Geissler et al. *EPJ Data Science* (2023) 12:35
https://doi.org/10.1140/epjds/s13688-023-00414-5

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Russian propaganda on social media during the 2022 invasion of Ukraine

Dominique Geissler[1,2*], Dominik Bär[1,2], Nicolas Pröllochs[3] and Stefan Feuerriegel[1,2]

*Correspondence:
d.geissler@lmu.de
[1]LMU Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
Full list of author information is available at the end of the article

Abstract

The Russian invasion of Ukraine in February 2022 was accompanied by practices of information warfare, yet existing evidence is largely anecdotal while large-scale empirical evidence is lacking. Here, we analyze the spread of pro-Russian support on social media. For this, we collected $N = 349{,}455$ messages from Twitter with pro-Russian support. Our findings suggest that pro-Russian messages received ∼251,000 retweets and thereby reached around 14.4 million users. We further provide evidence that bots played a disproportionate role in the dissemination of pro-Russian messages and amplified its proliferation in early-stage diffusion. Countries that abstained from voting on the United Nations Resolution ES-11/1 such as India, South Africa, and Pakistan showed pronounced activity of bots. Overall, 20.28% of the spreaders are classified as bots, most of which were created at the beginning of the invasion. Together, our findings suggest the presence of a large-scale Russian propaganda campaign on social media and highlight the new threats to society that originate from it. Our results also suggest that curbing bots may be an effective strategy to mitigate such campaigns.

Keywords: Social media; Online spreading; Propaganda; Bots; Russo-Ukraine war

1 Main

On February 24, 2022, Russia invaded Ukraine [1, 2], thereby escalating the Russo-Ukrainian war that began with the annexation of Crimea in 2014 [3]. As of now, the war has led to a major energy crisis [4], global food shortages [5], and one of the largest refugee crises with more than 7 million Ukrainian refugees [6]. The invasion was later deplored by the United Nations (UN) General Assembly, with 141 countries approving Resolution ES-11/1, 5 countries voting against (e.g., Belarus, North Korea), and 35 countries abstaining (e.g., India, South Africa, and Pakistan) [7].

A widespread concern is that practices of modern warfare in form of large-scale Russian propaganda campaigns are used to shape the narrative around the war, yet corresponding research is still nascent. On the one hand, the Russian government enforced new legislation exerting power over traditional media outlets to persuade citizens to support the war. As a result, domestic media outlets are forced to adopt the official narrative [8–10]. On the other hand, Russian propaganda has been suspected to influence other countries outside Russia, in particular, by using social media to promote hostility against the West. Here,

one goal could be to diminish the support for sanctions against Russia and to weaken the support for Ukraine, especially in countries that have abstained from approving the United Nations Resolution ES-11/1 deploring the invasion. However, evidence of Russian propaganda campaigns from the 2022 invasion of Ukraine is so far purely anecdotal, whereas rigorous empirical evidence is missing.

Russian propaganda has been documented in several Western countries during previous conflicts [11, 12]. Oftentimes, the underlying narratives are recycled from past propaganda campaigns [13, 14] and aim to destabilize democratic countries by sowing doubt and polarizing citizens [13]. With the rise of the Internet, propaganda campaigns increasingly make use of social media. This gives rise to growing concerns that social media may be strategically used to increase political division and influence public opinion as a tool of modern warfare [15–18]. For example, a coordinated social media campaign was launched by a Russian organization known as the Internet Research Agency (IRA) during the 2014 Russo-Ukrainian conflict [16, 19]. The IRA has also been suspected of meddling in several elections. Among others, the IRA aimed to influence the outcomes of the 2016 U.S. presidential election [20–26], even though the influence on voting behavior has been questioned [27]. Other examples of foreign influence operations through the IRA are, e.g., the U.K. Brexit Referendum [28], and the 2017 French presidential election [29]. Yet, the aforementioned works focus on historical tactics of the IRA, while it is likely that the tactics of Russian foreign influence operations have become more refined over time. For example, in 2016, the IRA primarily employed trolls (rather than automated accounts such as bots) to influence foreign events [30], and Twitter has taken actions to find and remove accounts associated with the IRA [31]. Hence, it is likely that social media campaigns such as from Russian propaganda have become more advanced over time and employ new tactics, which thus pose the need for new, large-scale empirical evidence.

A particular threat of social media is that propaganda campaigns can reach online exposure at an unprecedented scale. While previous campaigns from the IRA relied largely upon trolls to spread propaganda [30, 31], it is likely that current influence operations make increasing use of bots. Generally, bots allow producing high volumes of software-controlled social media profiles at low cost [32]. Previously, bots have been deployed to spread disinformation, fake news, and hate speech on social media [20, 33–35]. In particular, they aid in the spread of low-credibility content (e. g., misinformation, false news) by amplifying early-stage diffusion [20]. Despite that bots post and receive less retweets than humans in social media networks, bots still attract more attention than human accounts [36] and thus can proliferate content that would otherwise not go viral [35]. An example of this was seen by the role of bots in the 2016 U.S. presidential election [21, 24, 37]. The result is that bots have the potential to shape the online discourse, radicalize users, and amplify social division [16, 34]. In the context of Russian propaganda, anecdotal evidence suggests that Russia invested in automated disinformation tools and "bot farms" for many years [16, 19, 38]. This raises the concern that pro-Russian bots may fuel and amplify Russian propaganda efforts also during the 2022 Russian invasion of Ukraine.

In this paper, we analyze the spread of pro-Russian support on social media. For this, we collected N = 349,455 messages from February through July 2022 with pro-Russian content from Twitter. Our analysis is three-fold. First, we analyze the overall reach of the pro-Russian messages. We find that pro-Russian messages received more than ∼251,000 retweets and thereby reached ∼14.4 million users. Second, we analyze the strategy with

which pro-Russian messages were disseminated. In particular, we document a disproportionate role of bots, which suggests the presence of a coordinated campaign: ∼20.28% of the spreaders are classified as bots, and most of them were created at the beginning of the invasion. Third, we study between-country heterogeneity in the impact of bots and find pronounced bot activity in countries abstaining from voting on United Nations Resolution ES-11/1 such as India, South Africa, and Pakistan. Together, our findings provide evidence for a Russian propaganda campaign, which was disseminated widely on social media and was amplified by bots in the early diffusion. Finally, our findings have important implications for designing effective counter-strategies to mitigate societal threats from propaganda in modern warfare.

2 Methods

2.1 Data collection

The data for this study were collected from the social media platform Twitter (http://twitter.com). Twitter was chosen because it is widely used for news consumption (in addition to entertainment) [39] and because of its high popularity in various parts of the world including Western, African, and Asian countries [40]. This is different from other social media platforms that sometimes have only a narrow user base in a specific geographic region, whereas our choice should allow us to study cross-country heterogeneity in pro-Russian support.

We queried the Twitter API v2 (Academic Research track) [41] for messages (source tweets, retweets, and replies) from February 1, 2022 through July 31, 2022. For this, we first defined a "seed" search query which we then expanded iteratively. Specifically, we started with the hashtag `#istandwithrussia`, which was a widespread hallmark of pro-Russian support on Twitter and among the most trending hashtags on both March 2 and March 3, 2022. We then analyzed a random subsample of 1000 messages to search for other pertinent hashtags that may have been used to signal pro-Russian support. As a result, we identified three additional common hashtags with a clearly pro-Russian connotation (i. e., `#standwithrussia`, `#istandwithputin`, and `#standwithputin`), and we then queried Twitter also for these hashtags. Note that the above hashtags likely capture the bulk of messages with pro-Russian hashtags on Twitter. The reason is that other (less common) hashtags that may also be indicative of pro-Russian support are typically used in conjunction with at least one of these hashtags (see the example messages in Additional file 1 Table S1).

We decided to use hashtags, instead of keywords, as search terms for multiple reasons. (1) The chosen hashtags went surprisingly viral in March 2022 and were suspected to be part of a larger propaganda campaign [42–44]. Hence, to provide large-scale, empirical evidence of such a campaign, an analysis based on messages containing these hashtags is necessary. (2) The use of hashtags is more strict than the use of keywords as search terms. This way, we ensured to only record messages that were likely part of the coordinated propaganda campaign. (3) The query hashtags contain distinct pro-Russian stances that more general keywords do not cover. This ensures that we capture pro-Russian support on Twitter rather than a more general discussion of the invasion.

Overall, our dataset consists of N = 368,762 messages (i. e., source tweets, retweets, and replies) with pro-Russian hashtags that were posted by 139,591 different users. The majority of messages (80.93%) was written in English.

