ROBOTROLLING
2019. ISSUE 3

ROBOTROLLING

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

Executive Summary

In the period May—July 2019 bots accounted for 55% of all Russian-language messages on Twitter. This big increase in automated activity was largely driven by news-bots contributing to information effects around stories published by the Kremlin's propaganda outlet, Sputnik. On VK, the bot presence also increased, and currently accounts for one quarter of all users. 17% of English language messaging was done by bots.

Three military exercises were of particular interest for Russian-language bots on Twitter and VK: Spring Storm, Baltic Operations (BALTOPS), and Dragon-19. The level of Twitter activity during the month of July was less than half that observed for the period May–June.

Having studied robotic activity for almost three years, we see a clear pattern: whenever a military exercise takes place, coverage by hostile pro-Kremlin media is systematically amplified by inauthentic accounts. In this issue of Robotrolling we take a closer look at how manipulation has changed during the period 2017–2019 in response to measures implemented by Twitter. Since 2017 bot activity has changed. Spam bots have given way to news bots—accounts promoting fringe or fake news outlets—and mention-trolls, which systematically direct messaging in support of pro-Kremlin voices and in opposition to its critics.

We present an innovative case study measuring the impact political social media manipulation has on online conversations. Analysis of Russian Internet Research Agency posts to the platform Reddit shows that manipulation caused a short-term increase in the number of identity attacks by other users, as well as a longer-term increase in the toxicity of conversations.

The Big Picture

Robotrolling analyses the manipulation of information regarding the NATO presence in the Baltic States and Poland on the social media platforms Twitter and VK. Manipulation is achieved through automated accounts (bots) and coordinated, anonymous human accounts (trolls).

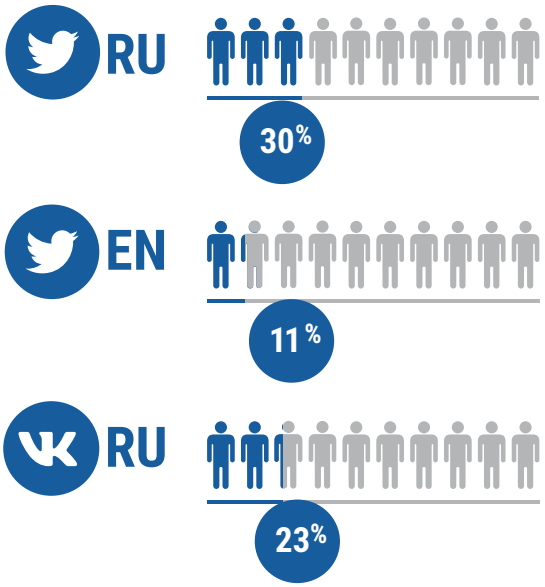
During the period May 1 to July 31, the volume of social media messaging and the number of users mentioning the NATO presence in the Baltics and Poland saw a substantial decrease in comparison to the period February 1 to April 30. On VK, the number of posts was 20% lower, at 22 400 mentions. On Twitter, there was an 11% decline. The number of active users also dropped by 15% on Twitter and 11% on VK.

On Twitter the reduction in absolute numbers was driven by a decrease in interest from English-language users. Russian-language activity increased by two percentage points over the previous quarter. Posts in Russian accounted for 42% of total mentions.

The increase in Russian-language content coincided with an uptick in automated messaging: bots accounted for 55% of all Russian-language messages compared to 47% in the previous quarter. English-language automation also saw a small increase.

The Russian-language space remained dominated by anonymous accounts, bots, and news aggregators. The relative presence of anonymous accounts, however, dropped by seven percentage points to 34% of all Russian-language users. On Twitter the percentage of bot users remained stable in comparison to the previous quarter with 11% of English Twitter accounts and 30% of Russian Twitter accounts. On VK, the percentage of bot accounts increased substantially—to 23%.

On VK, groups remain the main source of posts, accounting for more than 60%. The total number of views for VK posts mentioning NATO's presence in the Baltic States and Poland was 7.3 million. ■



Country Overview

This quarter, Russian-language activity diverged from English-language content. English-language users paid much less attention to military exercises in the region, focusing more on the US decision to deploy additional troops to Poland.

During peaks of activity, Russian-language bots and anonymous Twitter accounts made up over 50% of all posts. VK bots remained a major source of Russian-language messaging.

