Crime Science

RESEARCH

Open Access

The fight against disinformation and its consequences: measuring the impact of "Russia state-affiliated media" on Twitter

Jesús C. Aguerri[1*], Mario Santisteban[2] and Fernando Miró-Llinares[1]

Abstract

On February 28th, shortly after the Russian invasion of Ukraine on February 24th, Twitter announced the expansion of its labelling policy for "Russia state-affiliated media", in order to address disinformation in favour of the Russian government.. While this 'soft' approach does not include the removal of content, it entails issues for freedom of expression and information. This article investigates the consequences of this labelling policy for the range and impact of accounts labelled "Russia state-affiliated media" during the Ukrainian war. Using an iterative detection method, a total of 90 accounts of both media outlets and individual journalists with this label were identified. The analysis of these accounts' information and timeline, as well as the comparison of the impact of their tweets before and after February 28th with an ARIMA model, strongly suggests, that this policy, despite its limited scope, could have contributed to a reduction in the impact of the sampled tweets, among other concurrent events. These results provide empirical evidence to guide critical reflection on this content moderation policy.

Keywords Disinformation, Information warfare, Twitter, Content moderation, Accounts labelling, Flagging

Introduction

The circulation of false information on social media and other digital platforms has been a major concern for states and international institutions for almost a decade (Bennett & Livingston, 2018; Lazer et al., 2018). Alarm about the possible consequences of such false or distorted content was first raised during the 2016 US presidential campaign (McGonagle, 2017; Mihailidis & Viotty, 2017) and Brexit (Bastos & Mercea, 2019), then under the ambiguous term fake news. More recently, the COVID-19 crisis revitalized this issue, foregrounding the risks that misinformation poses to public health and safety (Ahmed et al., 2020; Interpol, 2020). The latest milestone in this brief chronology is the Russian war in Ukraine, which is also fought in communication and information media. The military invasion is accompanied by information warfare (Kalniete, 2022), a communication strategy that combines disinformation, partial information, and certain narratives to influence public opinion.

Both social media platforms and national states have taken measures to react to this strategy. Such measures

*Correspondence:
Jesús C. Aguerri
j.aguerri@crimina.es
[1] CRÍMINA Research Center for the Study and Prevention of Crime, Miguel Hernández University of Elche, Avda. de la Universidad s/n. Edif. Hélike, 03201 Elche, Spain
[2] Department of Business Law and Civil Law, University of the Basque Country, Bilbao, Spain

range from the withholding of the Twitter[1] accounts of *Russia Today* and *Sputnik*—following a legal requirement from the European Commission- to the deletion of accounts that pursue information warfare operations. Alongside these initiatives, there are other—allegedly less incisive—measures to restrain users' speech that social media platforms implement to fight disinformation. One of them is to label accounts as *state-affiliated media*. While this measure had been put in place by Twitter since 2020, it was expanded quickly after the invasion of Ukraine, now affecting a larger number of accounts, including journalists from media outlets connected to the Russian government. Few studies have shown interest in the effects of such flagging policies on disinformation or the consequences for the accounts concerned. This article is an attempt at filling this gap, examining Twitter's labelling policy from both an empirical and a normative perspective.

The specific focus will lie on the consequences that these labels entail for affected accounts in terms of effective range. To achieve this, we have analyzed 90 accounts tagged as affiliated with the Russian government. These were identified through an iterative review of the followers of these accounts and the gradual identification of the networks these accounts form. After the detection, we proceeded to download both their associated information and timelines from January 9th to March 15th 2022. The results of our analyses allow us to guide the discussion on the impact of the policy in question based on both the perspective of freedom of information and expression, and evidence regarding its scope and efficacy.

Background

Information warfare

Russian disinformation campaigns through social media platforms have been identified across several states (Golovchenko, 2020; Khaldarova & Pantti, 2016; Magdin, 2020; Retchtik & Mareš, 2021). At the same time, propaganda and (geo)political struggle for the control of public opinion, through all kinds of resources, are as old as the states themselves (Levi, 2019). In the current communicative ecosystem dominated by social media platforms and digital media outlets, however, the dissemination of false information by certain agents and distrust of the veracity of the information that contradicts one's beliefs (Van der Linden et al., 2020) seem to have gained prominence (Lazer et al., 2018). The literature suggests that the Russian Federation did attempt to interfere in the 2016 United States Presidential Election, to help President Trump by discrediting Hilary Clinton (Álvarez, et al., 2020; ICA, 2017; McKay & Tenove, 2021). Further interference has been identified in the Brexit referendum, although its effect on the vote has not been sufficiently accounted for (Intelligence & Security Committee of Parliament, 2020). The lack of information in this regard is not limited to Brexit. Rather, the real-world consequences of misinformation are generally not sufficiently understood (Miró-Llinares and Aguerri, 2023). The limited empirical evidence implies a rather limited scope of those disinformation campaigns that were hitherto identified; both in general (Allcott & Gentzkow, 2017; Grinberg et al., 2019; Guess et al., 2019) and those that were allegedly led by Russian actors (Erlich & Garner, 2021; Hjorth & Adler-Nissen, 2019).

This notwithstanding, the European Union has stressed the need to challenge Russia's disinformation efforts at least since 2015 (European Council, 2015). The European Commission has also stated that disinformation campaigns are used by domestic and foreign actors to sow distrust and create societal tensions (European Commission, 2018). In June 2020, the European Parliament created a special committee on foreign interference in all democratic processes in the European Union, including disinformation. This Committee elaborated a report for the European Parliament in which it highlighted the capacity of social media platforms to reinforce cognitive biases and to interfere with civic decision-making. It also expressed concern over the role of social media in information warfare (Kalniete, 2022).

Foreign interference tactics may not solely rely on disinformation (Kalniete, 2022). When used to influence foreign actors, disinformation is often described as a part of a broader strategy that attempts to dominate the public opinion in a state. Information warfare is characterized by the strategic use of information and disinformation to achieve political and military goals (Thornton, 2015). This may include strategies and techniques of deception such as the use of deep fakes or other technical innovations (Chesney & Citron, 2019), present during the invasion of Ukraine by Russia (Gleicher, 2022b). The term has also been used to describe the large-scale use of destructive force against information assets and systems that support critical infrastructure (Lewis, 1997). Information warfare can be waged via cyber-attacks that affect critical infrastructure, the illicit funding of political parties that match foreign actors' interests or the use of 'traditional' state-owned media for propaganda (Intelligence & Security Committee of Parliament, 2020).

Russia disseminates content via multiple communication channels and actors (i.e., media outlets, podcasts,

[1] In the process of revision of this article Twitter underwent significant changes as a social media platform, one of them being its acquisition by Elon Musk and a name change (it is now called "X"). In order to maintain clarity and consistency with the key concepts of the article, the authors decided to keep the former name throughout the paper.

social media accounts) that, while inconsistent in their messaging, help strengthen the overall narrative of the Russian authorities (Paul & Mattews, 2016). In recent years, Russia has followed the model of outlets such as CNN and the BBC, opting to create and promote international media to confront the Western mass media narrative. In this context, these outlets have been accused of being part of Russia's communication strategy. While the role of the media is a key element of dissemination for the strategy of the Kremlin, the role of individual users shouldn't be overlooked. Such users may not be linked with Russian authorities directly but end up sharing their narrative through interaction on social media (Golovchenko et al., 2018).

Modern technology allows the dissemination of disinformation in several ways and social media grants hitherto unknown ways to interfere with public discourse. Most recently, evidence has pointed to the use of bots or automated accounts to foster media campaigns beyond dissemination by actual users (Beskow & Carley, 2020). In a similar vein, information campaigns often utilize multiple accounts controlled by a single person to act in a coordinated way, publishing the same message or helping spread it.

The responses to disinformation: Twitter's labeling policy on Russian affiliated media

If there are well known examples of the criminalization of disinformation (Khan, 2021) most states have remained cautious, adopting soft responses to tackle this issue. This is mostly due to concerns regarding the conflict of interest between measures against disinformation and civic rights such as freedom of expression (OSCE, 2017). Taking that into consideration, the approach of different actors, ranging from national states to platforms is to counterfeit disinformation with methods that do not imply content deletion (High Level Expert Group on Fake News and Online Disinformation, 2018; Eu Code of Practice in Disinformation, 2018; McCarthy, 2020).[2]

Social media platforms generally do not remove content on the basis of its falsehood, even though there are some exceptions in the case of deep fakes, disinformation shared during electoral periods and disinformation on public health issues (Twitter, 2022). Instead, platforms seek to reduce its impact by implementing measures without deleting content, what is being known as soft content moderation. They retain accounts that are suspicious of sharing disinformation (Twitter transparency, 2022) or attack labels to content to dispute its veracity or promote other reliable sources (Papakyriakopoulos and Goodman, 2022). Indeed, Twitter labelled promoted tweets and accounts in order to distinguish them from other content and make paid advertisements identifiable (Twitter, 2019). To tackle disinformation about Covid-19, Twitter introduced labels and warning messages under tweets that contained disputed or misleading information, too. These labels furthermore contain links to additional information on the claims in question. In the case of warnings, users are shown text to indicate that the displayed information contradicts what is stated by health authorities before they can access the tweet (Roeth & Pickles, 2020).