2.2 Preprocessing

While our data collection allows for comprehensive coverage, the use of pro-Russian hashtags does not always equate to a pro-Russian stance. For example, users expressing an anti-Russian view sometimes employ pro-Russian hashtags to connect to the existing discourse. Similarly, Western news media report on the information warfare using the pro-Russian hashtags. After manual inspection, we found several false positives in our dataset, that is, Twitter messages that express an anti-Russian view or journalistic content, even though the message still uses a pro-Russian hashtag (e. g., `#istandwithrussia`). To remove false positives, we proceeded as follows. (1) We manually identified a list of 19 different anti-Russian and anti-Putin hashtags (e. g., `#stopputinnow`, `#stoprussia`). Note that we selected only hashtags that clearly shift the stance of a Twitter message, and, thus, one would not expect to find these hashtags in pro-Russian messages. The list is in Additional file 1 Table S4. (2) We discarded all messages containing one or more of the aforementioned hashtags. (3) We manually checked all verified accounts in our dataset and identified 44 Western news media outlets (e. g., NBC News, The Times). We used our common knowledge, as well as the biographies and queried messages of verified accounts to identify these news media outlets. The list is in Additional file 1 Table S5. We then discarded all messages from the aforementioned Western news outlets (as well as retweets of those messages) as they were merely reporting on Russian propaganda on Twitter using the query hashtags.

Overall, the filtering removed 19,307 messages (i.e., 5.24%). The resulting dataset contains N = 349,455 pro-Russian messages from 132,131 users, out of which 250,853 messages (71.78%) were retweets.

2.3 Dataset with pro-Ukrainian support

To compare pro-Russian and pro-Ukrainian support on Twitter, we collected a second dataset via the Twitter API v2 [41]. We performed the search analogous to the above; that is, we limited the search to the same time frame (February 1, 2022–July 31, 2022) and used a comparable set of hashtags in our search query: `#istandwithukraine`, `#standwithukraine`, `#istandwithzelensky`, and `#standwithzelensky`.

We applied the same preprocessing procedure to the messages with pro-Ukrainian support. To remove false positives, we identified five anti-Ukrainian hashtags that clearly shift the stance of the messages (see Additional file 1 Table S6). Overall, the filtering removed 461 messages. This left us with N = 9,818,566 messages (i.e., source tweets, retweets, and replies) posted by 2,079,198 users, which we consider as pro-Ukraine. Unless stated otherwise, all analyses in the main paper refer to the dataset with pro-Russian support (and not to the dataset with pro-Ukrainian support).

2.4 Human validation

We validated our preprocessing approach against human annotations following best practices [45]. Specifically, we recruited workers from Prolific (https://www.prolific.co/) and asked them whether a tweet was pro-Russia or pro-Ukraine. The annotators could select "pro-Russia", "pro-Ukraine", or "neutral/unclear/unrelated" as possible answers. For both datasets, we sampled 50 messages that were removed and 50 messages that remained after preprocessing. Messages that were removed from the pro-Russian dataset were considered pro-Ukrainian and vice versa. In accordance with best practices [45], we split the

validation into two batches of 100 messages each to avoid fatigue. Each dataset was annotated by three workers. The workers were subject to a strict screening procedure: residency in UK/US/AUZ, English as a first language; enrollment in an undergraduate, graduate, or doctoral degree; a minimum approval rate of 95%; and a minimum of 500 completed submissions on Prolific. We used the majority label for the final validation.

For the Russian dataset, we obtained a moderate agreement between the human annotators (Krippendorff's $\alpha = 0.49$ and Fleiss' $\kappa = 0.49$). The majority label from the annotators and the label from our preprocessing were in fair agreement (Cohen's $\kappa = 0.36$) when we considered the neutral/unclear label. When removing messages that were labeled as neutral/unclear, we obtained substantial agreement (Cohen's $\kappa = 0.7$) between the annotators and our preprocessing labels. Similarly, we obtained moderate agreement of annotators for the pro-Ukrainian dataset (Krippendorff's $\alpha = 0.52$ and Fleiss' $\kappa = 0.51$). The annotated majority label and our preprocessing label had moderate agreement (Cohen's $\kappa = 0.56$) when considering the neutral/unclear label and substantial agreement (Cohen's $\kappa = 0.71$) without the neutral/unclear label. Overall, this validates the reliability of our preprocessing approach.

2.5 Bot detection

We followed earlier research [20, 46, 47] and identified bots using Botometer [48]. Botometer is a supervised machine learning classifier that assesses the likelihood of an account being a bot using different features derived from the account, the friendship network, and different linguistic features. Previous research has empirically shown that bot detection via Botometer is highly accurate (area under the receiver operating curve [AUROC] of 0.96) [49]. Moreover, Botometer is well maintained, updated regularly to incorporate state-of-the-art data and methods, and has been widely adopted in research [50]. We directly accessed Botometer API [51] maintained by the Indiana University Observatory on Social Media. Botometer then returns the probability of an account being a bot. In line with previous research [20], we classified accounts with Botometer scores > 0.5 as bots. Overall, the Botometer API returned bot scores for 82,604 users (62.5%). Accounts that could not be matched onto human vs. bot due to Twitter's content moderation efforts were excluded from analyses that specifically differentiate between bot vs. human.

We validated the share of bots detected by Botometer [48] using Bot Sentinel [52]. Bot Sentinel is a machine learning classifier for inappropriate accounts on Twitter, which includes bots, trolls, and coordinated accounts (with a high accuracy of 95% [52]). It requires at least ten sample messages per account to make classifications, which highly limits the number of accounts in our dataset that can be validated. We let Bot Sentinel classify the subset that fulfilled the requirements, which amounted to 2661 accounts. The agreement between the classifications of Botometer and Bot Sentinel on this subset is 61.93%. This can be explained by the slightly different definitions by which accounts are flagged. Botometer is designed to detect bots, whereas Bot Sentinel is designed to detect inappropriate accounts, i.e., a much broader concept. Yet, the algorithms show a high agreement regarding the share of bots and humans: Botometer classifies 25.29% of the accounts as bots while Bot Sentinel classifies 26.53%. This suggests that the Botometer is able to accurately classify the share of bots in our data. Moreover, the main conclusions of our analysis did not change when considering only the validation set classified by Bot Sentinel in our analysis: India, South Africa, and the U.S. remain the main targets for bots.

2.6 Location analysis

To infer the geographic location where users are active, we applied the following procedure. (1) Users sometimes directly tagged their geolocation in messages in the form of a country code. In our dataset, this allowed us to identify the country location for 0.6% of the users in our dataset. (2) Otherwise, we analyzed the self-reported location in a user's Twitter profile. In our dataset, this information was available for around 59% of the users. We then entered the self-reported locations into Python Geocoder [53], which extracts real-world locations based on the OpenStreetMap API [54]. The API returns the spatial coordinates of the real-world location and an accuracy score, i.e., an estimate of how well the model was able to match the input to a real-world location. To account for incorrect or invalid locations, we filtered the results based on the accuracy that the Python Geocoder returns. We analyzed the distribution of the accuracy scores and found a bimodal distribution with a valley at 0.45. We manually inspected the self-reported locations with an accuracy below 0.45 and subsequently set the threshold accordingly, discarding all geocode annotations with an accuracy below 0.45. (3) For users with neither geotagged messages nor a valid, self-reported location, the location was determined using the following heuristic. Specifically, we assumed that users live in the same country as their followers and, we thus approximated a user's country location through the country location of their follower base. Hence, for the remaining users with no location, we extracted the top 1000 followers each using the Twitter API v2 Users Endpoint [55]. We then geocoded the self-reported locations of the followers (where possible) and computed their geometric median. Subsequently, we mapped the spatial coordinates onto country codes using the "naturalearth geometry" in GeoPandas v0.11.1 [56], which we then used as the estimated country of residence. Overall, the steps (1)–(3) yielded location information for 70.19% of all users in our dataset. The relative frequency of bots in each country was computed as the mean number of bots among the classified users of that country in our dataset. We later also perform robustness checks: we plot only the accounts from steps 1 and 2 without the followers proxy (see Additional file 1 Figure S3) and we plot humans, bots and accounts without bot score information separately (see Additional file 1 Figure S4).

We further validated our approach in two steps. First, we validated the accuracy of the Python Geocoder [53]. For this, we sampled 750 users for which Twitter was able to obtain a geotag of the message and that also provided a self-reported location. Analogous to above, we entered the self-reported locations into the Python Geocoder and obtained spatial coordinates for 599 users after applying a threshold of 0.45. We obtained an almost perfect agreement (Cohen's $\kappa = 0.92$) between the country code provided by Twitter and the country code provided by the Python Geocoder [53]. This proves that the country codes obtained through the Python Geocoder are highly accurate. Second, we validated our assumption that users live in the same country as their followers. For this, we extracted the top 1000 followers for the same sample of 750 users as above. Analogously, we geocoded the self-reported locations of the followers and computed the geometric median to obtain the country of residence. This yielded an estimated location for 452 users, which were in almost perfect agreement (Cohen's $\kappa = 0.81$) with the true country of the validation set.

