Exploring Russian Anti-War Discourse on Twitter during Russia's full-scale invasion of Ukraine: Dynamics, Influence, and Narratives

Iuliia Alieva[1][0000-0001-6270-8985] and Kathleen M Carley[1][0000-0002-6356-0238]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
ialieva@cmu.edu

Abstract

This paper explores the complexities of Russian anti-war discussions on Twitter following the invasion of Ukraine in 2022. Our research seeks to uncover the fundamental dynamics of this discourse within the Russian-speaking Twitter community, illuminating influential figures and interconnected communities, while closely examining the prevalence of bot activity. Through a comprehensive mixed-method approach, encompassing data collection via the Twitter API, social network analysis, bot identification, and community detection, alongside natural language processing, we analyze the narratives and dissemination strategies utilized by two key factions: the anti-war opposition and pro-government propagandists.

Introduction

In the lead-up to Russia's full-scale invasion of Ukraine in 2022, the political landscape of Russia was characterized by a hybrid regime, displaying a blend of democratic and autocratic elements (Petrov, Lipman, and Hale 2014). This period witnessed a quasi-balance between democratic ideals and authoritarian practices, with ruling elites continuously adjusting their strategies in response to shifting circumstances. Of particular scrutiny was the media sector, where the Russian government, faced with growing demand for opposition voices, maintained a mixed media environment comprising both state-controlled and independent outlets (Denisova 2017).

However, the onset of Russia's full-scale invasion of Ukraine in 2022 marked a pivotal moment, precipitating a series of restrictive measures that severely curtailed freedom of speech. This regulatory crackdown underscored a strategic shift in the government's approach, revealing a heightened sensitivity to perceived threats to the established regime and prompting a swift response from ruling elites.

In the face of escalating consequences, including imprisonment for any form of dissent, a burgeoning number of individuals turned to online platforms as a means to anonymously voice anti-war sentiments. Understanding the dissemination and impact of Russian anti-war discourse on social media assumes critical importance in elucidating public attitudes and willingness to protest amid government oppression.

Related Works

The role of social media is increasingly prominent in both national and international political arenas, especially within the realm of social movements, demonstrations, and protests. New digital media emerged as a tool for protest in 21st century Russia, notably during the protests in 2011-2012, when social media platforms played a significant role in organizing anti-government activities (Denisova 2016; Spaiser et al. 2017). Subsequently, Russian state propaganda intensified its online presence, disseminating their narratives through social media channels (Kiriya 2021; Linvill & Warren 2020; Mejias & Vokuev 2017). Due to strict government control over most traditional news outlets such as broadcast, radio, and newspapers in Russia (Kiriya 2021), individuals tend to express their criticisms online through open public platforms and social media. However, with the enactment of censorship laws in Russia after the 2024 invasion of Ukraine, criticizing the government or the war has become more perilous due to the risk of imprisonment.

Previous research indicates that state propaganda reduces protest rates, and its effects endure over time (Carter & Carter 2021). Pro-government propaganda can alter individuals' beliefs regarding the regime, the autocrat's performance, official national positions on global issues, and public discourse. Additionally, propaganda showcases power and authority; its mere use and manner of dissemination can signal an autocrat's inclination to use force: "Such propaganda aims not to 'brainwash' people with specific content praising the government but rather to signal the government's strength through the act of propaganda itself" (Huang 2015). In autocratic nations, social media is often viewed as a tool for liberation, providing uncensored spaces for disenfranchised voices to express dissent. However, social media can also serve as a tool of oppression, as autocratic regimes implement these

technologies to propagate their narratives (Spaiser et al. 2017). Furthermore, social media may aid in isolating and marginalizing opposition voices, which can be advantageous for autocratic governments (Denisova 2016).