Nonetheless, Twitter's labelling policies affect individual users as well. Twitter started labelling individual users during the 2018 midterm US election, identifying accounts of candidates who qualified for the general election (Coyne, 2018). In 2020, Twitter initiated a labelling policy for accounts that were related to governments, a measure that was first initiated by YouTube in 2018. This policy covered accounts of key government officials, including foreign ministers, institutional entities, ambassadors, official spokespeople, and key diplomatic leaders. Besides accounts belonging to state-affiliated media entities, their editors-in-chief, and/or their senior staff and journalists were also labelled.

According to Twitter, the goal of this labelling policy was to provide users with context on the source of information, fostering informed judgement on the visualized content (Twitter Support, 2020). However, the platform has since decided that this labelling policy affects the overall audience that labelled accounts can reach (Twitter Support, 2020). Twitter announced that it wouldn't keep amplifying state-affiliated media accounts or their tweets through recommendation systems including the home timeline, notifications, and search.

Since February 28th, Twitter has expanded this policy, labelling more accounts that shared links to *Russia state-affiliated media* websites (Benson, 2022). This has resulted in new accounts being labelled as affiliated with the Russian government. As mentioned above, the label entails that tweets published from these accounts are not recommended in the home timeline, notifications, and other places on Twitter (Gleicher, 2022a). According to Twitter, the labelling policy has led to a 30% reduction in the reach of the content (McSweeney, 2022). But Twitter itself has not shared any information about the impact of

[2] This 'soft' European approach to disinformation has changed due to the conflict between Russia and Ukraine. The European Union has suspended the broadcasting activities of Russia Today (RT) and Sputnik, alleging that their propaganda constitutes a significant and direct threat to the Union's public order and security (Council Regulation (EU) 2022/350 of 1 March 2022, Recital 8). It can be argued that the law is respectful of freedom of expression and is thus a legitimate regulation of media (see Baade, 2022), matching the requirements that the CJUE has imposed on the Council to adopt these measures (CJUE, case T-262/15). However, all restrictive regulations of speech must be carefully scrutinized, analyzing whether its underlying objectives are legitimate in a democratic society and whether the chosen resources are necessary and suited to achieve these objectives.

the labels on state-affiliated media accounts, and Twitter's reasons for this decision are unclear.

The empirical impact of this policy change has not been described yet either. To the best of our knowledge, no study has measured the quantitative impact of this policy on affected accounts in terms of their scope. The majority of research has attempted to find out whether tagging news as disinformation increases disbelief in the affected sources (Mena, 2020; Pennycook et al., 2020). Other studies have measured the impact of state-affiliated media labels on users' distrust of the content published by the affected outlets (Nassetta & Gross, 2020).

If hard moderation remedies have been the main point of discussion regarding platforms power on controlling user speech, different scholars have been paying attention to other restrictions that do not entail content deletion but limiting the presence of content on news feeds or other content pools (Gillespie, 2022). Some authors have used the term "reduction" to refer to this content moderation measure (Gillespie, 2022) or "demotion" as some legal texts (Leerssen, 2023). Besides, these technique is often linked to the broad term "shadow banning", that refers to some content moderation decision that are not communicated to the public and can be only spotted indirectly (Le Merrer et al., 2021; Leerssen, 2023; Savolainen, 2022), and which existence has been generally disputed by platforms and scholars.

By contrast, there is numerous examples on content demotion or reduction that have been acknowledged by platforms. Generally, they use this method of soft moderation to target content that is problematic but that does not infringe its community guidelines (Gillespie, 2022), as disinformation or click bait. For instance, Facebook debunks disinformation on their feeds, content previously identify by fact checkers (Meta, 2022). Twitter stated in 2020, that some Covid-19 disinformation that could cause public harm "may not be shared on Twitter" (Roeth and Pickles, 2020). The state affiliated media policy joins now these examples, even if is not clear the type of problematic content that wants to address.

If content demotion does not pose the same level of harm to speech as hard moderation remedies, it is influence in the exercise of this liberty cannot be overlook. As Gillespie outlines, "reduction avoids none of the legal or societal problems that already haunt content removal: the presumptive power of the arbiter, the possible biases, the inequitable impact on different user communities, and the implications for free speech" (2022). Indeed, there is growing concern among institutions about the influence of social media algorithms on the content that users visualize (Council of Europe, 2012). Beyond the claims that filters shape users' serendipity, even if the empirical evidence is mixed in supporting the influence of social media algorithms in user polarization (Kubin & von Sikorski, 2021), it is clear that search engineers and social media algorithms play a considerable role in the content that is finally accessed by users (Ekström et al., 2022).

Current legal frameworks don't overlook this matter. As some authors have pointed out, shadow banning is forbidden by recent legal texts like the European Union's Digital Service Act (Leerssen, 2023). Under the umbrella of this regulation every restriction on user generated content shall be informed to users, including the demotion of content (art. 17.1). Furthermore, social media must consider the effects of the enforcement of their policies on fundamental rights (art. 14.4 of the DSA). In this sense, Article 19 of the International Covenant on Civil and Political Rights (Khan, 2021) and Article 10 of the European Convention of Human Rights allow for limitations to the right of freedom of expression to tackle disinformation (Pitruzzella & Pollicino, 2020). However, these limitations must comply with adequate democratic guarantees, such as the observance of the principles of proportionality and foreseability of the restrictive measure. Additionally, they must pursue a legitimate aim to be enforced, something that is lacking in the explanations of Twitter behind this policy.

Data and method

The aim of the study is to enrich policy debates regarding the Twitter labeling policy on the accounts labeled as *Russia state-affiliated media* with data on the effects of this policy on the affected accounts. To that end we compiled a list of Twitter 90 accounts labelled "Russia state-affiliated media"[3] by the platform. Since no public list of thusly labelled accounts is available, the list was constructed via a system similar to a snowball sampling: starting from one labelled account, we proceeded to review all the accounts followed by the first one, registering those labelled as "Russia state-affiliated media". We iterated the procedure on the resulting list of tagged accounts, and so forth recursively until no new tagged accounts were identified, our *saturation point*.

Twitter's population can be represented as a network of accounts following one another (Himelboim et al., 2017), creating multiple paths that link the different accounts both directly and indirectly, through followers and followers of followers respectively. The revision of followed allows us to navigate through the network thus reaching all nodes in the network in a limited number of iterations. However, it is not necessary here to traverse the entire network; Twitter is built through affinity relationships (of very different types) that are vertebrated

[3] The label name in English will be used, but the research also includes accounts in which this label appears in other languages. Specifically, this label was found in: English, Spanish, French, Russian and Arabic.

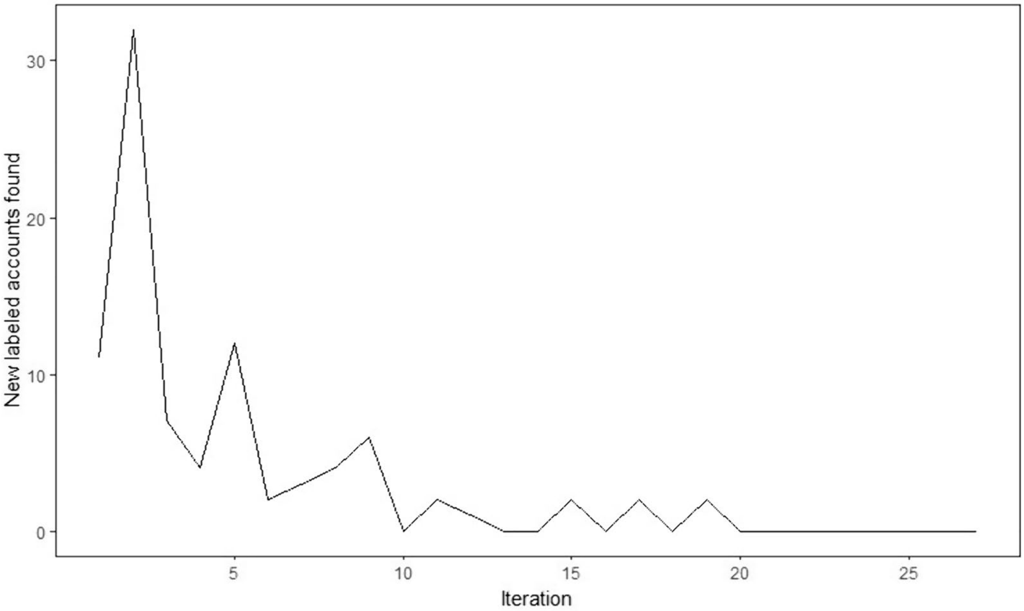

Fig. 1 New accounts detected by iteration

through "follows" relationships (Hahn, 2015). So, starting at one of the nodes, traversing all the paths that connect it with others and iterating the process, we can map out a sub-network of nodes matching our selection criteria. This procedure allows us to find all the accounts labelled by Twitter as "Russia state-affiliated media", under the assumption that no unconnected second sub-graph of tagged accounts exists. While the possibility cannot be excluded (see limitations), it doesn't seem plausible that such a second unrelated group of "Russia state-affiliated media" exists.