2.7 Retweet network

To visualize the retweet network, we represented individual users as nodes and retweets as edges. We colored the nodes and edges based on the country of origin of the corresponding

accounts (India = purple, U.S. = blue, and South Africa = green). In our implementation, we built the network using networkx [57] and used the software Gephi v0.9.7 [58] for visualization. For better readability, we applied a weighted degree filter of 10 and used the filter "giant components", so that only nodes with a large number of retweets remained. Since the retweet network is undirected, the weighted degree refers to the number of in- and outgoing retweets a node has. We later also perform a robustness check where we plot the retweeting network for users where no bot score information was available (see Additional file 1 Figure S5).

3 Results

3.1 Pro-Russian support on social media

Our analysis is based on Twitter messages posted between February through July 2022 that used the hashtags `#istandwithrussia`, `#standwithrussia`, `#istandwithputin`, and `#standwithputin`. We applied further filtering rules to select only messages where the content was pro-Russian (see Methods). Overall, this yielded N = 349,455 messages. The messages further generated nearly 1 million likes. To measure the global exposure to pro-Russian messages, we estimated the overall readership based on the number of unique users that followed authors of pro-Russian messages in our dataset [59], amounting to ∼14.4 million users.

The messages in our dataset are fairly diverse (see Additional file 1 Table S1). For example, some messages contain only a series of hashtags (e. g., "*#IStandWithPutin #isupportrussia #Putin #standforrussia #StandWithPutin #IndiaWithRussia*"), while others state verbal affirmations of support for Putin or hate against Ukraine or NATO countries. Examples of the latter are: "*@RWApodcast I literally love Putin. The most honest leader in the world. #istandwithrussia*" and "*US is responsible for more than 81% conflicts in the world. The real war criminal is US. US should be completely isolated on the global stage #IStandWithPutin #RussiaArmy #IStandWithPutin*". By analyzing popular hashtags, we also see that several of them are unique to expressing a pro-Russian sentiment (see Additional file 1 Table S3). Examples are, e. g., `#hypocrisy` (posted 5682 times), `#doublestandards` (posted 2552 times), and `#stopnato` (posted 2156 times).

Pro-Russian messages showed distinctive temporal patterns (see Fig. 1) that coincided with the day that the United Nations General Assembly adopted Resolution ES-11/1 deploring the invasion (March 2, 2022). For example, peaks in the message volume occurred on March 2, 2022 (64,738 pro-Russian messages), March 3, 2022 (103,772 pro-Russian messages), and March 4, 2022 (66,794 pro-Russian messages), respectively. A fine-grained analysis showing temporal dynamics of the number of bot and human messages can be found in Additional file 1 Figure S1.

Further, on the day of the UN vote (March 2, 2022), ∼41.7% of the posted messages can be traced back to India, followed by Pakistan (∼5.9%) and Nigeria (∼2%). In contrast, on the day after the UN vote (March 3, 2022), the majority of the messages were posted from the U.S. (∼14.1%), Nigeria (∼10.5%), and India (∼10%). Apparently, messages from the U.S. were surprisingly rare on the day of the UN vote, despite that the majority of the Twitter user base is from the U.S. [40]. This suggests that pro-Russian support was potentially disseminated through a campaign targeting specific countries, for which we provide evidence in the following.

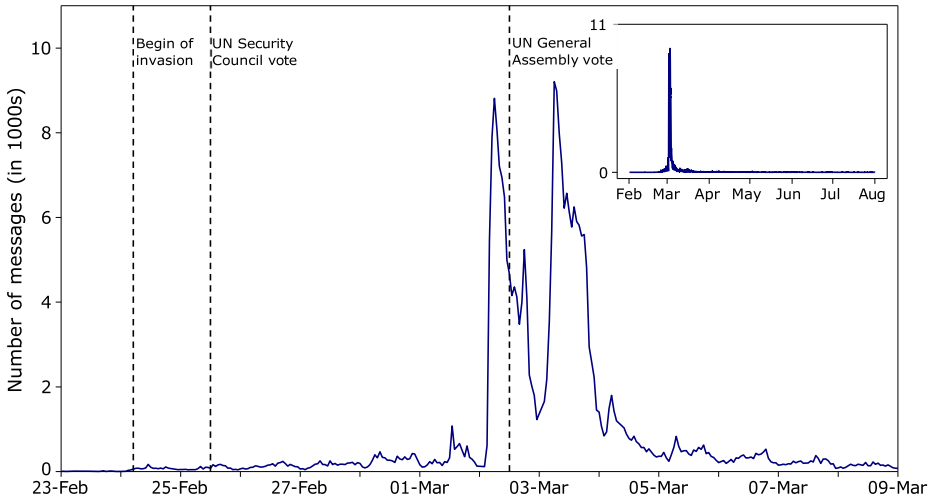

Figure 1 Temporal dynamics of pro-Russian support. The plot shows the number of pro-Russian messages during the first two weeks of the invasion. The peak on March 2, 2022, coincides with the day the United Nations General Assembly adopted Resolution ES-11/1 deploring the invasion. Inset: volume of pro-Russian messages for the entire time period of the dataset

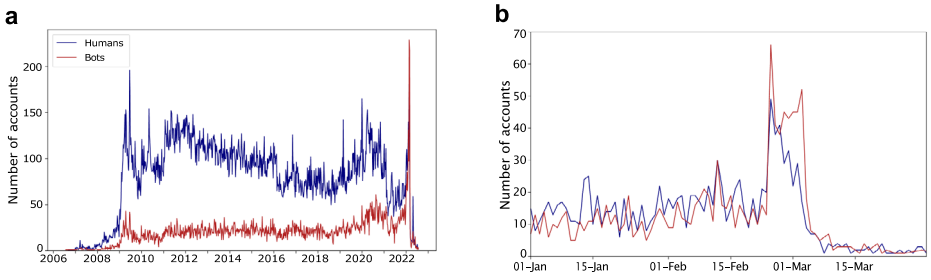

Figure 2 Spreaders of pro-Russian messages. (a) Dates on which accounts were created. Here, the time axis starts with the inception of Twitter in 2006. (b) Dates on which accounts were created. Here, the time axis starts shortly before the beginning of the 2022 Russian invasion

3.2 Spreading dynamics of pro-Russian support

Pro-Russian messages have been spread by 132,131 accounts (see Additional file 1 Table S7 for a list of influential accounts). To analyze the role of bots in the spread of pro-Russian messages, we used Botometer [48] to classify accounts according to humans and bots. For each account, we computed a bot score ($\rho \in [0,1]$), which can be interpreted as the level of automation of that account [20]. A threshold of 0.5 is typically used to classify an account as likely human or likely bot (see Methods for details). Using this method, 20.28% of the accounts were categorized as bots. Hence, bots played a critical role in spreading pro-Russian messages.

Accounts from humans and bots showed a clear difference in when the accounts were created (Fig. 2a). Accounts classified as bots tended to have been created more recently than accounts classified as humans. Notably, there also was a clear peak in the number of newly created bots, which coincided with the beginning of the invasion on February 24, 2022 (Fig. 2b). A robustness check showing the creation dates of accounts for which a bot score could not be assigned is provided in Additional file 1 Figure S2.

To further quantitatively characterize the spreading dynamics of pro-Russian support, we collected an additional dataset with pro-Ukrainian messages that were posted on Twitter between February 2022 through July 2022 (see Additional file 1 Table S2). We first compared the number of bots spreading pro-Russian messages (20.28%) with the number of bots spreading pro-Ukrainian messages (14.25%). Here, we find that pro-Ukrainian support was spread by significantly less bots than pro-Russian support (Kolmogorov-Smirnov (KS) test [60]: $D = 0.062$, $p < 0.001$). We then compared spreaders of pro-Russian vs. pro-Ukrainian support in terms of the number of followers (Fig. 3a): Pro-Russian supporters had a substantially smaller number of followers with a mean of only 1690 followers, whereas the mean number of followers was 2248 for pro-Ukrainian supporters (KS test: $D = 0.049$, $p < 0.001$). The number of followers is typically interpreted as a proxy for the social influence of online users [59], implying that spreaders of pro-Russian support had a comparatively smaller social influence than spreaders of pro-Ukrainian support.