Analysis of Twitter discourse related to the protests following the Duma and presidential elections in Russia in 2011-2012 revealed that pro-government users employed a variety of communication strategies to manipulate political discourse and marginalize oppositional voices on the platform (Spaiser et al. 2017). The recent advancements in algorithmic capabilities of social media platforms have facilitated the proliferation of computational propaganda tools, including bots and troll factories, allowing for the manipulation of online social discourse (Kiriya 2021; Linvill & Warren 2020; Mejias & Vokuev 2017). Computational propaganda involves the deliberate dissemination of particular narratives across social media platforms through the use of algorithms, automation, and human curation (Woolley & Howard, 2017). Russia's state influence operations have been detected in relation to global events, where trolls attempted to control discourse and promote Kremlin narratives, thereby fostering social discord (Badawy et al. 2019; Nyst & Monaco 2018; Linvill & Warren 2020). Trolls and bots disseminate vast amounts of false information and state propaganda narratives in multiple languages with the aim of polarizing society and promoting pro-Russia official stances. Furthermore, recent laws, regulations, and numerous cases of arrests for expressing anti-war sentiments online in Russia have deterred people from voicing critical opinions against the regime. It's crucial to grasp how Russian anti-war discourse spreads on social media and its effect, as it provides understanding of public sentiments and readiness to protest despite governmental suppression. This evolving landscape underscores the imperative of monitoring and comprehending the dynamics of digital narratives and dissent within an increasingly constrained political environment.

To unravel the impact of Russian anti-war discourse on Twitter, this study posits several research questions:

RQ1: How has the Russian anti-war discourse on Twitter evolved over time?

RQ2: To what extent are bot accounts involved in shaping the Russian anti-war Twitter discourse?

RQ3: Who are the most influential actors and communities driving the discourse?

RQ4: What are the prevailing narratives propagated by these actors and communities?

Data and Method

The study's data collection and methodologies follow a set of procedures established by the social cyber-security studies (Alieva and Carley 2021; Alieva, Moffitt, and Carley 2022; Alieva, Ng, and Carley 2022; Alieva, Robertson, and Carley 2023; Uyheng and Carley 2019). To address our research inquiries, we collected a dataset of tweets using the Python package *twarc* through an archive search with the most recent version of the Twitter API, version 2 and academic access.

For this study, the focus is on examining tweets in the Russian-language segment of Twitter to analyze the anti-war discourse related to the Russia's full-scale invasion of Ukraine. We collected tweets about anti-war discourse starting from February, 2022 till November, 2022. The keyword search includes the phrase "нет войне" and the similar hashtag #нетвойне ("no to war") in order to identify the anti-war tweets for our dataset. We converted the raw Twitter data into a meta-network consisting of user-to-user communication networks, user-to-tweet, and user to various tweet artifacts (hashtags, URLs) networks with ORA software to conduct network analysis (Carley 2014). A mixed-method approach was employed for data analysis, including quantitative and computational methods for data collection and data analysis, as well as qualitative investigation of tweets, users and narratives.

We used bot detection tool such as Bot-Hunter (Beskow and Carley 2018) to identify bot activities. Bot-Hunter is a tiered supervised machine learning approach for bot detection and characterization. The methodology also includes Twitter data network analysis to detect key actors and influencers, including super spreaders and super friends, along with Leiden clustering to identify communities within the network. We then identify topics with BERTopic for discussions by two prominent communities such as anti-war opposition community discourse and pro-government propaganda discourse. We also conducted a qualitative analysis of the most influential agents and communities in the network, their corresponding tweets, and narratives surrounding communities of agents.

For the analysis of the influencers, we used ORA software to compute network metrics and identify lists of the most influential accounts. Influencers are accounts whose tweets wield substantial influence on the social network, owing to their substantial follower count and strategic network position. The narratives they propagate hold the power to sway the opinions of fellow users within the network. Recognizing these pivotal influencers is essential for comprehending the potential adverse effects of information operations. ORA software is used to identify Twitter influencers based on computed network metrics for Twitter data and communication networks between users in the dataset. This study focuses on two specific categories: super spreaders (identified using a combination of network analysis metrics such as out-degree centrality, page rank centrality, and k-core) and super friends (determined through a combination of total degree centrality and k-core).

