover politically sensitive topics in the media. Unlike the other two platforms, OK demonstrated a growing scale of occurrences for this vocabulary from early August onwards suggesting that this platform was the primary target.

Public perception. The analysis of publications by professional users, who produce over 20 publications a week about Ukraine, and regular users, who might represent the general public, showed that the latter were less eager to portray Ukrainians as the enemy. In Figure 3, the flat orange line represents relatively low and stable frequencies for dehumanising vocabulary (which are still very high in comparison with visible anti-war stance in other graphs).

In a subcorpus built from the top 5% of publications based on engagement scores, anti-war topics (e.g. Ukrainian military success, framing Russian actions in Ukraine as war, occupation or invasion) are more frequent, while pro-war rhetoric is noticeably less dense than in the entire corpus.

Out of 215 authors who produced the publications with the highest engagement score, 80 were classified as anti-war (479 out of 1000 most popular posts), 24 authors as neutral with (27 posts), 111 authors were classified as pro-war (494 posts). However, in the top 50 most popular posts, we found only two posts by pro-war authors. The first 38 posts by the level of engagement and 48 posts in total (out of 50) were written by anti-war authors. As a result, the level of involvement for anti-war posts is 1.5 times higher than pro-war posts (8.5 mln reactions vs 5.7 mln reactions).

6 Discussion and consolidation

By tracing the frequencies of pro-war and anti-war topics across social media platforms over time, linking them to external events and checking for repetitive content, we were able to reveal signs of information manipulation aimed at shaping public opinion by controlling the agenda and framing the concepts to fit the current ideology.

As we demonstrate, some of the major themes artificially promoted by the state include: (i) the ideas of denazification and demilitarisaion and (ii) dehu-

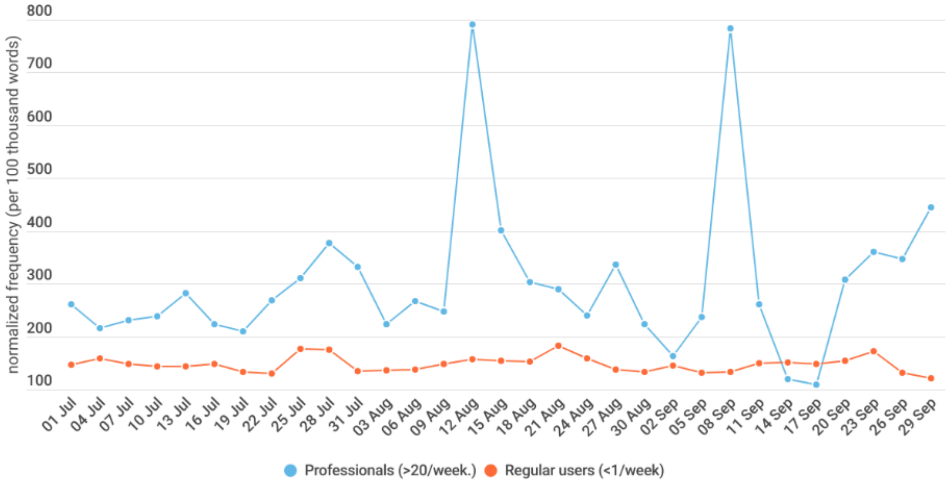

Figure 3: Anti-Ukrainian hate speech is publications by most prolific accounts and regular users

manization of Ukrainians. Based on this dataset, we also identify several other promoted ideas, such as (iii) the existential threat posed by NATO and (iv) framing Ukraine as controlled by the hostile ‘collective West’. However, we do not include them in the analysis due to limited space.

The strategies of public opinion manipulation can include the attempts to undermine trust in mass media, frame any source of information except state-controlled as spreading ‘fakes’, and appropriate opponents’ vocabulary and diluting its meaning. In this analysis, we demonstrate one of these strategies: imitation of popular support for promoted ideas on social media.

Our findings point at possible mechanisms behind the Kremlin’s digital war propaganda. Instead of attempting to reach war opponents or users without clear preferences, the regime’s astroturf communication seems to flourish in a predominantly pro-war environment. In line with both classical research on media effects (Lazarsfeld et al., 1960) and contemporary research on the effects of propaganda in authoritarian Russia (Shirikov, 2022), these findings suggest that the main strategy of the regime’s astroturf online communication might be similar to the one of authoritarian propaganda: to reinforce beliefs of those who are already pro-regime rather than to win new supporters.