We further find heterogeneity in the online virality of pro-Russian and pro-Ukrainian support. For this, we compared the number of likes, replies, and retweets that pro-Russian

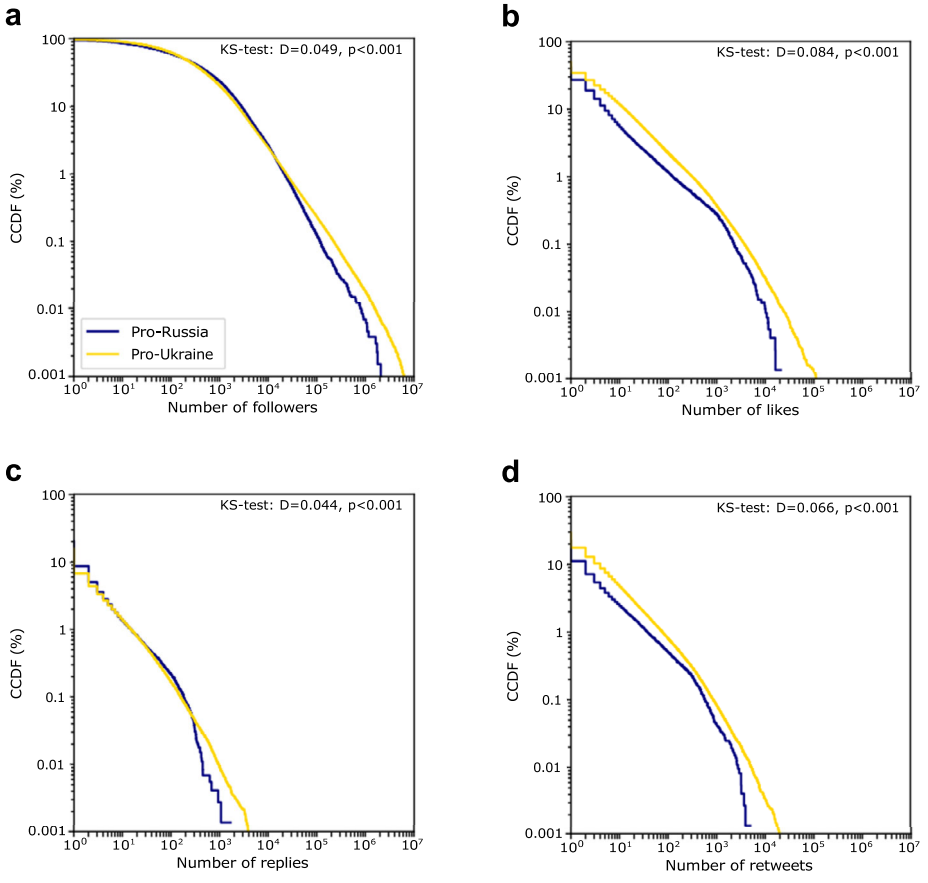

Figure 3 Online virality of pro-Russian vs. pro-Ukrainian support. Here, we compare complementary cumulative distribution functions (CCDFs) for: (a) the number of followers; (b) the number of likes; (c) the number of replies; and (d) the number of retweets. The former was computed at the user level, while the latter three were computed at the tweet level. Statistical comparisons are based on Kolmogorov-Smirnov (KS) tests [60]. All distributions span several orders of magnitude, implying that there was a substantial share of pro-Russian messages that went viral

vs. pro-Ukrainian source tweets received (Fig. 3b–d). On average, pro-Russian source tweets received 12.97 likes, 1.16 replies, and 3.38 retweets. The corresponding numbers were significantly smaller than for pro-Ukrainian source tweets, which, on average, received 28.35 likes, 1.22 replies, and 6.56 retweets (KS tests: $D = 0.084$, $p < 0.001$; $D = 0.044$, $p < 0.001$; and $D = 0.066$, $p < 0.001$, respectively). Thus, pro-Russian support tended to be less viral than pro-Ukrainian support. Note however that, for both pro-Russian and pro-Ukrainian support, we observe very broad distributions spanning several orders of magnitude. Hence, there was still a substantial proportion of pro-Russian messages that went viral.

3.3 Cross-country heterogeneity in the exposure to pro-Russian support

To analyze the cross-country heterogeneity in pro-Russian support, we inferred the geographic location of the underlying user accounts (see Methods). Evidently, the countries with the most accounts spreading pro-Russian messages were India, the United States, South Africa, and Nigeria (Fig. 4a). The pronounced role of these English-speaking countries in spreading pro-Russian messages may be partially explained by the use of English hashtags as search queries. However, these countries also show a high percentage of pro-Russian supporters in comparison to the overall number of Twitter users in that country (see Table 1). Moreover, pro-Russian support was disproportionally high in countries that abstained from voting on the United Nations Resolution ES-11/1 (such as India, South Africa, and Pakistan) relative to other English-speaking countries (such as the United

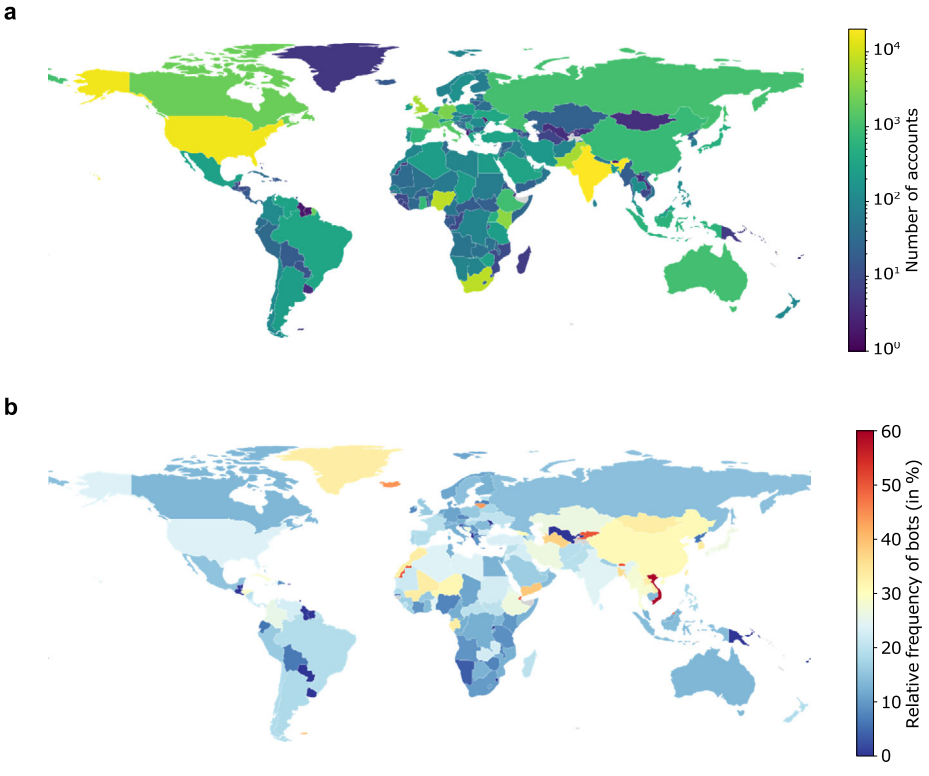

Figure 4 Cross-country differences in the spread of pro-Russian support. Here, we inferred the geographic location of accounts (see Methods). (a) Number of users per country (log scale). (b) Relative frequency of bots per country (in %)

Table 1 Total number of Twitter users per country (in millions) and the relative frequency of pro-Russian supporters in our dataset. The total number of Twitter users is based on data from 2022 [40, 61–63]. We selected the ten leading countries with the highest number of Twitter users. In addition, we included Nigeria, South Africa, and Pakistan, due to their relevance to our analysis

Country	Twitter users (in millions)	pro-Russian supporters (in %)
Nigeria	0.32	2.290
South Africa	2.85	0.263
Pakistan	3.40	0.161
India	23.60	0.085
United Kingdom	18.40	0.030
Canada	7.90	0.028
United States	76.90	0.021
Indonesia	18.45	0.003
Saudi Arabia	14.10	0.002
Mexico	13.90	0.002
Turkey	16.10	0.002
Brazil	19.05	0.002
Japan	58.95	0.001

States, the United Kingdom, and Australia). Subsequently, we computed the relative frequency of bots across countries (Fig. 4b). Several of the countries with many pro-Russian messages also showed a pronounced role of likely bot activity: 24.2% of the accounts in India were bots, 23.9% in the United States, 10.2% in South Africa, and 7.9% in Nigeria. The patterns remained robust across different methods for inferring geographic locations (Additional file 1 Figure S3). We also conducted a robustness check in which the locations of humans, bots, and accounts without bot scores were mapped separately and found robust patterns (see Additional file 1 Figure S4).

Overall, countries that abstained from the UN vote had the highest relative frequency of bots (20.3%), in comparison to countries that voted against (14.9%) or approved (16.6%) the UN Resolution ES-11/1 (one-way ANOVA test: $F = 84.73$; $p < 0.001$). Hence, countries abstaining from the UN vote (e.g., India, South Africa) have been prime targets of bots circulating pro-Russian support. Supplement A provides a content analysis that further substantiates the connection between countries and the UN vote.