Super spreaders list includes users consistently producing content that is widely and efficiently circulated throughout the network. The super friends list encompasses users engaged in regular two-way communication, thereby playing a role in the development of extensive and resilient communication networks.

To identify network communities participating in the conversations on Twitter, we use Leiden clustering method. Leiden clustering algorithm involves network partitioning and node movement that guarantees well-connected communities. Leiden algorithm was proved to be more efficient than others, such as Louvain; it is also faster and uncovers better partitions (Traag, Waltman, and Van Eck 2019). After identifying the communities, qualitative methods were used to compare content and user characteristics between groups. We conduct influencer analysis on the largest Leiden communities, pinpointing the primary attitudes expressed by influencers within each group. Due to the substantial number of communities and actors, we opt to concentrate on the most influential users, given that their content enjoys widespread dissemination within each community. We also used Leiden community clustering to identify two groups for more detailed investigation and comparison such as anti-war opposition community and a community with highly influential pro-government propaganda accounts.

BERTopic modeling is employed to identify major topics for each of the two groups (anti-war opposition and pro-government propaganda communities) using the textual content of tweets. BERTopic has demonstrated superiority over other topic modeling strategies, such as LDA, specifically for short-text documents (de Groot, Aliannejadi, and Haas 2022). The BERTopic pipeline involves utilizing an embedding model, typically BERT-based, to generate vector representations of text. This is followed by a dimensionality reduction step and a clustering step to group similar documents (Grootendorst 2022). We used a multilingual embedding model, which allowed us to process Russian tweets. For this study, we use BERTopic to identify topics in each of two groups as well as narratives implemented by bots in each of the two groups.

Results

As a result of data collection, our dataset contains 657,548 tweets with 497,431 retweets and 163,205 users (agents) (see Figure 1 for details on tweets over time and bot/not bot users).

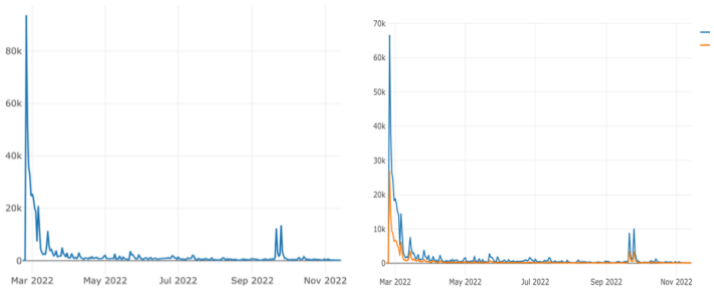

Figure 1: Number of tweets over time (on the left) and number of tweets from bots and not bots over time with orange line showing tweets from bots (on the right).

Figure 1 illustrates a significant decline in the number of tweets beginning in March-April, with two spikes observed around October 2022. These patterns are indicative of the blocking of Twitter in Russia in March 2022, rendering it more challenging to access this social network without a VPN. Additionally, they could be attributed to the restrictive laws enacted by the Russian government, which prohibit labeling Russia's invasion of Ukraine as a war and implement other censorship laws related to warfare, such as the law on the "discrediting" of the Russian armed forces, which criminalizes the dissemination of information deemed to be defamatory or damaging to the reputation of the Russian military. According to this law, any critical information about the Russian armed forces and its operations could result in punishment of up to fifteen years of imprisonment, potentially dissuading individuals from tweeting about the invasion. The spike observed in October could be associated with news regarding the Russian government's decision to declare a partial mobilization of individuals capable of participating in military combat. The number of tweets from bots reflects the trends in the overall discourse and aligns with the tweets from all users.