This quarter, the volume of conversation was distributed similarly across the four countries. Lithuania received the highest number of mentions and bot activity on VK, whilst Twitter mentions in English predominantly referred to Poland and Lithuania.

Human users on Twitter paid special attention to Poland, whereas on VK they wrote about Lithuania and Estonia. Russian-language bot levels on Twitter were the highest when messaging about Estonia.

Estonia

Estonia received the highest number of mentions on Russian-language Twitter—an increase of 16% compared to last quarter. The Spring Storm and BALTOPS military exercises were disproportionately targeted by Russian-language bots on Twitter, with Russian-language bots producing more than half of all messages.

On English-language Twitter, the main Estonia-related coverage occurred on 11 June in response to reports that British aircraft stationed in Estonia had scrambled to intercept two Russian military transport planes. In contrast, Russian-language Twitter messaging centred on Sputnik's coverage of an incident involving a NATO soldier during an exercise in early May.

Latvia

On Twitter, bot activity about Latvia increased with peaks occurring during the Spring Storm, BALTOPS, and Dragon-19 exercises. There were, however, no major spikes in mentions of events involving Latvia. The main coverage occurred during the otherwise quiet month of July. On 28 July, Russian Navy Day, VK bots disseminated Russian state media content and images praising the Russian Navy and the country's military prowess. Simultaneously on VK, a large number of anti-Kremlin bots shared a message emphasising the consistent lack of government investment in the Russian Navy.

Lithuania

In June 2019, Lithuania hosted the multi-national military exercise Iron Wolf 2019. On Twitter, most automated messaging about the exercise pointed to news articles. On VK, Lithuania received the highest number of mentions and was the main target of bot activity. Russian-language Twitter volume increased by 30% compared to the previous quarter. Together with anonymous profiles, bots accounted for over 50% of Russian-language mentions overall.

Poland

Conversations in the Russian-language environment centred on damage suffered by the Polish transport ship ORP Gniezno during BALTOPS and a forest fire in Drawsko Pomorskie, the site of the Dragon-19 exercises.
Russian-language Twitter and VK activity targeting Poland was considerably lower than in the previous quarter. The highest Russian-language activity on both Twitter and VK was registered on 18 June, coinciding with the Dragon-19 military exercise. English-language activity peaked on 13 June when media reported that the US would deploy an additional 1 000 troops to Poland.

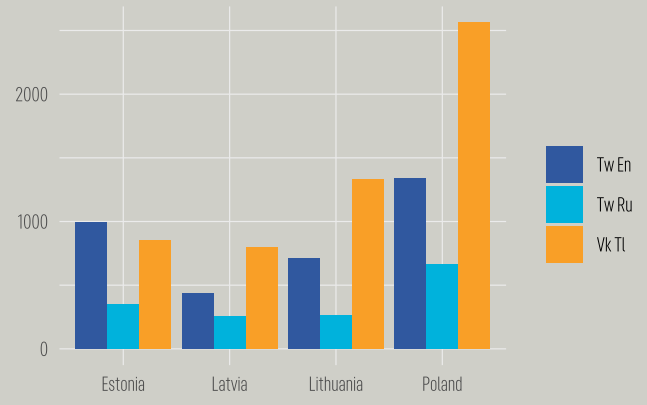


Figure 2: Posts from bots and anonymous accounts for English-language Twitter, Russian-language Twitter, and VK timelines (excluding group posts). Generic references to the Baltics are not included

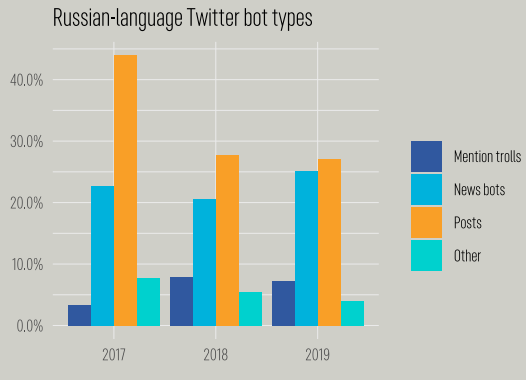


Figure 3: Bot types over time in Russian-language Twitter conversations

Themes

Having studied robotic activity for almost three years, we see a clear pattern: whenever a military exercise takes place, coverage by hostile pro-Kremlin media is systematically amplified by inauthentic accounts. Incidents involving NATO troops in Estonia and Poland were the subject of widely disseminated reports by the propaganda outlet Sputnik.