The saturation point was reached at the 20th iteration (Fig. 1). Seven additional iterations were carried out to verify that no new labelled accounts appeared. It was therefore not necessary to review the followers of all the sampled accounts, given that their number would have required an exceeding amount of manual work. Since the Twitter API does not reveal any information regarding account tagging, the review had to be performed by visual inspection of the accounts in question. Finally, we identified 90 tagged accounts across 27 iterations, conducted between March 11th and March 14th of 2022, having reviewed a total of 36,924 accounts.

Once these 90 accounts were identified, we used R software (R Core Team, 2020) and the AcademicTwitteR package (Barrie & Ho, 2021) to download account information and their timeline (tweets published and retweeted) from January 9th to March 15th, via the Twitter API v2. After assembling the database, several descriptive analyses were carried out to identify the characteristics and activity of the accounts. We selected the 19 days from the start of the invasion and the 19 days before as the time frame for descriptive analyses. Similarly to other studies, we measured the number of retweets received by the tagged accounts to have a representation of the impact that these accounts reached. Retweets are the tweets that a user reposts after seeing them in their timeline. These have been used to measure the interest of the public on certain content (Keib et al., 2018; Lee & Xu, 2018), existing a general assumption that the number of retweets affects the audience that a tweet can reach (Blankenship, 2018). In this study we used the number of retweets alongside the original tweets to measure this impact, even though original tweets are excluded in the predictive analyzes (see Fig. 5).

An ARIMA model was created (Hyndman & Athanasopoulos, 2018), taking the period from January 9th to February 28th as a baseline of comparison.[4] This point in time corresponds to Twitter's announcement that it was starting to tag personal accounts.

ARIMA models are one of the most common methods for time series analyses (Hyndman & Athanasopoulos, 2018). This method allows for predicting future values of variables from their past values. In this article, the predictions made by the model for the days following

[4] This period was selected to feed the ARIMA with more data points differing with the period select for the data analyses that was constrained to the 19 days prior and after the beginning of the invasion, that did not go beyond to avoid that other events unrelated to the study would cause oscillations that would have disturbed the sample.

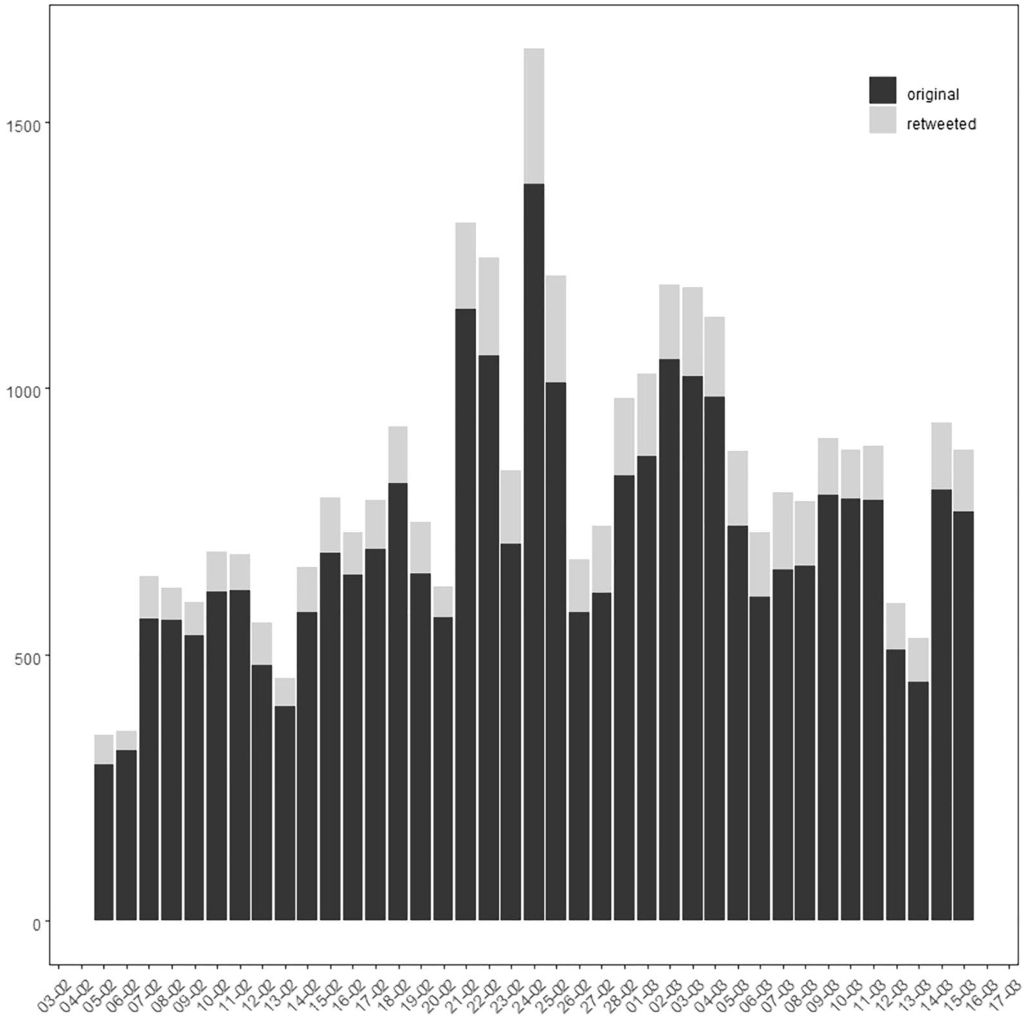

Fig. 2 Number of tweets published by day

February 28th will be taken as a reference to determine whether the change in Twitter's tagging policy has had a significant impact on the reach of tagged accounts. That is, we assume that the ARIMA model produces a 'prediction' of the impact that the analyzed tweets would have had, had Twitter not updated its policies. To build and select the model that best fits the data, we followed Kemp et al. (2021) and used the *auto.arima* function from the Forecast package (Hyndman et al., 2020) for R. This function allows for automatically searching for the best model by applying a variation of the Hyndman-Khandakar algorithm (Hyndman & Khandakar, 2008) to estimate univariate time-series. This algorithm constructs various ARIMA models using different estimates for the components p (the order of the auto-regressive model), d (the order of differencing) and q (the order of the moving average) and selects the one with the lowest AICc (a corrected version of the Akaike Information Criterion).

Finally, to ensure that the observed tendencies were not a consequence of a drastic shift of attention from some themes to others, the tweets were classified with Structured Topic Modelling (STM).[5] This allows us to detect the main themes of the tweets inside the database, classify those tweets considering the main theme that is presented in each of them, as well as stablishing the more common terms within a theme and the ones that are more representative, understood as the ones that are more likely to appear in a theme and not in the others. In this article we have used STM, in its version implemented in the stm package for R software (Roberts et al., 2019), to build a document classification algorithm that takes as parameters the text of the tweet and the day of

[5] Is relevant to outline, that, due to language limitations, the STM covered only tweets in English, thus, covering 21% of the content of the identified accounts (See limitations).

Table 1 Accounts descriptive statistics

	Minimum	Mean/Proportion	Median	Maximum
Creation date	2007-09-03	2012-12-03	2011-09-19	2021-05-25
Verified accounts		0.45 *(26)*		
Total tweets	92	67,429,3	10,571	676,105
Followers	198	210,425,64	23,430	2,998,731
Location (self-reported)				
Russia		0.43 *(25)*		
Germany		0.07 *(4)*		
France		0.05 *(3)*		
United Kingdom		0.07 *(4)*		
United States of America		0.17 *(10)*		
Spain		0.02 *(1)*		
Unknown		0.19 *(11)*		
Total accounts		58		

publication. To determine the number of topics, we took as reference the reduction of held-out likeliness and residuals, and the maximization of semantic coherence (Roberts et al., 2014), being 5 the number of topics that best balanced the three parameters.

Results

Accounts description

Table 1 shows 58 of the 90 identified accounts. The rest were discarded for analysis due to inactivity during the reference period (from February 5th to March 15th). Most accounts are more than 10 years old and exhibit a large number of tweets published (median = 10,571) as well as a high number of followers (median = 23,430). As for the geographical distribution of the users, it is observed that almost half of them report being located in Russia. Among the accounts with known locations, the second largest group is located in the USA, followed by a group of accounts located across Europe (UK, Germany, France and Spain). Despite the small number of locations, the tweets collected are written in 33 different languages. 68% are written in Russian and 21% in English. Spanish is the third most used language but is used in only 5% of the tweets.

Table 2 Accounts activity by period

	Median	Amount/Mean	Maximum
Peace period			
Original tweets		11,990	
Total tweets		13,666	
Retweets gotten		233,508	
Retweets gotten by tweet	3	19.5	8547
War period			
Original tweets		15,957	
Total tweets		18,640	
Retweets goten		544,467	
Retweets per tweet	9	34.1	5664

Overall impact

The evolution of the number of tweets published per day shows a peak on February 24th, the day of the invasion (Fig. 2). Quantitative analysis of the number of tweets published (Table 2) shows that activity remained high in the following weeks, with more tweets published than in the previous weeks. It should be noted that the accounts in our sample are mainly dedicated to publishing original content rather than retweets (Fig. 2). The main difference between the two periods of data collection is the increase in the impact of the messages published. The total number of retweets obtained after the invasion is higher than in the previous weeks. In addition, the number of retweets obtained by each tweet increased significantly.