Our cross-platform analysis indicates that discussions on OK are largely influenced by astroturfing. TG remains a relatively free space devoid of official rhetoric, while VK users exhibit a mild tendency to re-produce official narratives about the war.

Aiming at revealing the level of support for promoted ideas and the effectiveness of the said strategies, our analyses based on user group activity and reactions on social media demonstrated that many of these narratives fall flat on domestic audiences. Modest numbers of Russians participating in public discussions show that a lot of communication online is one-way, with people withdrawing from the public space. The public reactions that are available in WarMM-2022 demonstrate that the extent of support for the promoted ideas is rather limited.

7 Conclusion

This study reports details of textual data collection and analysis in the interests of social sciences. We release WarMM-2022, a corpus of public online communications collected from a large number of mass media websites and social media platforms, which was used to obtain the results reported in this paper.

Our analysis relies on expert-curated lists of words and phrases which are used to cross-examine topical content in posts from a wide range of Russian mass media and most popular social media, published in July-September 2022. Informed by the previous work on data-driven propaganda detection, we aimed to assess the scale and societal impact of media manipulation in wartime Russia. In particular, we were interested in the distribution of, and support for, selected topics reflecting opposite viewpoints on the events in Ukraine. We revealed that the distribution of topics in social media (unlike traditional media, including TV) was largely affected by state-controlled interventions that varied in scale across the three social media compared in this work. The patterned nature of these interventions and their alignment with the Kremlin's intentions expressed in recommendations for the press suggest that these are signs of "networked authoritarianism", a system of measures to exert control over the internet. The study was focused on the pro-war and anti-war themes in social networks and revealed a considerable amount of "astroturfing" (imitation of public support online). Our results support the idea that the Kremlin employs a digital propaganda ecosystem including networks of state-controlled accounts – bots and paid influencers – across Russia's main social media platforms. This ecosystem is engaged in an organised manner to shape public opinion on current or forthcoming events. The frequency patterns of topics related to the Russia-Ukraine war reveal the artificial nature of online communication on Russian social media in July-September 2022 and help us to identify the key messages infused by the astroturfing campaigns, such as the existential threat posed by NATO, the need for patriotic unity against the hostile West and dehumanisation of the Ukrainians. Although anti-war voices are largely silenced by censorship and the threat of persecution, these opinions are heard and get more public attention than any propagandist content. At the same time, the number of Russians who get publicly involved in online participation as authors is ridiculously small. As users, Russians have to navigate an increasingly volatile, noisy, and restrictive environment infused with highly repetitive pro-war content.

Limitations

By construction, our corpus is not representative in any sense of the general population of messages related to the war on RuNet. Platforms blocked in March 2022 (notably, Facebook and Instagram) remained largely absent. Also, we restricted our analysis to Russian-based users while many Russian-speaking digital communities remained active from outside the country. Lexicon-based approaches are necessarily limited by the scope of topics that they are able to cover. Similarly, data collection decisions reduce the claims that can be made in the findings to the observations relevant to the given dataset. We admit that the explored topics do not exhaust the ideas that circulated online within the given time frame, and it is likely that we missed other important themes. Besides, despite care was taken to avoid including ambiguous keywords, and unexpected frequencies were checked in manual analysis at the stage of constructing the lexicons, simple word statistics cannot identify contexts. In fact, elements of pro-war narratives can occur in essentially anti-war publications as examples or mockery of the opponents' discourse. Keywords can generate high frequencies from a few repetitive documents in a time series, too. A better approach would be to operate at the level of documents and report results for complete statements rather than words. The manual analysis and annotation attempts demonstrated that social media content is very dependent on multimedia. However, we focused on the textual content and discarded linked images or videos. Finally, given the large number of comparisons that we carried out, we did not see it feasible to perform a proper statistical analysis of the differences between various subcorpora.