This quarter, the bulk of the conversation on social media regarding NATO's presence in the Baltics and Poland was closely connected to three military exercises conducted in the region between May and June: Spring Storm (19 April to 10 May in Estonia), BALTOPS 19 (9–21 June, across the region), and Dragon-19 (starting 13 June in Poland and the Baltic Sea). The low volume of posts registered in July compared to May and June shows how social media activity correlates with military exercises in the region.

The decline in the levels of activity observed during July is consistent with an absence of NATO exercises in the region. Deeper analysis shows that incidents arising from or coinciding with military exercises formed the principal focus of conversation.

In the Russian-language space the proportion of tweets from automated accounts declined from 2017 to 2018, but in 2019 this number stabilised and is now on the rise again. This rise is driven primarily by fake news bots and accounts used to artificially promote external content as shown in Figure 3.

Fake Russian-language activity, especially on Twitter but also on VK, centred on events highlighted by the Russian propaganda outlet Sputnik. Sputnik stories about NATO tend to be rebroadcast by dozens of other news outlets, and then widely promoted by social media bots. This was the case on 22 June when a cascade of posts reacted to a Sputnik report, itself based on a report published four days earlier by the NATO Defense College—'Why the Baltics Matter. Defending NATO's North-Eastern Border'.

Another example is a Sputnik News article from 5 May about a NATO soldier who climbed a monument commemorating fallen Soviet soldiers with a grenade launcher during Spring Storm. This received substantial attention from Russian-language social media users. On the day of publication, bot and VK group activity together accounted for nearly 70% of mentions. This story was picked up by other Russian media outlets such as the Russian version of RT (formerly Russia Today) and RIA Novosti, keeping the conversation alive. According to research by DFRLab, the story was widely disseminated among Russian media—54 news stories in Russian and four in English from 6–8 May 2019. VK continued amplifying the story until at least 10 May.

The conversation on both Twitter and VK focused on the risks posed to Russia and to global stability by the NATO presence near the border during these exercises. The manner in which Russian-language human users, bots, and groups engaged with this theme was largely consistent. However, during this period English-language users were less interested in military exercises than has been the norm. Consequently, the Russian-language portion of the conversation increased. ■

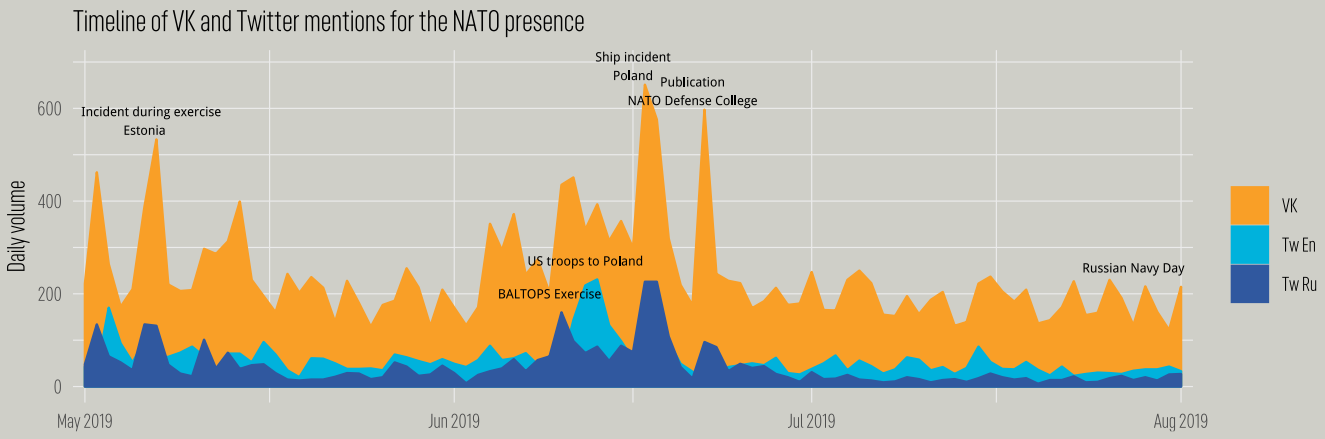


Figure 4: Timeline showing volume of posts on Twitter and VK.