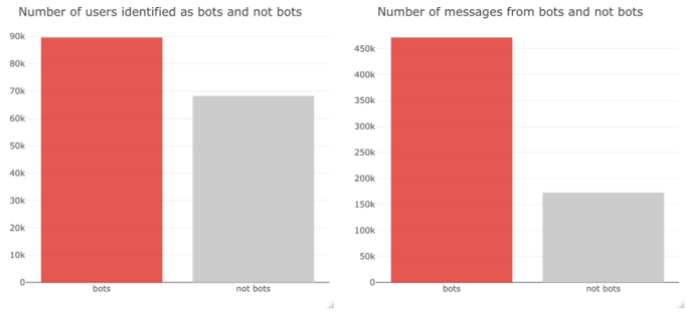

Figure 2: Number of users identified as bots/not bots (left) and the number of tweets posted from bots/not bots (right).

Figure 2 demonstrates the predominant presence of bots in this specific discussion. The quantity of users identified as bots surpasses that of non-bot users. Additionally, the number of messages posted by bots significantly exceeds the number of messages posted by non-bot users.

Twitter Influencers

In our dataset, various accounts and individuals emerged as top super spreaders, including members of the Russian opposition such as those associated with the Anti-Corruption Foundation, an organization founded by opposition leader Alexei Navalny. The list also comprised European politicians, human rights defense groups, news and opinion accounts, independent newsrooms, as well as anti-war activist groups.

Furthermore, we observed an anomaly among the top super spreader accounts mentioning the "No war" slogan alongside anti-war, pro-Ukraine, pro-opposition, and pro-Navalny accounts. Among these, there was an account ZavtraRu, which positions itself as a conservative and extreme-right newspaper in Russia, espousing ultranationalist and anti-capitalist views. Additionally, another similar account, @DenTvRu, was identified as a super spreader. Den TV is an extreme-right television media outlet with a YouTube channel and website promoting Russian state propaganda and conspiracy theories, sharing similar ultranationalist and anti-capitalist perspectives.

After analyzing the hashtags and tweets spread by various accounts, differences in narratives were identified between the anti-war channels and accounts such as Zavtra and Den TV. While anti-war channels used anti-war hashtag #nowar in combination with other similar hashtags such as Stop Putin and No war with Ukraine, accounts similar to Zavtra and Den TV used 'No war' hashtag in combination with #Z and similar propaganda hashtags such as Yes to the victory (#ДаПобеде), Glory to Russia (#СлаваРоссии), and other state propaganda narratives. After further investigation, extensive hashtag hijacking was identified in both pro-war propaganda and anti-war groups. Hashtag hijacking is a form of cyber content attack in which a hashtag is used for a purpose other than its original intent such as labeling messages with undesirable content and promoting this content to a target audience (Xanthopoulos et al. 2016). We visualized networks for hashtags used in the dataset and identified muptiple pro-government hashtags that were used in combination with anti-war hashtags (see Figure 3).

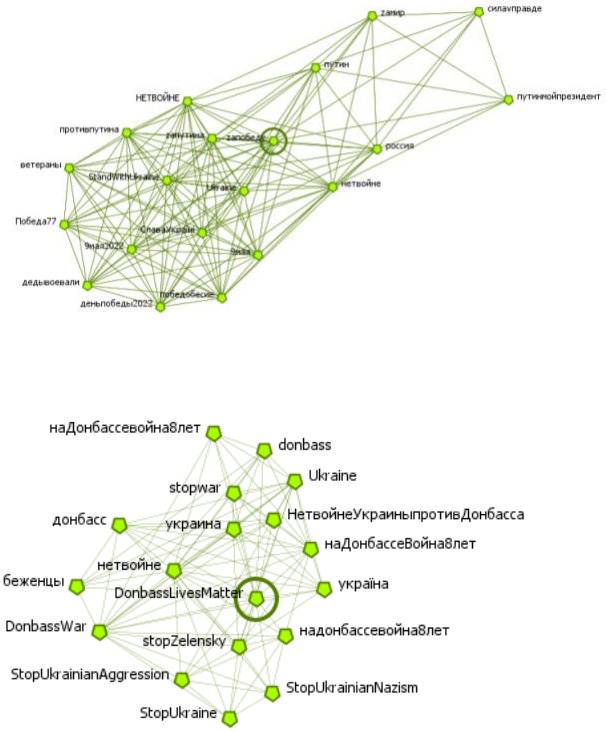

Figure 3: Examples of hashtag networks for two pro-government hashtags that are used in combination with anti-war hashtags.