Analysis of the daily evolution of the impact of the sampled accounts, as measured by the sum of the number of retweets obtained and tweets published, shows that the impact of these accounts is greater after the Russian invasion of Ukraine (Fig. 3). We can however appraise that the impact of these accounts during the lead-up to the invasion had already been increasing. Growth seems to slow down from February 28th onward and reverts to a decrease later on.

Consequences for journalists' accounts

It should be noted that 33 of the 58 sampled accounts correspond to media outlets ($n_{media} = 33$), which had previously been subject to Twitter's tagging policy. As was mentioned above, the fundamental change in the account tagging policy occurred on February 28th, when Twitter

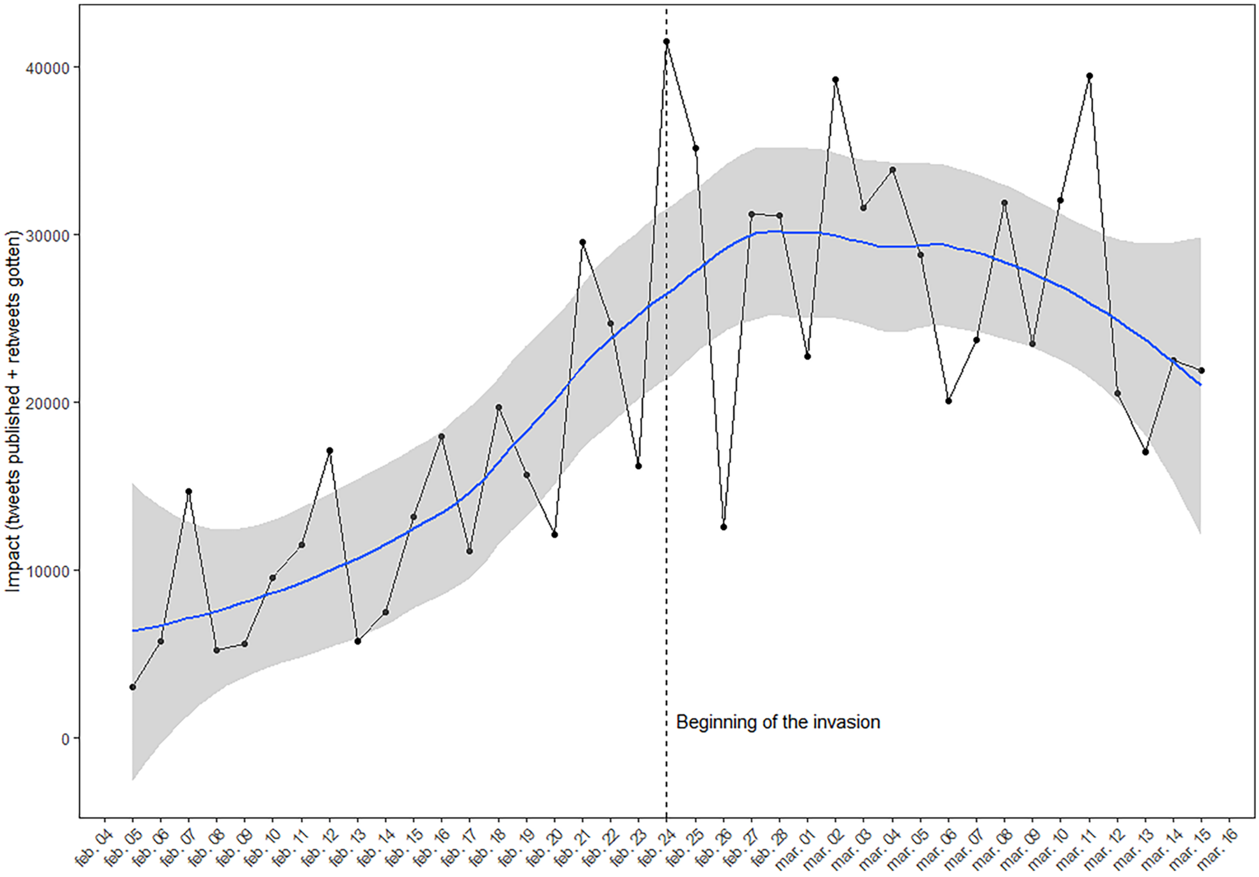

Fig. 3 Total impact by day (loess model, span = 0,6, degree = 1)

announced that the policy would be extended to users who "shared information" from those outlets. However, other events were relevant to the evolution of the impact of these accounts as well. On March 3rd, Twitter banned various accounts of RT and Sputnick media in Europe. On March 4th, the Russian government restricted access to Twitter in Russia. Figure 4 shows that the impact of the media accounts does not start to decrease until March 3rd or 4th, whereas journalists' personal accounts ($n_{journalists} = 25$) are negatively affected from February 28th onwards (Fig. 4). The evolution of the impact is consistent with the hypothesis that the impact of different accounts was affected by different events. Most notably, journalists' accounts and media outlets' accounts seem to have been selectively affected by the extension of Twitter's tagging policy on February 28th and the bilateral restriction of access to and from Russian media on Twitter on March 3rd and 4th respectively.

To verify the impact of Twitter's policies on the reach of journalists' accounts without conflating it with the effects produced by the restrictions imposed by the Russian Government, an ARIMA (2, 1, 1) was constructed to model the evolution of the sum of retweets gotten by journalists' accounts up to February 28th, excluding tweets in Russian. The observed values are only partially accounted for by the ARIMA prediction. We observe an interruption of the upward trend after February 28th, and a few days later the observed values start to be lower than the model's expectations (Fig. 5). On average, the model predicts 5306 retweets more than the ones that were finally obtained[6]. In consequence, it´s hard to explain these values only by the normal dynamics of the sampled accounts. This observation is consistent with the view that changes in Twitter's policy are related to an apparently slightly reduction of the impact of the sampled accounts and corresponding tweets.

In addition, the classification model built on the English-language tweets posted by journalists has allowed us to see that there are 5 themes running through these tweets—Table 3, among which themes 2 and 4 stand out. Both refer to Ukraine, but theme 2 seems to focus

[6] The difference between the real data and the data predicted by the model was statistically significant ($t = 2.97$, p-value < 0.05).