We also compared the cross-country heterogeneity of pro-Russian support to pro-Ukrainian support (see Fig. 5). We find a larger focus of pro-Ukrainian support in the U.S. and European countries. Countries that were highly active in spreading pro-Russian support such as South Africa, Pakistan, and Nigeria were not as active in spreading pro-Ukrainian support. Furthermore, we compared the relative frequency of bots of pro-Ukrainian supporters. Similarly to pro-Russian support, we found a pronounced bot activity in India (28.57%) and South Africa (16.67%). In contrast, the United States and Nigeria showed less to no bot activity (11.42% and 0%, respectively).

3.4 Retweet network

We analyzed the network diffusion patterns of pro-Russian support and especially how bots promoted its spread. First, we examined the retweet dynamics with which pro-Russian messages were disseminated across different account types (Fig. 6a). humans tended to primarily retweet other humans rather than bots. bots, in return, tended to mainly retweet humans but retweeted other bots only rarely. This indicates that bots drove the spread of pro-Russian support primarily by exposing humans to human-generated, pro-Russian messages.

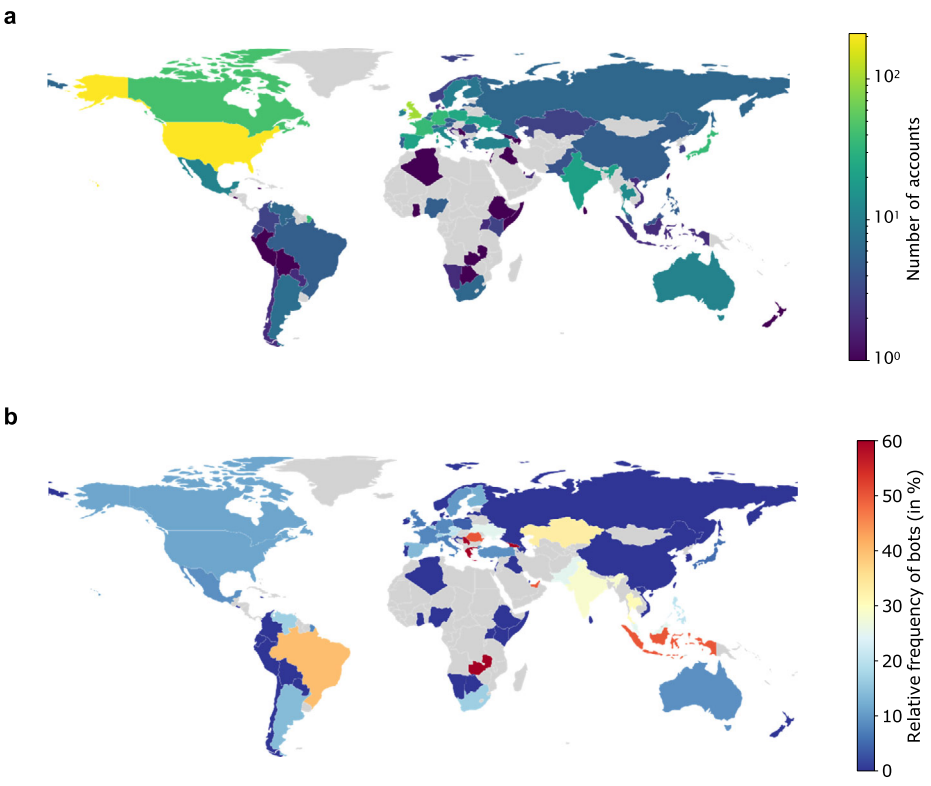

Figure 5 Cross-country differences in the spread of pro-Ukrainian support. Here, we inferred the geographic location of a subsample of pro-Ukrainian accounts (see Methods). (a) Number of users per country (log scale). (b) Relative frequency of bots per country (in %)

The retweet network of individual accounts revealed several clusters in which pro-Russian messages primarily circulated (Fig. 6b–d). By matching accounts to their geographic location, we find that some of the clusters were of large geographic homogeneity. In particular, we could map two of the clusters to users from India and South Africa, both of which were two major countries that abstained from the UN vote. These countries exhibited relatively isolated retweet networks in which pro-Russian messages were able to infiltrate the local online communities with little external influence. In comparison, accounts from the U.S. did not show the same geographic clustering but were more broadly scattered over the retweet network. This suggests that there may have been differences in the coordination behind the pro-Russian support across countries as India and South Africa were specifically targeted by pro-Russian supporters. Accounts from the U.S. retweeted accounts from all over the network, whereas accounts from South Africa and India discussed the invasion mostly with accounts from their country. The content analysis in Supplement A further substantiates that discussions in India and South Africa were held at a local scale and focused on national issues. We also performed a robustness check of the retweeting networks on the accounts that did not have bot information and corroborated our findings (see Additional file 1 Figure S5).

3.5 Amplification of pro-Russian support spreading through bots

We further examined how bots contributed to the spreading of pro-Russian support (e.g., by automatically making pro-Russian hashtags go viral or retweeting other accounts) and,

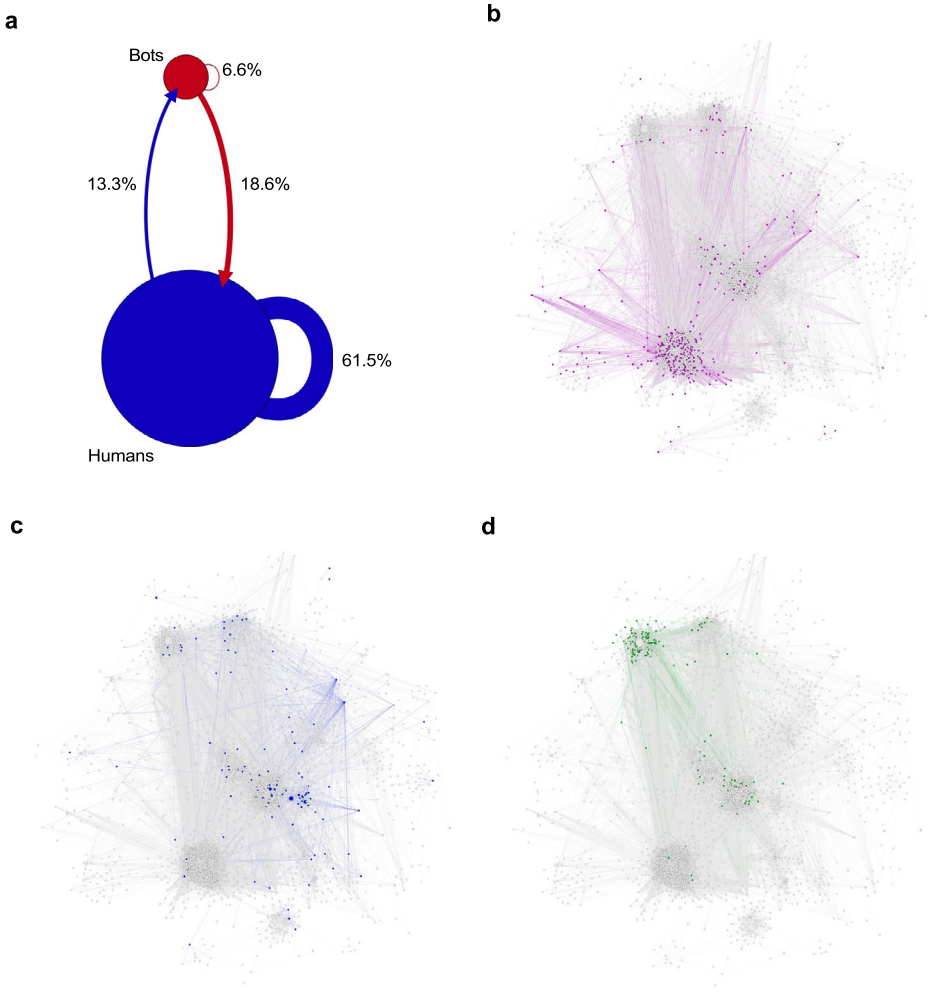

Figure 6 Retweeting network. (a) Spreading patterns between humans vs. bots (blue = humans, red = bots). The node size represents the number of humans vs. bots. The edges represent the direction and relative frequency of retweets. (b) Retweet network with accounts from India colored in purple. (c) Retweet network with accounts from the U.S. colored in blue. (d) Retweet network with accounts from South Africa colored in green. The retweet networks were visualized using Gephi [58] (see Methods)

to this end, analyzed differences in the online behavior of humans vs. bots. bots were responsible for only 20.82% of the source tweets, while 79.18% of the source tweets originated from humans (see Additional file 1 Figure S6). Hence, most of the content generation was done by humans. However, even though 20.28% of the accounts were categorized as bots in our sample, they were responsible for 25.72% of the retweets. As a measure of popularity, we analyzed the number of likes that messages of humans and bots received. Messages from bots received 17.46% of the likes that pro-Russian messages received overall. Hence, messages from bots were slightly less popular than messages from humans (Mann-Whitney U test: $U = 2 \cdot 10^9$; $p < 0.001$ with $\mu_{\text{bot}} = 9.75$ and $\mu_{\text{human}} = 10.02$).