Ethics Statement

In considering the ethical aspects of this study, we strive to avoid any potential harm to individual Internet users or publishing outlets, to protect their privacy, and to respect their right to the created texts. These considerations motivated the following practical decisions. First, we used the publications that were publicly available at the time of collection. Second, the corpus is made available only as a list of links augmented with non-revealing attributes, such as date, media type, source (website or platform), region and engagement score (for social networks subcorpus) [5], with the actual textual content deleted from this version of the corpus. It is done to protect the users' right to take down their content and to avoid violating their copyright. While these restrictions imply reduced replicability of our results and additional efforts for researchers associated with the necessity to recollect the data, they were considered ethical to avoid potential harm to individuals. Third, we do not distribute any metadata or publish any considerable parts of the collected texts that can be used to identify individuals with particular political beliefs at the moment or in the future. This is particularly important, given the current scale of prosecution for anti-war publications and reactions expressed online in Russia. Finally, while we admit that research on propaganda strategies can be used to improve ways of information manipulation, we think that uncovering and describing these practices serves

[5] https://github.com/kunilovskaya/WarMM-2022

the greater social good of raising the awareness of the public about types of disinformation and potentially delusive environments that can be created online. When presenting the results of the analyses, care was taken to avoid any wording that can be interpreted as promoting particular political beliefs, where possible.

References

Giovanni Da, San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeño, Cede~ Cedeño, and Preslav Nakov. 2020. Prta: A System to Support the Analysis of Propaganda Techniques in the News. In *ACL 2020*.

RJ Deibert, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain. 2010. *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*. The MIT Press.

Larry Diamond and Marc F Plattner. 2012. *Liberation technology: Social media and the struggle for democracy*. JHU Press.

Ashrakat Elshehawy, Konstantin Gavras, Nikolay Marinov, Federico Nanni, and Harald Schoen. 2021. Illiberal Communication and Election Intervention during the Refugee Crisis in Germany. *Perspectives on Politics*, pages 1–19.

Sheena Chestnut Greitens. 2013. Authoritarianism online: What can we learn from internet data in nondemocracies? *PS: Political Science & Politics*, 46(2):262–270.

Sergei Guriev and Daniel Treisman. 2020. A theory of informational autocracy. *Journal of public economics*, 186:104158.

Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1960. How the voter makes up his mind in a presidential campaign. *The people's choice*.

Darren L Linvill and Patrick L Warren. 2020. Troll factories: Manufacturing specialized disinformation on twitter. *Political Communication*, 37(4):447–467.

Rebecca MacKinnon. 2011. Liberation technology: China's "networked authoritarianism". *Journal of democracy*, 22(2):32–46.

Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. Voynaslov: a data set of russian social media activity during the 2022 ukraine-russia war. *arXiv preprint arXiv:2205.12382*.

Margaret E Roberts. 2018. Censored. In *Censored*. Princeton University Press.

Sergey Sanovich, Denis Stukal, and Joshua A Tucker. 2018. Turning the virtual tables: Government strategies for addressing online opposition with an application to russia. *Comparative Politics*, 50(3):435–482.

Anton Shirikov. 2022. *How Propaganda Works: Political Biases and News Credibility in Autocracies*. Ph.D. thesis, The University of Wisconsin-Madison.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

Denis Stukal, Sergey Sanovich, Richard Bonneau, and Joshua A Tucker. 2017. Detecting bots on russian political twitter. *Big data*, 5(4):310–324.

Denis Stukal, Sergey Sanovich, Richard Bonneau, and Joshua A Tucker. 2022. Why botter: how pro-government bots fight opposition in russia. *American political science review*, 116(3):843–857.

Dmitri Teperik, Grigori Senkiv, Giorgio Bertolin, Kateryna Kononova, and Anton Dek. 2018. Virtual russian world in the baltics. *NATO StratCom COE*, 21.

Thomas Zerback and Florian Töpfl. 2022. Forged examples as disinformation: The biasing effects of political astroturfing comments on public opinion perceptions and how to prevent them. *Political Psychology*, 43(3):399–418.

A Appendix

Individual search items and topics

- Context
  - war with/on Ukraine, etc
  - special/military operation, etc
- Aims and Explanations
- War with NATO
- Economic worries
- Isolation from world
- Nuclear threat
- Society Polarisation
- Fake (фейк as a noun)
- Undermining trust in media
- Spoiler Crisis
  - Ukrainian crisis
  - other (gas, food, economic crisis)
- Spoiler Wars
- Phrases from Circulated Recommendations
- Dehumanisation
- Lack of Ukraine independence
- #nopanic
- Annexation
- Russian invasion
- Against mobilisation
- Evading the draft
  - fleeing the country
  - dodging call-up
- Ukraine military success