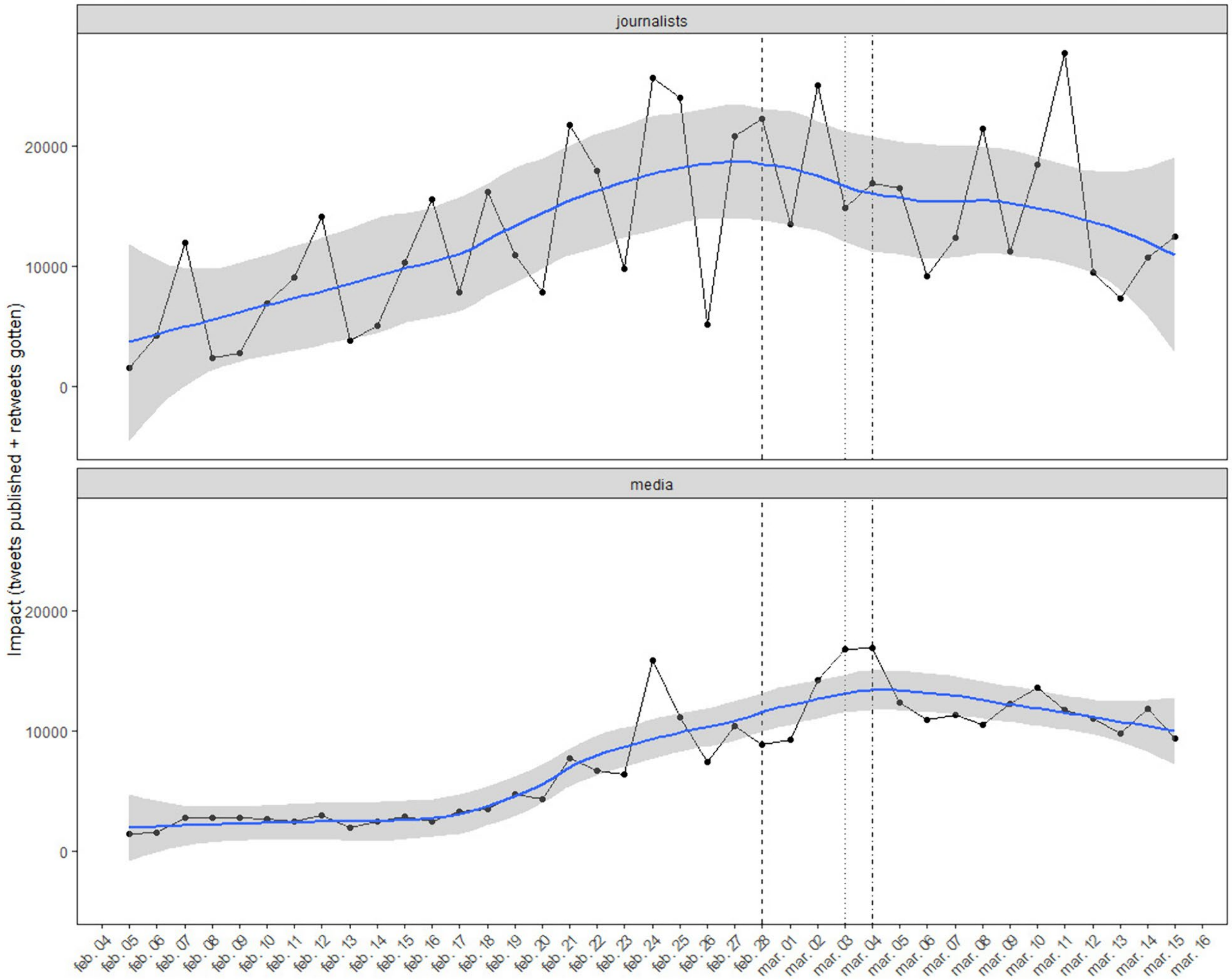

Fig. 4 Total retweets by day and type of account (loess model, span = 0.5). *Dashed line (28-02-2022): Twitter announce extension of labelling policy to journalists´ accounts; Dotted line (03-03-2022): Twitter withheld RT and Sputnick accounts in Europe; Dotted and dashed lines (04-02-2022): Rusia limits the access to Twitter in Russia. *Blue line represents the values predicted by local weighted regression (span = 0,6, degree = 1) fitted using the observed values. The gray area represents the regression confidence intervals

on NATO's actions, while theme 4 focuses more explicitly on the conflict in Ukraine. Topics 1 and 3 also have some relevance, with Topic 1 being primarily related to the media and containing, for example, messages criticizing the ban of certain outlets; while Topic 3 contains criticisms of the actions and measures taken by the USA and Europe. Lastly, Topic 5 gathers a small number of messages, which mainly ask for support or subscriptions on other platforms or the sharing of videos.

In none of the five topics (Fig. 6) is there a significant increase in the number of average retweets obtained by each tweet that could compensate for the fall in certain topics, which makes it difficult to consider that the changes in the impact of the tweets analyzed could be due to changes in the public's interest. This fact is especially relevant if we consider that topics such as 1 or 3 seem to refer to issues of particular current importance at that time, such as the restrictions imposed on certain media on those dates. Likewise, if we look at the evolution of the total impact per topic—Fig. 7—we can also see how from day 28 onwards the growing trend in all topic's stops.

Discussion

In view of our results, it would appear that the scope and effect of disinformation on Twitter are limited. However, Russia likely uses additional strategies to disseminate messages on social media platforms. But based on our sample data, one can hardly speak of a global disinformation network with the capacity to substantially influence

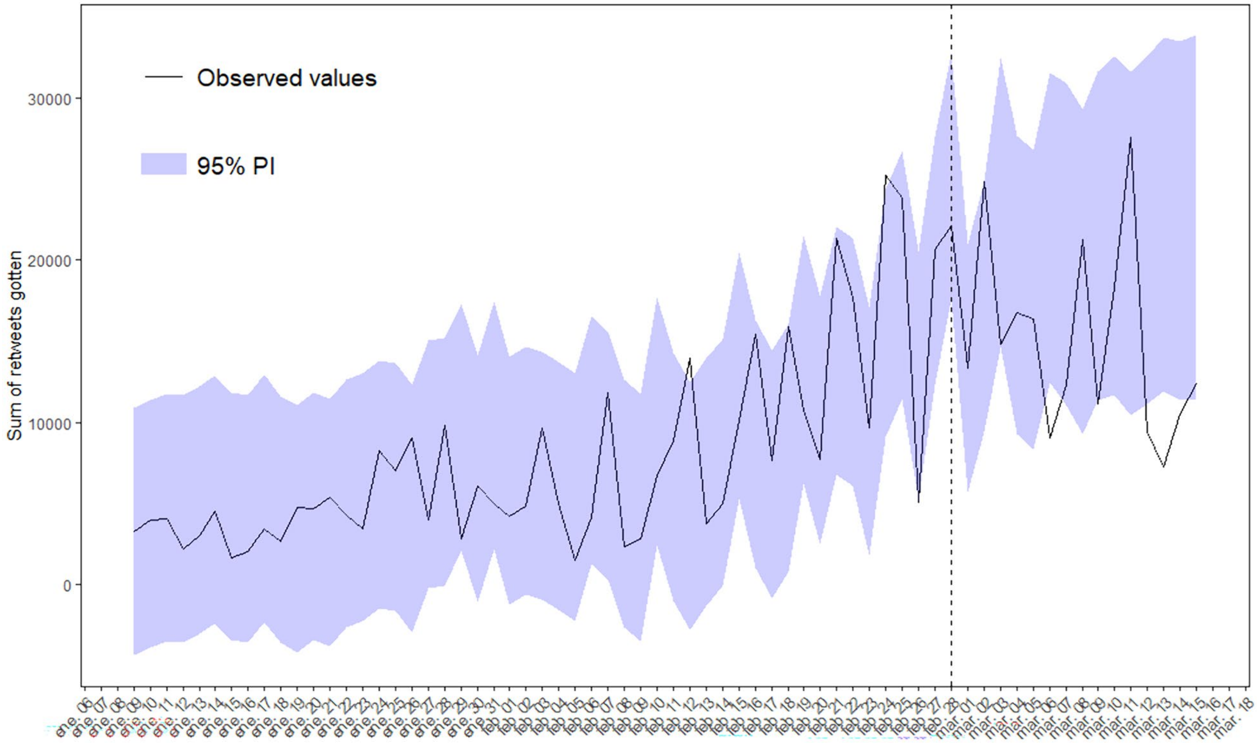

Fig. 5 Total impact of journalists' accounts (excluding tweets in Russian) and ARIMA's 95% Prediction intervals. *Dashed line (28-02-2022): Twitter announces extension of labelling policy to journalists' accounts

public opinion worldwide. Few accounts, while having a large number of followers, have Russian as their main language, meaning that their scope is mainly directed at users who understand this language. This is not to say that there are no networks run by the Russian government to benefit its socio-political interests. But it does allow us to affirm that Twitter's labelling policy has limited scope within information warfare.

The analysis of the sample showed considerable growth in both their activity and impact after the invasion. In fact, this growth could be observed during the lead-up to the invasion, and eventually slowed down during the week of February 28th, the date after which the growing trend reversed. This reversal may have been influenced by various events, such as Russian restrictions on the use of Twitter or the retention of the RT and Sputnick Twitter accounts in Europe. Our data even suggest that the reduction in reach was triggered by different political measures in the case of newly labelled journalists' accounts and established "Russia state-affiliated media" outlets such as Sputnick and RT. While the former saw a drop in reach after February 28th, the latter didn't lose prominence until March 3rd and 4th. We take this to be indicative of the dissociation of the two effects, suggesting that Twitter's policies are not only suitable to reduce the impact of tweets, but that different measures can be designed to target specific types of media agents (i.e., outlets vs. individual journalists).

Additionally, the results of the topic modeling applied to the corpus of tweets in English show that there are different themes in the conversation, ranging from the ban of Russian media outlets by the European Union and social media, the invasion of Ukraine and the Russia state affiliated media policy of Twitter. If these are diverse topics, the interruption of the growing tendency was found in all of them, suggesting that it was not caused by its content but external factors.

As Twitter has admitted, the platforms content moderation strategies go beyond the usual binary leave-up-take-down strategy (McSweeney, 2022). Between the measures Twitter can enforce we can find content reduction, which limits visibility and therefore impose a restriction on speech. Regardless of whether the platforms are legally bound by fundamental rights, one could argue that they have a responsibility to respect them (Jones, 2019). This commitment is held by platforms that claim adherence to and protection of human rights via the enforcement of their policies or by having signed the Guiding Principles on Business and Human Rights (Sissons, 2021). Organizations and researchers