We further explored the messaging activity of humans vs. bots. Specifically, we studied the distribution of bot scores across authors of source tweets and retweets (Fig. 7a). Here, we again find that humans took a leading role in content creation. We also explored how humans interacted with messages shared by bots. This provides insights into whether bots

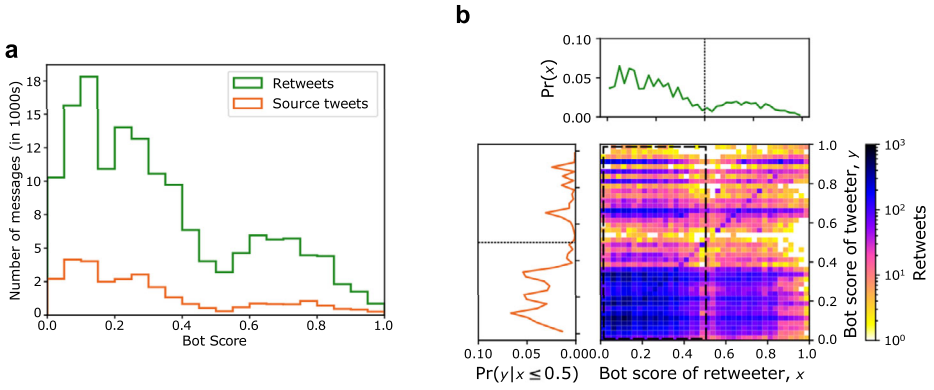

Figure 7 Impact of humans and bots. (a) Distribution of bot scores for source tweets and retweets. The two groups had significantly different bot scores (Mann-Whitney U test: $U = 2 \cdot 10^9$; $p < 0.001$, and Mood's median test: $\chi = 666.84$; $p < 0.001$ with median bot score of source tweeters = 0.22 and median bot score of retweeters = 0.27), implying that retweeters were more likely to be bots. (b) Joint distribution of bot scores of authors of source tweet-retweet pairs (heatmap). The top subplot shows the distribution of bot scores for retweeters. The left subplot shows the distribution of bot scores for accounts that were retweeted by accounts classified as humans (using a threshold of 0.5). We find that most source tweets were posted by humans. They were also active retweeters, but so were bots. Different from the spread of low-credibility content [20], we find that a significant proportion of retweeters were bots and that they tended to retweet humans rather than other bots.

were able to elicit human interactions such as retweeting. For this, we computed the distribution of bot scores for each source tweet–retweet pair and thus analyzed who retweets whom (Fig. 7b). Generally, humans did most of the tweeting (Fig. 7b, top). humans were also active in retweeting but bots were relatively more active (see Additional file 1 Figure S6). Moreover, many accounts retweeted themselves to amplify their own messages, a tactic that was commonly used by bots (23.5% of the 1653 accounts that retweeted themselves were bots). The results confirmed our findings from the retweeting network: humans tended to retweet other humans, while bots were more inclined to retweet humans. However, humans rarely retweeted bots. This is a crucial difference from earlier work on low-credibility content for which humans have been found to frequently retweet bots [20], implying that it is difficult for bots to make pro-Russian messages go viral among humans.

Given this evidence, we further examined whether there were different temporal dynamics in the retweeting behavior of bots and humans. For this, we compared the bot score distribution of retweeters across different time lags for retweets (Fig. 8). We find that humans were retweeted equally fast by bots and humans (Fig. 8a), while bots were retweeted by other bots with a disproportionately small time lag (Fig. 8b). This suggests that bots systematically retweeted other bots early in the diffusion to promote the proliferation of pro-Russian support.

Previous work found that a key strategy for bots is to spread content by mentioning influential accounts (e. g., "*@UN*", "*@cnnbrk*", or "*@RusEmbEthiopia*"), in the hope that they reshare and thus boost credibility [20]. To systematically analyze whether pro-Russian bots employ such a mentioning strategy, we computed the mean number of followers of the mentioned accounts (Fig. 8c). We find that humans tended to mention accounts with substantially more followers than bots. Recall that the number of followers is a common proxy for the social influence of online users [59], which implies that bots tended to mention users with a smaller social influence in their messages. Notably, this finding differs

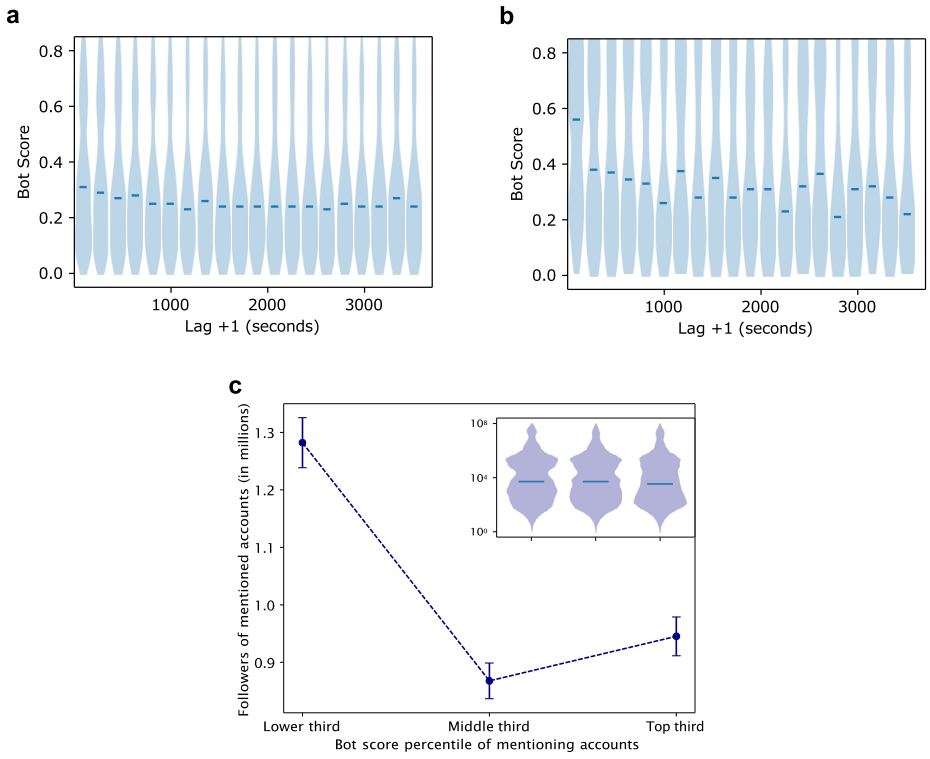

Figure 8 Bot strategies. (a) Distribution of bot scores of accounts that retweeted human-made source tweets grouped by different time lags between source tweet and the corresponding retweet. (b) Distribution of bot scores of accounts that retweeted likely bot-made source tweets grouped by different time lags between the source tweet and the corresponding retweet. Hence, bots (but not humans) tended to retweet bots to promote the early diffusion of pro-Russian messages. (c) Here, we plot the average number of followers of mentioned accounts to analyze whether bots specifically targeted influential users. The mentioning accounts are grouped by their bot score percentile. Error bars indicate the standard errors. Inset: violin plot showing the distribution of follower counts for the mentioned user accounts in each bot score group. In violin plots, the width of a contour represents the probability of the corresponding value, and the median is marked by a colored line

from earlier research studying the spread of low-credibility content through bots, where bots – and not humans – target influential users to make messages strategically go viral [20].

An alternative proxy for the social influence of users is their centrality in a retweet network, computed as their PageRank [34]. Consistent with the above findings, we find that bots mentioned users with lower PageRank (mean PageRank of 0.002) than humans (mean PageRank of 0.0022). This difference is statistically significant (Mann-Whitney U test: $U = 2 \cdot 10^{10}$; $p < 0.001$) and, again, differs from earlier findings, where bots have been found to target influential users at the center of retweeting networks when promoting the spread of inflammatory content [34].