Both groups were using hashtags from the opposite group for persuasion in order to promote their message (see Figures 4-5).

Figure 4: Examples of pro-government propaganda tweets found in anti-war discourse using anti-war hashtag

along with propaganda hashtags (with automatic Twitter translation).

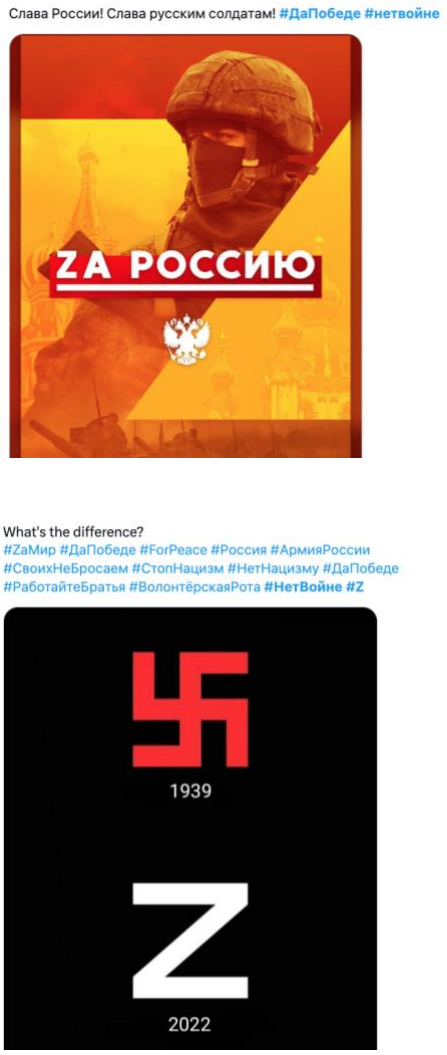

Figure 5: An example of pro-government Russian propaganda tweets using anti-war hashtag along with propaganda hashtags and an example of anti-war tweet using anti-war hashtags along with pro-government Russian propaganda hashtags.

Leiden clustering for community detection

After applying Leiden clustering to the entire communication network, the first ten groups were examined. The four largest groups demonstrated authentic anti-war sentiments and comprised of opposition leaders and independent media such as Alexey Navalny, Ilya Yashyn, MediaZona, and OVD Info, among others. However, groups #5 and #6 were predominantly composed of Russian politicians, state-funded media and influencers, including mfa_russia (Russia's ministry of foreign affairs), rian_ru (Russia's state news agency), and dumagovru (Russia's government). Group #7 featurs FBK and other opposition activists as top influencers, while group #8 consists of pro-Ukraine users and Ukrainian politicians. Group #9 includes Russian anti-war singers and performers as the main influencers. Lastly, group #10 also had anti-war accounts as the primary influencers.

As a result of Leiden clustering analysis, we selected two communities that represent discourse for two sides such as pro-opposition and anti-war accounts as well as a community that represents Russian pro-government pro-war discourse. In terms of composition of these two groups, we found that both of them have users identified as bots by our tool. For the opposition and anti-war group we found 56% of bot accounts in the community (with 75% of messages posted by bots). In government propaganda community, BotHunter found 60% of the group are bot users (with 72% of tweets posted by bots). This analysis demonstrates that both groups used bots to spread their messages and amplify their narratives.

Topic modeling with BERTopic

Among the main topics found by BERTopic in all tweets, the first one covered emotional language people expressed about war with all the fear, worrying and sadness about friends and relatives in both countries. The second one was mainly negative slurs about the Russian president, while the third one covers anti-war protests and police detentions of protesters. Other topics covered the concepts of fascism and Nazism (used often by both sides), as well as anti-war sentiments and news.