Table 3 Topics of Tweets in English

Topic 1 Media	
Terms with highest frequency	go, rt, amp, media, ban, state, underground, countri, world, use
More representative terms	rt, underground, ban, full, state, use, go, interview, union, channel
*Example tweet**	*- China: The US has 336 labs in 30 countries under its control, including 26 in Ukraine alone. It should give a full account of its biological military activities at home and abroad and subject itself to multilateral verification.* https://t.co/oLnwPejksC *- YouTube has no problem hosting ISIS propaganda videos and promoting abusive content to children, yet it is blocking access to RT-related channels… The voices that have consistently brought you the coverage the mainstream media won't are now being wiped from the internet*
Proportion of tweets that have this topic as the main theme	0.15
Topic 2 N.A.TO.	
Terms with highest frequency	ukrain, us, nato, russian, russia, uk, new, report, nation, eu
More representative terms	uk, foreign, presid, new, report, nation, syria, nato, british, eu
*Example tweet**	- The Russian Foreign Ministry has called on Western media outlets to publish a full list of dates on which Russia will invade Ukraine for the year ahead, so Russian diplomats can schedule their vacations accordingly. This is not satire. They did this - US official Nuland—infamous for her part in the 2014 Ukraine coup which would kill 14,000 now admits the US has biowarfare labs on Russia's borders—real WMD not the Iraqi kind?! https://t.co/2YKPwB6owr
Proportion of tweets that have this topic as the main theme	0.43
Topic 3 Critics to measures	
Terms with highest frequency	amp, peopl, us, just, work, can, like, anti, want, get
More representative terms	watch, hate, imperi, anti, imperialist, dont, woke, work, know, make
*Example tweet**	*- My personal twitter account is NOT state-affiliated media. Nobody at RT or in Russia tells me what to tweet on this account. This is an attempt to discredit me & prevent people from hearing an anti imperialist message. Shame on you Twitter!* https://t.co/lqM8A33AAV *- I'm heartbroken. My RT America colleagues are some of the most incredible people I have had the privilege of working with. These are real people whose lives are being impacted. "Journalism is printing what someone else does not want published—everything else is public relations"*
Proportion of tweets that have this topic as the main theme	0.17
Topic 4 Invasion of Ukraine	
Terms with highest frequency	russia, amp, peopl, ukrain, ukrainian, year, nazi, donetsk, lugansk, russian
More representative terms	donetsk, lugansk, donbass, nazi, shell, year, republ, unsc, power, ukrainian
*Example tweet**	*-European universities in France, Belgium and the Czech Republic have begun expelling Russia students* -❗ *Russia says the final straw was President Zelensky's declaration of intent to restore Ukraine as a nuclear power. This, they say, sealed the invasion*
Proportion of tweets that have this topic as the main theme	0.24
Topic 5 Social media	
Terms with highest frequency	Will, time, talk, live, pm, thank, video, join, end, even
More representative terms	pm, video, time, talk, join, thank, will, tune, stream, movement
*Example tweet**	*- YouTube is straight up lying. They are removing entire tv channels, entertainment and documentaries, anti imperialist shows critical with Russia, past shows that ended long time ago and even a video services agency.* https://t.co/f5Lm4bW0e5 *- Thank you, thank you for all of the support this week! If you want to follow my work, please subscribe to my channel on Rokfin for exclusive content:* https://t.co/8oAVTEwxLu *Follow my Telegram page:* https://t.co/t9DbRdc4yy *Join my Telegram group:* https://t.co/Qjn2P69NfC *https://t.co/rs6lEHOA0Q*
Proportion of tweets that have this topic as the main theme	0.01

[*] Example tweets correspond to the tweet with the most retweets within the topic and with a gamma higher than 0.4 (which ensures a high coherence between the content of the tweet and the topic)

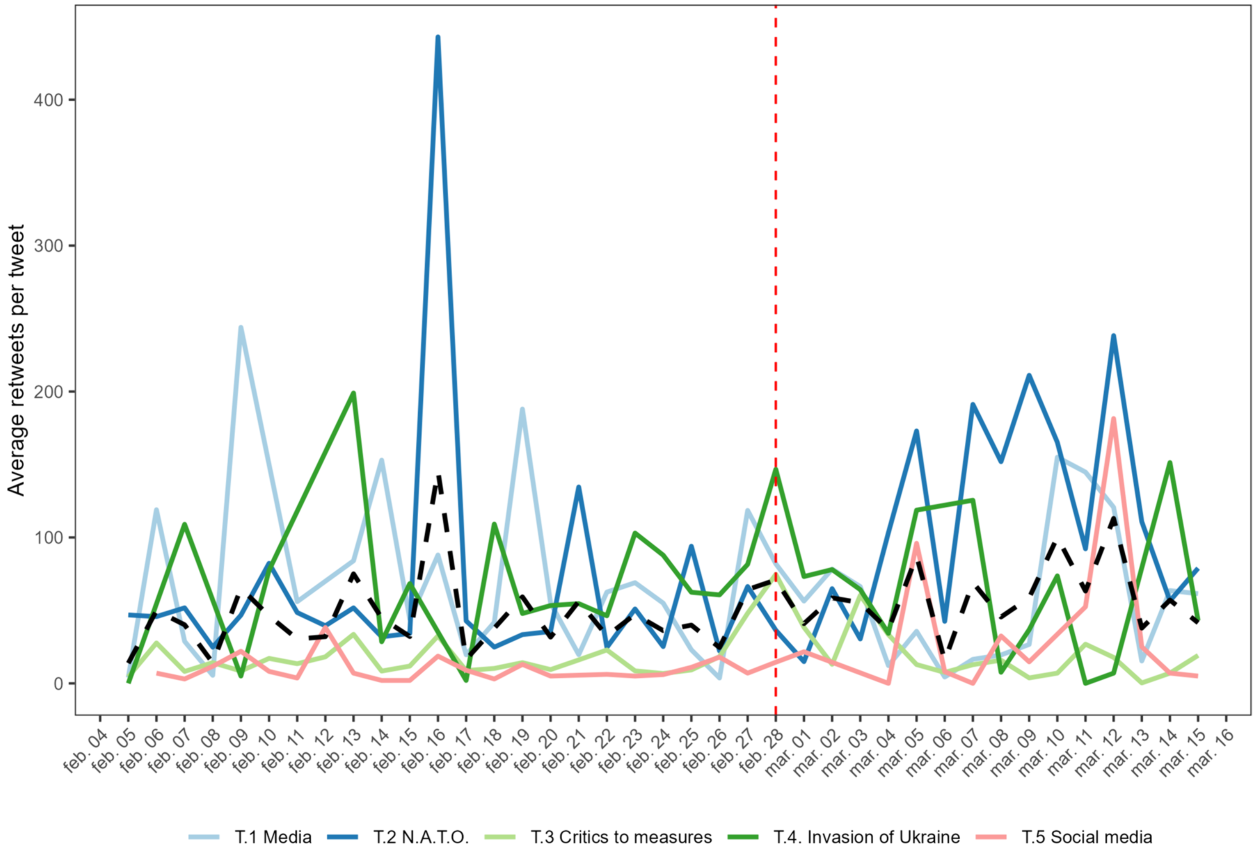

Fig. 6 Average retweets per tweets of journalists' accounts by topic (only tweets in English). *Dashed red line (28-02-2022): Twitter announces extension of labelling policy to journalists' accounts. * Blue line represents the average retweets gotten by all the topics

push social media to comply with a series of principles that ultimately seek to avoid arbitrariness (The Santa Clara Principles, 2022). In a similar vein, legislators in the European Union are moving towards limiting the powers of private actors to curate content and the Digital Service Act states that they must consider fundamental rights in the enforcement of their policies.

Content demotion could be justified under this legal framework. However, we cannot welcome the enforcement of Twitters Russia stated affiliated media policy cause the social media did not sufficiently justify this measure. Labelling accounts as affiliated with a particular government provides context for users to make informed decisions about what they read. This could reduce the impact of disinformation or propaganda, due to the contextualization provided by the additional information (Nassetta & Gross, 2020). In contrast, demoting news algorithmically does not provide additional context about what they see. It prevents users from accessing other sources of information. Demotion might be justified in the case of repeated engagement in disinformation and false information from a given account, which would have to be proven by fact checks. However, without such an ex-ante judgment, Twitter does not have sufficient evidence to justifiably claim that a given account is harmful or shows misleading content, and fighting misinformation, despite its importance, shouldn´t be used as a carte blanche.

In this sense, we believe that Twitter should make an effort to justify its change in policy and revise it in case no compelling arguments can be brought forth to maintain it. We also believe it would be appropriate for the social media platform to provide additional information, such as a list of affected accounts and reports on the performance of their tweets, enabling further assessment of the effects of this policy. This would further foster compliance with the platform's commitment to transparency, covering a wide range of other aspects such as information regarding the removal of content.

Limitations

This study presents several limitations that deserve detailed discussion.