4 Discussion

The massive spread of online propaganda has been identified as a major threat to democracies [64]. While propaganda is a tool that has been used since ancient times, social media has made its spreading faster and more scalable, thereby presenting particularly fertile ground for sowing propaganda. Prior research provides evidence of systematic social

media propaganda campaigns that aim to influence geopolitical events such as elections [16, 20, 22, 27]. Online propaganda has also become a concerning tool in modern warfare. Here, a particular threat is that social media amplifies the spread of misinformation and helps propaganda campaigns to shape false narratives around wars [65]. So far, however, there is little systematic, scientific research that analyzed the spread of pro-Russian support during the 2022 Ukraine invasion, which is our contribution. Unlike earlier research on historical tactics of the IRA [20–27], we focus on a recent foreign influence operation that employed state-of-the-art and novel tactics to proliferate propaganda (e.g., by making large-scale use of automation through bots).

We find robust support for a Russian propaganda campaign, defined as systematic and coordinated efforts to manipulate beliefs and behaviors in the propagandists' interests [66]. Pro-Russian messages have been spread on Twitter disproportionately through bots, which interacted in highly-connected retweet networks. The retweet networks showed distinctive clusters in countries that are of key interest for Russian politics (e. g., India and South Africa) and thus suggest a coordinated effort. The accumulation of messages on the day of the UN vote on Resolution ES-11/1 gives rise to concerns that countries that abstained from the UN voting were targeted by Russian propaganda efforts. Strikingly, many bots that spread pro-Russian messages were created shortly before the UN vote, which indicates an intentional and planned manipulation of public opinion on Twitter as part of a Russian propaganda campaign.

Our findings demonstrate that bots are an important driver in the early diffusion of bot-created propaganda on social media. Bots were more active retweeters than humans and acted together in a coordinated manner. Unlike spreaders of low-credibility content [20] and inflammatory content [34], bots mentioned users with less social influence than humans when spreading pro-Russian messages. A possible explanation for this strategy behind Russian propaganda is that, because bots were rarely retweeted by humans (cf. Fig. 7b), they did not target individuals. Instead, bots primarily aimed to expose users to organic, pro-Russian messages from humans. By creating traffic around Russian propaganda, certain hashtags appeared as so-called "trending topics" on the front page of Twitter and were thus visible to all users [42–44]. This is especially alarming, since repeated exposure can lead people to perceive misinformation as accurate [67].

Crucial differences between the spread of propaganda and the spread of low-credibility content [20, 34, 68] by bots become evident. On the one hand, we identified bots as amplifiers of propaganda rather than content creators. bots in propaganda were more inclined to retweet than to produce "original" content (e. g., source tweets). On the other hand, bots did not specifically target influential users. Instead, they aimed at broad exposure to maximize the number of people that see their message. Previously, such an amplification strategy has been conjectured to be a mature tactic of the IRA [69]. The likely goal is to augment the prominence and activity level of organic accounts that naturally act in ways that are aligned with the objectives of the propaganda campaign.

As with other research, ours is not free of limitations, which presents opportunities for future research. First, our results are based on a single social media platform. However, Twitter is a platform with a particularly large and international audience, which makes it a fertile ground for planting propaganda and, hence, presents a common focus in earlier research [11, 16, 21, 25, 35]. Second, our data covers mostly messages in English since we searched messages based on English hashtags. However, these hashtags went reportedly

viral in March 2022 [42–44] and, subsequently, were widely used as search terms as well as to strategically flag corresponding messages. Third, the pro-Ukrainian support on Twitter is much larger than the pro-Russian support in absolute terms. This is likely the case since the main user base of Twitter is located in the West, which mostly supports Ukraine in the conflict. By primarily analyzing pro-Russian support, we focus on a minority of all tweets around the Russo-Ukrainian war. However, there is anecdotal evidence that there is a coordinated propaganda campaign behind the pro-Russian support on Twitter, and not behind the pro-Ukrainian support. Fourth, another limitation of our study is the possibility that Twitter may have removed some particularly egregious pro-Russian messages through content moderation efforts. However, messages that were removed by Twitter are also those that were hindered to go viral and that humans were thus not exposed to. Fifth, the accuracy of our analysis depends on the accuracy of other tools such as Botometer [48]. However, these tools have been shown to achieve a high accuracy [48] and are widely used in research [20, 24, 29, 70]. Sixth, while the scale of the pro-Russian support on Twitter is impressive in absolute terms (e.g., reached ∼14.4 million users), it may not have infiltrated online communities to an extent that swayed public opinion. Research largely still lacks an understanding of the real-world effects of social media propaganda [16], which future work should explore. In particular, additional research with complementary research methods (e.g., survey approaches [27]) is needed to better understand the impact of exposure to propaganda on opinion formation and public discourse.

Our results have direct implications for society and democracies. First, our results are alarming as social media platforms present substantial vulnerabilities that propaganda campaigns can exploit strategically. Without significant effort by social media platforms to curb the spread of disinformation, toxic content can spread widely and virally [71–76]. Here, more research is needed to understand the mechanism behind the pro-Russian propaganda campaign [77], as well as machine learning for detection [78]. Second, our results suggest that an effective countermeasure to curb the spread of propaganda is to reduce the influence of bots. Here, it may be likely that counter-measures from fake news mitigation can be adapted [79–81]; yet this requires further research to establish the effectiveness of such interventions. Third, propaganda on social media may influence public opinion and increase political division. It is thus important that policy-makers are aware of the potential threats that social media propaganda poses to modern societies. As such, it will be critical to continuously monitor and actively counter the proliferation of online propaganda in the future.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00414-5.

Additional file 1. (PDF 1.1 MB)

Acknowledgements
Codes for plotting are based on Shao et al. [20], which we gratefully acknowledge.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Abbreviations
AUROC, area under the receiver operating curve; IRA, Internet Research Agency; KS test, Kolmogorov-Smirnov test; U.K., United Kingdom; UN, United Nations; U.S., United States of America.

Availability of data and materials
The data and code that support the findings of our study are available on GitHub (https://github.com/DominiqueGeissler/Russian_Propaganda_on_social_media).

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed to conceptualization, results interpretation, and manuscript writing. DG contributed to data analysis. All authors approved the manuscript.

Author details
[1]LMU Munich, Munich, Germany. [2]Munich Center for Machine Learning (MCML), Munich, Germany. [3]University of Giessen, Giessen, Germany.

Received: 6 February 2023 Accepted: 25 August 2023 Published online: 12 September 2023

References

1. United Nations. Security Council, 8974th meeting
2. Lister T, Kesa J Ukraine says it was attacked through Russian, Belarus and Crimea borders. CNN (24 February 2022). https://edition.cnn.com/europe/live-news/ukraine-russia-news-02-23-22/h_82bf44af2f01ad57f81c0760c6cb697c
3. Nations U. Security Council, 7683rd meeting
4. Kirby P EU leaders consider how to cap gas prices. BBC News (6 October 2022). https://www.bbc.com/news/world-europe-63130645
5. The Economist The coming food catastrophe (19 May 2022). https://www.economist.com/leaders/2022/05/19/the-coming-food-catastrophe
6. United Nations High Commissioner for Refugees Situation Ukraine refugee situation (19 September 2022). https://data.unhcr.org/en/situations/ukraine
7. Nations U General Assembly, 11th emergency special session, 5th & 6th meetings (am & pm) (2 March 2022)
8. Sloane W (2022) Putin cracks down on media. Br Journal Rev 33:19–22
9. Alyukov M (2022) Propaganda, authoritarianism and Russia's invasion of Ukraine. Nat Hum Behav 6:763–765
10. Troianovski A, Safronova V Russia takes censorship to new extremes, stifling war coverage. The New York Times (4 March 2022). https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html
11. Alieva I, Moffitt JD, Carley KM (2022) How disinformation operations against Russian opposition leader Alexei Navalny influence the international audience on Twitter. Soc Netw Anal Min 12:80
12. Golovchenko Y (2020) Measuring the scope of pro-Kremlin disinformation on Twitter. Humanit Soc Sci Commun 7:176
13. Yablokov I (2022) Russian disinformation finds fertile ground in the West. Nat Hum Behav 6:766–767
14. Sanovich S Computational propaganda in Russia: the origins of digital misinformation. Oxford Internet Institute
15. Ratkiewicz J et al (2011) Detecting and tracking political abuse in social media. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 5, pp 297–304
16. Bail CA et al (2020) Assessing the Russian Internet Research Agency's impact on the political attitudes and behaviors of American Twitter users in late 2017. Proc Natl Acad Sci USA 117:243–250
17. Golovchenko Y, Hartmann M, Adler-Nissen R (2018) State, media and civil society in the information warfare over Ukraine: citizen curators of digital disinformation. Int Aff 94:975–994
18. Del Vicario M et al (2016) The spreading of misinformation online. Proc Natl Acad Sci USA 113:554–559
19. Doroshenko L, Lukito J (2021) Trollfare: Russia's disinformation campaign during military conflict in Ukraine. Int J Commun 15:4662–4689
20. Shao C et al (2018) The spread of low-credibility content by social bots. Nat Commun 9:4787
21. Badawy A, Ferrara E, Lerman K (2018) Analyzing the digital traces of political manipulation: the 2016 Russian interference Twitter campaign. In: IEEE/ACM international conference on advances in social networks analysis and mining, pp 258–265
22. Guess AM, Nyhan B, Reifler J (2020) Exposure to untrustworthy websites in the 2016 US election. Nat Hum Behav 4:472–480
23. Luceri L, Giordano S, Ferrara E (2020) Detecting troll behavior via inverse reinforcement learning: a case study of Russian trolls in the 2016 US election. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 14, pp 417–427
24. Bessi A, Ferrara E (2016) Social bots distort the 2016 U.S. presidential election online discussion. First Monday 21
25. Dutta U et al (2021) Analyzing Twitter users' behavior before and after contact by the Russia's Internet Research Agency. Proc ACM Hum-Comput Interact 5:1–24
26. Arif A, Stewart LG, Starbird K (2018) Acting the part: examining information operations within #BlackLivesMatter discourse. Proc ACM Hum-Comput Interact 2:1–27
27. Eady G et al (2023) Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. Nat Commun 14:62
28. Grčar M, Cherepnalkoski D, Mozetič I, Kralj Novak P (2017) Stance and influence of Twitter users regarding the Brexit referendum. Comput Soc Netw 4:6
29. Ferrara E (2017) Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday 22