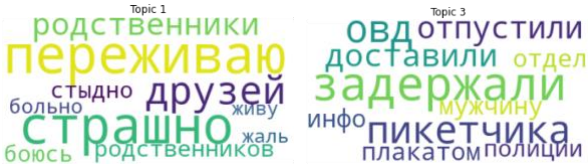

Figure 6. Examples of BERTopics for all tweets in the datasets.

We also checked topics disseminated by all bot accounts in the dataset and found that they mostly amplified pro-government state propaganda narratives about the predominant presence of fascism and Nazism in Ukraine, amplifying pro-government hashtags and support for the Russian military. However, we also found anti-Putin and anti-war topics amplified by bots as well.

Figure 7. Examples of BERTopics for all tweets from bots in the datasets.

For the topics in the opposition and anti-war group, we identified narratives about anti-war protests, statements by the opposition leader Alexey Navalny, news about protest activities in various regions in Russia. Bots in this community mostly amplified anti-war news and news from independent media related to anti-war activities and attitudes in Russia.

Figure 8: Examples of BERTopics for all tweets from opposition group in the datasets.

For the narratives in the pro-government accounts discourse, the main topics covered the war in the occupied territories of Ukraine including Donbas and Luhansk regions where pro-Russia separatists took over the power. These tweets mostly blamed Ukraine for the war and presenting Russia's so called special operation as liberation of these territories. These accounts used hashtag #nowar in the different context labeling invasion as a civil war in Ukraine between Ukrainian government (identified as Nazi regime by the pro-government propaganda accounts) and Eastern territories of Ukraine, claiming that Russia is preventing war and defending Russian speaking population. Their posts mostly mock the opposition and criticize anti-war opinions and influencers for the lack of support of the pro-government stances. Bots within this community amplify identical topics and attribute the suffering of the Donbas people to the disregard for the actions of the Ukrainian government by the global community.

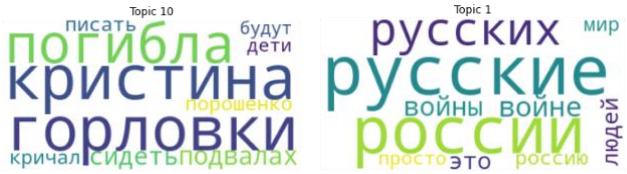

Figure 9: Examples of BERTopics for all tweets from pro-government group in the datasets.

Conclusion

Following our analysis of Russian anti-war discourse on Twitter, several significant insights were identified. We observed a notable decline in tweet activity, likely stemming from the impact of Twitter blockage in Russia and restrictive laws enacted by the Russian government regarding the portrayal of the war in Ukraine. Additionally, our investigation revealed a substantial presence of bots in the discussion, surpassing non-bot users in both quantity of users and messages posted.

Further exploration identified various accounts and individuals as top super spreaders, spanning from those associated with the Russian opposition to pro-government accounts. Analysis of communication networks using Leiden clustering unveiled distinct communities, including an opposition anti-war community and a pro-government community. BERTopic analysis highlighted prevalent topics, such as emotional expressions regarding the war and discussions on anti-war protests and police actions, as well as pro-war topics emphasizing Russia's pro-government narratives. The anti-war community shared messages condemning the conflict and criticizing the Russian government, while pro-government tweets propagated anti-Ukraine and anti-Western sentiments. Our research revealed that both communities employed bots to advance their respective agendas. Bots within the opposition group primarily amplified anti-war content and news from independent media outlets covering anti-war sentiments and activities in Russia. Conversely, bots associated with the pro-government community amplified anti-Ukraine narratives, including accusations against the Ukrainian government for crimes committed against people in Donbas and labeling the official Ukrainian government as a Nazi regime.

Overall, our comprehensive analysis provides insights on the intricate dynamics of online discourse surrounding the Russia's invasion of Ukraine, demonstrating the prevalence of diverse narratives and the influence of bots. This study underscores the scholarly importance of understanding the dynamics of the discourse and its influence on communication and persuasion within the current geopolitical landscape following Russia's full-scale invasion of Ukraine.