Firstly, the selection of the sample of accounts. While sampling was transparently described and is reproducible, it is based on two premises: (i) the sampled accounts

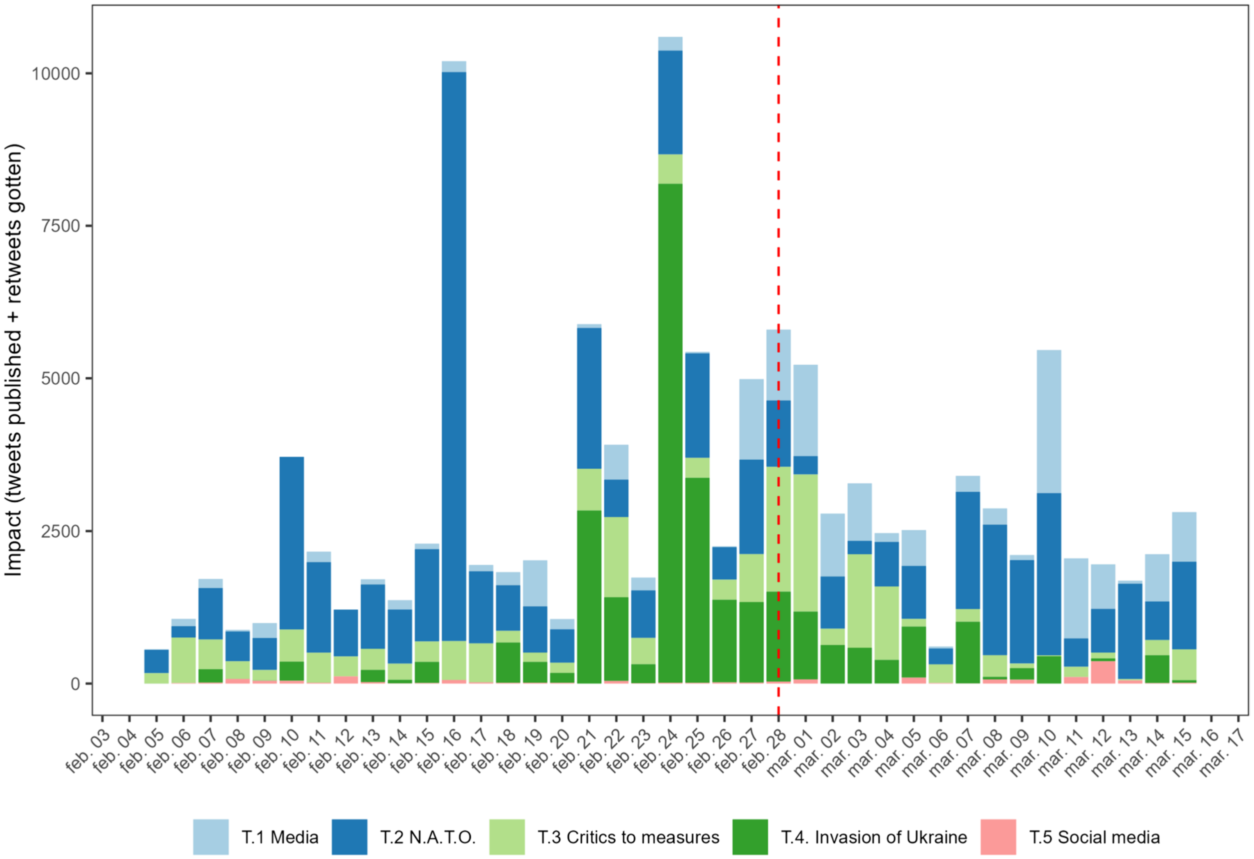

Fig. 7 Total impact by topic (only journalists' accounts and tweets in English). *Dashed red line (28-02-2022): Twitter announces extension of labelling policy to journalists' accounts

labelled "Russia state-affiliated media" have some kind of connection, albeit indirect, between them, and (ii) there can be no completely separate network (i.e., no independent subgraph) of accounts labelled "Russia state-affiliated media" that is unrelated to our sample. While there is no indication of evidence against these premises, they present the main caveat for our argument. If, on the other hand, these assumptions turned out to be false, the results obtained would continue to shed light on the consequences of Twitter's tagging policies, on the level of a specific network of users. Our findings would, however, be restricted to a very specific context and lose generalizability.

Furthermore, the ARIMA model's predictions are not enough to probe a causal relationship between Twitter's policy and the impact of the labelled accounts. Indeed, the consequences of Twitter's labelling policy from the withholding of Sputnick and RT accounts in Europe, as well as the restrictions imposed by the Russian Government might have had an influence in the results. The construction of the ARIMA model without consideration of tweets in Russian and the fact that the upward trend of the journalists' media impact (cf. comparison Fig. 4) slowed down before Twitter's media ban on Russian state media might dispute this claim.

It must be acknowledged as well that even if Twitter did announce the extension of its media affiliated accounts policy to journalists in February 28 there is no guarantee that the tagging of accounts did not start in previous days, which can compromise the validity of the findings. However, theme 3, which includes criticism of the West and also the denunciation of account tagging, has its peak in both average retweets and number of tweets on 28 February, which seems to indicate that a good part of the accounts were tagged on that day.

Acknowledgements
We thank the anonymous reviewers for their helpful comments and thanks Tobias Gretenkort for his help with the lengauge revision

Author contributions
All authors read and approved the final manuscript.

Funding
MS contribution to the article has been funded by the University of the Basque Country, as a result of the PIV UPV/EHU 2020 (https://www.ehu.eus/es/). JA contribution has been funded by the State Research Agency of Spain

(MCIN/AEI/https://doi.org/10.13039/501100011033) (https://www.aei.gob.es/en) and the European Union ("NextGenerationEU"/PRTR") through the grant FJC2020-042961-I. In addition, the research is part of the project #FakePenal (PID2021-125730OB-I00) funded MCIN/AEI/https://doi.org/10.13039/501100011033/FEDER, UE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

All data are available from: https://github.com/jcaguerri/disinformation

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 5 September 2022 Accepted: 30 May 2024
Published online: 15 June 2024

References

Ahmed, W., Vidal-Alaball, J., Downing, J., & López-Seguí, F. (2020). COVID-19 and the 5G conspiracy theory: Social network analysis of Twitter data. *Journal of Medical Internet Research, 22*(5), 1–9. https://doi.org/10.2196/19458

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Álvarez, G., Choi, J., & Strover, S. (2020). Good news, bad news: A sentiment analysis of the 2016 election Russian Facebook Ads. *International Journal of Comunication, 14*, 3027–3053.

Baade, B. (2022, March 8). The EU's "Ban" of RT and Sputnik: A lawful measure against propaganda for war. In: *Verf Blog*. https://doi.org/10.17176/20220308-121232-0

Barrie, C., & Ho, J. (2021). academictwitteR: An R package to access the Twitter Academic Research Product Track v2 API endpoint. *Journal of Open Source Software, 6*(62), 3272. https://doi.org/10.21105/joss.03272

Bastos, M. T., & Mercea, D. (2019). The Brexit botnet and user-generated hyperpartisan news. *Social Sciences Computer Review, 37*(1), 38–54. https://doi.org/10.1177/0894439317734157

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication, 33*(2), 122–139. https://doi.org/10.1177/0267323118760317

Benson, S. (2022, February 28). Twitter to label all state-affiliated Russia media. In: *Poltico* Retrieved June 5, 2024, from https://www.politico.com/news/2022/02/28/twitter-label-state-affiliated-russia-media-00012351

Beskow, D. M., & Carley, K. M. (2020). Characterization and comparison of Russian and Chinese disinformation campaigns, at disinformation, misinformation, and fake news. In K. Shu (Ed.), *Social media emerging research challenges and opportunities*. Springer.

Blankenship, E. B., Goff, M. E., Yin, J., Tse, Z. T. H., Fu, K. W., Liang, H., Saroha, N., & Fung, I. C. (2018). Sentiment, contents, and retweets: A study of two vaccine-related twitter datasets. *The Permanente Journal, 22*, 17–138. https://doi.org/10.7812/TPP/17-138

Chesney, R. & Citron, D. (2019). Deepfakes and the new disinformation war. In: F*oreign affairs*. Retrieved June 5, 2024, from https://perma.cc/TW6Z-Q97D.

Council of Europe. (2012). Recommendation CM/Rec(2012)3 of the Committee of Ministers to member States on the protection of human rights with regard to search engines, adopted by the Committee of Ministers on 4 April 2012. Council of Europe. Retrieved June 5, 2024, from https://www.coe.int/en/web/freedom-expression/committee-of-ministersadopted-texts/-/asset_publisher/aDXmrol0vvsU/content/recommendation-cmrec-2012-3-of-the-committee-of-ministers-to-member-states-on-the-protectionof-human-rights-with-regard-to-search-engines-adopted-by

Coyne, B. (2018, May 23). Introducing US Election Labels for Midterm Candidates. In: *Twitter Blog*. https://blog.twitter.com/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates

Ekström, A. G., Niehorster, D. C., & Olsson, E. J. (2022). Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports, 7*, 100226. https://doi.org/10.1016/j.chbr.2022.100226

Erlich, A., & Garner, C. (2021). Is pro-Kremlin Disinformation? Effective evidence from ukraine. *The International Journal of Press/politics*. https://doi.org/10.1177/19401612211045221

EU Code of Practice on Disinformation. (2018). Retrieved June 5, 2024, from https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

European Council (2015, March 19, 20). *European Council conclusions*. Retrieved June 5, 2024, from https://www.consilium.europa.eu/media/21888/european-council-conclusions-19-20-march-2015-en.pdfAccessed 5 Jun 2024.

European Commission. (2018). *Tackling online disinformation: a European approach*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0236&from=EN.Accessed 5 Jun 2024.

Gillespie, T. (2022). Do not recommend? Reduction as a form of content moderation. *Social Media + Society*. https://doi.org/10.1177/20563051221117552

Gleicher, N. (2022a). Updates on our security work in Ukraine. In: *About Meta*. Retrieved June 5, 2024, fromhttps://about.fb.com/news/2022/02/security-updates-ukraine/.

Gleicher, N. (2022b). "1/Earlier today, our teams identified and removed a deepfake video claiming to show President Zelensky issuing a statement he never did. It appeared on a reportedly compromised website and then started showing across the internet". In: *Twitter*. Retrieved June 5, 2024, from https://twitter.com/ngleicher/status/1504186935291506693?s=20&t=J_r9eb3j_y1SE2-blk2tAQ.