30. Golovchenko Y, Buntain C, Eady G, Brown MA, Tucker JA (2020) Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election. Int J Press/Polit 25:357–389
31. Twitter Update on Twitter's review of the 2016 US election (31 January 2018). URL. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html
32. Ferrara E, Varol O, Davis C, Menczer F, Flammini A (2016) The rise of social bots. Commun ACM 59:96–104
33. Chen W, Pacheco D, Yang K-C, Menczer F (2021) Neutral bots probe political bias on social media. Nat Commun 12:5580
34. Stella M, Ferrara E, de Domenico M (2018) Bots increase exposure to negative and inflammatory content in online social systems. Proc Natl Acad Sci USA 115:12435–12440
35. Caldarelli G, de Nicola R, Del Vigna F, Petrocchi M, Saracco F (2020) The role of bot squads in the political propaganda on Twitter. Commun Phys 3:81
36. González-Bailón S, de Domenico M (2021) Bots are less central than verified accounts during contentious political events. Proc Natl Acad Sci USA 118:e2013443118
37. Badawy A, Lerman K, Ferrara E (2019) Who falls for online political manipulation? In: Companion Proceedings of The World Wide Web Conference, pp 162–168
38. Stukal D, Sanovich S, Bonneau R, Tucker JA (2017) Detecting bots on Russian political Twitter. Big Data 5:310–324
39. Mitchell A, Shearer E, Stocking G (2021) News on Twitter: consumed by most users and trusted by many. Pew Research Center. https://www.pewresearch.org/journalism/2021/11/15/news-on-twitter-consumed-by-most-users-and-trusted-by-many/
40. Dixon S (2022) Countries with most Twitter users 2022. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
41. Twitter (2022). Twitter API v2. https://developer.twitter.com/en/docs/twitter-api
42. The Economist Russia is swaying Twitter users outside the West to its side (14 May 2022). https://www.economist.com/graphic-detail/2022/05/14/russia-is-swaying-twitter-users-outside-the-west-to-its-side?utm_medium=social-media.content.np&utm_source=twitter&utm_campaign=editorial-social&utm_content=discovery.content
43. Gragnani J, Arora M, Ali S Ukraine war: the stolen faces used to promote Vladimir Putin. BBC News (10 May 2022). https://www.bbc.com/news/blogs-trending-61351342
44. Miller C Who's behind #IStandWithPutin? The Atlantic (5 April 2022). https://www.theatlantic.com/ideas/archive/2022/04/russian-propaganda-zelensky-information-war/629475/
45. Song H et al (2020) In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. Polit Commun 37:550–572
46. Broniatowski DA et al (2018) Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. Am J Publ Health 108:1378–1384
47. Wojcik S, Messing S, Smith A, Rainie L, Hitlin P (2018) Bots in the Twittersphere. Pew Research Center. https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/
48. Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 11, pp 280–289
49. Sayyadiharikandeh M, Varol O, Yang K-C, Flammini A, Menczer F (2020) Detection of novel social bots by ensembles of specialized classifiers. In: Proceedings ACM International Conference on Information and Knowledge Management, pp 2725–2732
50. Yang K-C, Ferrara E, Menczer F (2022) Botometer 101: social bot practicum for computational social scientists. J Comput Soc Sci 5:1511–1528
51. OSoMe Botometer Python API. https://github.com/IUNetSci/botometer-python
52. Sentinel B (2022). Platform developed to detect and track political bots, trollbots, and untrustworthy accounts. https://botsentinel.com
53. Carriere D (2013) Geocoder. https://geocoder.readthedocs.io
54. OpenStreetMap Wiki (2021). OSMPythonTools. https://wiki.openstreetmap.org/w/index.php?title=OSMPythonTools&oldid=2150829
55. Twitter Twitter API v2 Users Endpoint. https://developer.twitter.com/en/docs/twitter-api/users/follows/api-reference/get-users-id-followers
56. Jordahl K et al (2022) Geopandas v0.11.1. https://geopandas.org/en/stable/
57. Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the python in science conference, pp 11–15
58. Bastian M, Heymann S, Jacomy M (2009) Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 3, pp 361–362
59. Cha M, Haddadi H, Benevenuto F, Gummadi K (2010) Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 4, pp 10–17
60. Massey FJ (1951) The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc 46:68–78
61. Kemp S (2022) Digital 2022: South Africa. Datareportal. https://datareportal.com/reports/digital-2022-south-africa
62. Kemp S (2022) Digital 2022: Nigeria. Datareportal. https://datareportal.com/reports/digital-2022-nigeria
63. Kemp S (2022) Digital 2022: Pakistan. Datareportal. https://datareportal.com/reports/digital-2022-pakistan
64. Aral S, Eckles D (2019) Protecting elections from social media manipulation. Science 365:858–861
65. Scott M As war in Ukraine evolves, so do disinformation tactics. Politico (10 March 2022). https://www.politico.eu/article/ukraine-russia-disinformation-propaganda/
66. Jowett G, O'Donnell V (2012) What is propaganda, and how does it differ from persuasion? In: Propaganda & persuasion. SAGE, Los Angeles
67. Pennycook G, Cannon T, Rand DG (2018) Prior exposure increases perceived accuracy of fake news. J Exp Psychol Gen 147:1865–1880
68. Vosoughi S, Roy D, Aral S (2018) The spread of true and false news online. Science 359:1146–1151
69. Linvill DL, Warren PL (2022) Engaging with others: how the IRA coordinated information operation made friends. Harvard Kennedy School Misinformation Review
70. Suárez-Serrato P, Roberts ME, Davis C, Menczer F (2016) On the influence of social bots in online protests. Int Conf Soc Inform 10047:269–278

71. Bär D, Pröllochs N, Feuerriegel S (2023) New threats to society from free-speech social media platforms. Commun ACM
72. Pröllochs N, Bär D, Feuerriegel S (2021) Emotions in online rumor diffusion. EPJ Data Sci 10
73. Pröllochs N, Bär D, Feuerriegel S (2021) Emotions explain differences in the diffusion of true vs. false social media rumors. Sci Rep 11
74. Pröllochs N, Feuerriegel S (2023) Mechanisms of true and false rumor sharing in social media: collective intelligence or herd behavior? ACM Conference On Computer-Supported Cooperative Work And Social Computing
75. Robertson CE et al (2023) Negativity drives online news consumption. Nat Hum Behav
76. Naumzik C, Feuerriegel S (2022) Detecting false rumors from retweet dynamics on social media. In: Proceedings of the ACM Web Conference, pp 2798–2809
77. Geissler D, Feuerriegel S (2023) Analyzing the strategy of propaganda using inverse reinforcement learning: evidence from the 2022 Russian invasion of Ukraine. arXiv:2307.12788
78. Maarouf A, Bär D, Geissler D, Feuerriegel S (2023) HQP: a human-annotated dataset for detecting online propaganda. arXiv:2304.14931
79. Pennycook G et al (2021) Shifting attention to accuracy can reduce misinformation online. Nature 592:590–595
80. Gallotti R, Valle F, Castaldo N, Sacco P, de Domenico M (2020) Assessing the risks of 'Infodemics' in response to COVID-19 epidemics. Nat Hum Behav 4:1285–1293
81. Ducci F, Kraus M, Feuerriegel S (2020) Cascade-LSTM: a tree-structured neural classifier for detecting misinformation cascades. In: The ACM SIGKDD conference on knowledge discovery and data mining, pp 2666–2676

Publisher's Note