References

Alieva, I., and Carley, K. M. 2021. Internet trolls against Russian opposition: A case study analysis of Twitter disinformation campaigns against Alexei Navalny. In *2021 IEEE International Conference on Big Data (Big Data)*, 2461-2469.

Alieva, I., Moffitt, J. D., and Carley, K. M. 2022. How disinformation operations against Russian opposition leader Alexei Navalny influence the international audience on Twitter. *Social Network Analysis and Mining*, *12*(1), 80.

Alieva, I., Ng, L. H. X., and Carley, K. M. 2022. Investigating the Spread of Russian Disinformation about Biolabs in Ukraine on Twitter Using Social Network Analysis. In *2022 IEEE International Conference on Big Data - Big Data* (pp. 1770-1775). IEEE.

Alieva, I., Robertson, D., and Carley, K. M. 2023. Localizing COVID-19 Misinformation: A Case Study of Tracking Twitter Pandemic Narratives in Pennsylvania Using Computational Network Science. *Journal of Health Communication*, *28*(sup1), 76-85.

Badawy, A., Addawood, A., Lerman, K., and Ferrara, E. 2019. Characterizing the 2016 Russian IRA influence campaign. *Social Network Analysis and Mining*, *9*, 1-11.

Beskow, D. M., and Carley, K. M. 2018. Bot-hunter: a tiered approach to detecting & characterizing automated activity on Twitter. *SBP-BRiMS: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, vol. 3*.

Carley, K. M. 2014. ORA: A Toolkit for Dynamic Network Analysis and Visualization. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of Social Network Analysis and Mining*, Springer.

Carter, E. B., and Carter, B. L. 2021. Propaganda and protest in autocracies. *Journal of Conflict Resolution*, *65*(5), 919-949.

de Groot, M., Aliannejadi, M., and Haas, M. R. 2022. Experiments on generalizability of BERTopic on multi-domain short text. *arXiv preprint arXiv:2212.08459*.

Denisova, A. 2017. Democracy, protest and public sphere in Russia after the 2011–2012 anti-government protests: Digital media at stake. *Media, Culture & Society*, *39*(7), 976-994.

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

Petrov, N., Lipman, M., and Hale, H. E. 2014. Three dilemmas of hybrid regime governance: Russia from Putin to Putin. *Post-Soviet Affairs*, *30*(1), 1-26.

Huang, H. 2015. Propaganda as Signaling. *Comparative Politics*, 47(4), 419-44.

Kiriya, I. 2021. From "troll factories" to "littering the information space": Control strategies over the Russian internet. *Media and Communication*, *9*(4), 16-26.

Linvill, D. L., and Warren, P. L. 2020. Troll factories: Manufacturing specialized disinformation on Twitter. *Political Communication*, *37*(4), 447-467.

Mejias, U. A., and Vokuev, N. E. 2017. Disinformation and the media: the case of Russia and Ukraine. *Media, culture & society*, *39*(7), 1027-1042.

Nyst C., and Monaco N. 2018. How governments are deploying disinformation as part of broader digital harassment campaigns. *Institute for the Future*. https://t.ly/bkn-- .

Spaiser, V., Chadefaux, T., Donnay, K., Russmann, F., & Helbing, D. 2017. Communication power struggles on social media: A case study of the 2011–12 Russian protests. *Journal of Information Technology & Politics*, *14*(2), 132-153.

Traag, V. A., Waltman, L., and Van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports*, *9*(1), 1–12.

Uyheng, J., and Carley, K. M. 2019. Characterizing bot networks on Twitter: An empirical analysis of contentious issues in the Asia-Pacific. *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, Springer, pp. 153-162.

Woolley, S., and Howard, P. 2017. Computational propaganda worldwide: Executive summary. Oxford, UK: Project on Computational Propaganda. demtech.oii.ox.ac.uk.

Xanthopoulos, P., Panagopoulos, O. P., Bakamitsos, G. A., and Freudmann, E. 2016. Hashtag hijacking: What it is, why it happens and how to avoid it. *Journal of Digital & Social Media Marketing*, *3*(4), 353-362.