Golovchenko, Y. (2020). Measuring the scope of pro-Kremlin disinformation on Twitter. *Humanities and Social Science Communications, 7*, 176. https://doi.org/10.1057/s41599-020-00659-9

Golovchenko, Y., Hartmann, M., & Adler-Nissen, R. (2018). State, media and civil society in the information warfare over Ukraine: Citizen curators of digital disinformation. *International Affairs, 94*(5), 975–994. https://doi.org/10.1093/ia/iiy148

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*, 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), 1–8. https://doi.org/10.1126/sciadv.aau4586

Hahn, K. S., Ryu, S., & Park, S. (2015). Fragmentation in the Twitter following of news outlets: The representation of south korean users' ideological and generational cleavage. *Journalism & Mass Communication Quarterly, 92*(1), 56–76. https://doi.org/10.1177/1077699014559499

High Level Expert Group on Fake News & Online Disinformation. (2018). *A multi-dimensional approach to disinformation*. Retrieved June 5, 2024, from https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-be1d-01aa75ed71a1/language-en

Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying twitter topic-networks using social network analysis. *Social Media + Society*. https://doi.org/10.1177/2056305117691545

Hjorth, F., & Adler-Nissen, R. (2019). Ideological asymmetry in the reach of pro-Russian digital disinformation to united states audiences. *Journal of Communication, 69*(2), 168–192. https://doi.org/10.1093/joc/jqz006

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F. (2020). *forecast: Forecasting functions for time series and linear models (R package version 8.13)*. https://pkg.robjhyndman.com/forecast/.Accessed 5 Jun 2024.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). OTexts.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast Package for R. *Journal of Statistical Software, 27*(1), 1–22. https://doi.org/10.18637/jss.v027.i03

ICA. (2017). *Assessing Russian activities and intentions in recent US elections*. https://www.dni.gov/files/documents/ICA_2017_01.pdf

Intelligence and Security Committee of Parliament. (2020). *Report Presented to Parliament pursuant to section 3 of the Justice and Security Act 2013*. Retrieved June 5, 2024, from https://isc.independent.gov.uk/wp-conte nt/uploads/2021/03/CCS207_CCS0221966010-001_Russia-Report-v02-Web_Accessible.pdf.

Interpol. (2020). *Cybercrime: COVID-19 impact*. Interpol.

Jones K. (2019). *Online Disinformation and Political Discourse Applying a Human Rights Framework. Chatham House*. Retrieved June 5, 2024, from https://www.chathamhouse.org/2019/11/online-disinformation-and-political-discourse-applying-human-rights-framework

Kalniete, S. (2022). *REPORT on foreign interference in all democratic processes in the European Union, including disinformation (2020/2268(INI))*.

Keib, K., Himelboim, I., & Han, J.-Y. (2018). Important tweets matter: Predicting retweets in the #BlackLivesMatter talk on twitter. *Computers in Human Behavior, 85*, 106–115. https://doi.org/10.1016/j.chb.2018.03.025

Kemp, S., Buil-Gil, D., Moneva, A., Miró-Llinares, F., & Díaz-Castaño, N. (2021). Empty streets, busy internet: A time-series analysis of cybercrime and fraud trends during COVID-19. *Journal of Contemporary Criminal Justice, 37*(4), 480–501. https://doi.org/10.1177/10439862211027986

Khaldarova, I., & Pantti, M. (2016). Fake news: The narrative battle over the Ukrainian conflict. *Journalism Practice, 10*(7), 891–901. https://doi.org/10.1080/17512786.2016.1163237

Khan, I., (2021). Disinformation and freedom of opinion and expression. *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression. A/HRC/47/25*. Retrieved June 5, 2024, from https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/085/64/PDF/G2108564.pdf?OpenElement.

Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association, 45*(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, J., & Xu, W. (2018). The more attacks, the more retweets: Trump's and Clinton's agenda setting on Twitter. Public Relations Review, 44(2), 201–213. https://doi.org/10.1016/j.pubrev.2017.10.002

Leersen, P. (2023). An end to shadow banning? Transparency rights in the Digital Services Act between content moderation and curation. *Computer Law and Security Review, 48*, 1–13.

Le Merrer, E., Morgan, B. & Trédan, G. (2021). Setting the record straighter on shadow banning. In *IEEE INFOCOM2021-IEEE Conference on Computer Communications* (pp. 1–10).

Levi, S. (2019). *#FakeYou: Fake news y desinformación [#Fake you: Fake news and disinformation]*. Rayo Verde.

Lewis, B. C. (1997). Information warfare and the intelligence community. In E. Cheng & D. C. Snyder (Eds.), *The final report of the Snyder Commission, Woodrow Wilson School policy conference 401a: Intelligence reform in the post-cold war era*. Princeton Unversity.

Magdin, R. (2020). Disinformation campaigns in the European Union: Lessons learned from the 2019 European Elections and 2020 Covid-19 infodemic in Romania. *Rominian Journals of European Affairs, 20*(2), 49–61.

McCarthy, T. (2020, May 28). Zuckerberg says Facebook won't be 'arbiters of truth' after Trump threat. *The Guardian*. Retrieved June 5, 2024, from https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump.

McGonagle, T. (2017). 'Fake news': False fears or real concerns? *Netherlands Quarterly of Human Rights, 35*(4), 203–209. https://doi.org/10.1177/0924051917738685

McKay, S., & Tenove, C. (2021). Disinformation as a threat to deliberative democracy. *Political Research Quarterly, 74*(3), 703–717. https://doi.org/10.1177/1065912920938143

McSweeney, S. (2022, March 16). Our ongoing approach to the war in Ukraine. *Twitter blog*. Retrieved June 5, 2024, from https://blog.twitter.com/en_us/topics/company/2022/our-ongoing-approach-to-the-war-in-ukraine.

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy and Internet*. https://doi.org/10.1002/poi3.214

Meta. (2022). Our approach to misinformation. Meta Transparency Center. Retrieved June 5, 2024, from https://transparency.fb.com/es-es/features/approach-to-misinformation/.

Mihailidis, P., & Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in 'post-fact' society. *American Behavioral Scientist, 61*(4), 441–454. https://doi.org/10.1177/0002764217701217

Miró-Llinares, F., & Aguerri, J. C. (2023). Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'. *European Journal of Criminology, 20*(1), 356–374. https://doi.org/10.1177/1477370821994059

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreign misinformation. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-45

OSCE. (2017, March 3). Joint declaration on freedom of expression and "fake news", disinformation and propaganda. Retrieved June 5, 2024, from https://www.osce.org/files/f/documents/6/8/302796.pdf.

Papakyriakopoulos, O. & Goodman. E. (2022). The Impact of Twitter Labels on Misinformation Spread and User Engagement: Lessons from Trump's Election Tweets. In Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3485447.3512126

Paul, C., & Matthews, M. (2016). *The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter it*. RAND Corporation.

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science, 66*(11), 4944–4957. https://doi.org/10.1287/mnsc.2019.347

Pitruzzella, G., & Pollicino, O. (2020). *Disinformation and hate speech: A European constitutional perspective*. BUP.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved June 5, 2024, from https://www.r-project.org/.

Retchtik, M., & Mareš, M. (2021). Russian disinformation threat: Comparative case study of czech and slovak approaches. *Journal of Cmparative Politics, 14*(1), 4–19.

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software, 91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., & Rand, D. G. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.

Roeth, Y., & Pickles, N. (2020). Updating our approach to misleading information. Retrieved June 5, 2024, from https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*. https://doi.org/10.1177/01634437221077174

Sissons, M (2021, March 16). Our commitment to human rights. In: *About Meta*. Retrieved June 5, 2024, from https://about.fb.com/news/2021/03/our-commitment-to-human-rights/.

The Santa Clara Principles. (2022). The Santa Clara principles on transparency and accountability on content moderation. Retrieved June 5, 2024, from https://santaclaraprinciples.org/.

Thornton, R. (2015). The changing nature of modern warfare. *The RUSI Journal, 160*(4), 40–48. https://doi.org/10.1080/03071847.2015.1079047

Twitter. (2019). Twitter progress report: Code of practice on disinformation. Retrieved June 5, 2024, from https://ec.europa.eu/information_society/newsroom/image/document/2019-5/twitter_progress_report_on_code_of_practice_on_disinformation_CF162219-992A-B56C-06126A9E7612E13D_56993.pdf.

Twitter. (2022). The twitter rules. Retrieved June 5, 2024, from https://help.twitter.com/en/rules-and-policies/twitter-rules.

Twitter Support. (2020, August 6). New labels for government and state-affiliated media accounts. In: *Twitter Blog*. Retrieved June 5, 2024, from https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.

Twitter Transparency. (2022). *Transparency*. Retrieved June 5, 2024, from https://transparency.twitter.com/en/reports/information-operations.html.

Van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: Political bias in perceptions of fake news. *Media, Culture & Socierty*. https://doi.org/10.1177/0163443720906992

Publisher's Note