Robo-topics

The Robotrolling series has studied bot activity on Twitter since March 2017. Over this period we have observed some sharp changes in the types of manipulation most common on the platform. In 2017 Twitter made a major intervention to remove bot accounts from their platform. Since then a number of follow-up initiatives have attempted to root out the problem. These interventions have, on the one hand, had the effect of reducing the overall amount of automated activity, and, on the other hand, resulted in the unintended effect of encouraging purveyors of automation to shift to bot types less likely to be detected.

Figure 3 shows the distribution of bot types across three periods: Mar–Dec 2017, Jan-Dec 2018, and Jan–Jul 2019. For each period we show the proportion of all tweets posted by different types of automated accounts. To create the categories we identified all accounts our algorithm believed to be bots, then looked at what type of activity they performed.

In 2017 the majority of bots, labelled 'post-bots', are automated accounts which posted only plain-text messages containing news headlines or links to other platforms. Over time the proportion of post-bots has declined, but not as sharply as one might expect given how primitive these accounts are. In 2019 the volume of post-bots stabilised and has begun increasing again in the latest quarter.

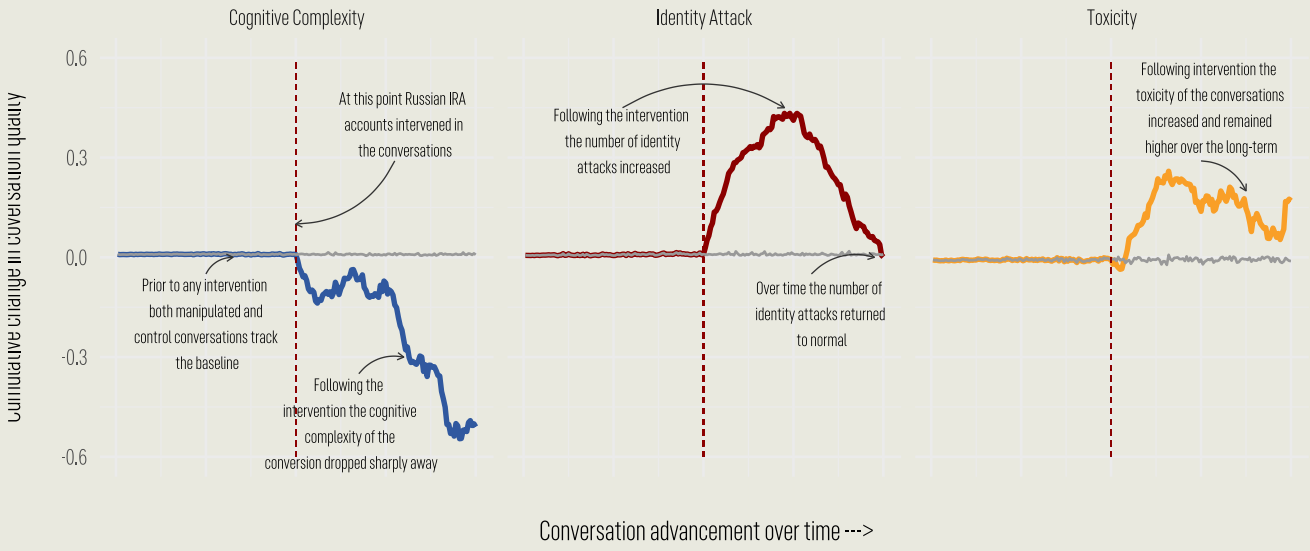
The second category is news-bots. This includes legitimate media outlets such as BBC Russian or Ria Novosti, whose accounts automatically cross-post news articles from their website. However, the vast majority of messages from such accounts promote content from fringe or outright fake news outlets. In 2019, the proportion of such posts has increased, driven in large part by accounts promoting a series of strange news aggregators hosted on subdomains to tula.su (e.g. vk.tula.su). Accounts adopting the fig-leaf of a news-related screen-name are more likely to be tolerated by the platform than other bots.

Mention-trolls are fake accounts that only tweet at (@) other users. They often flood accounts belonging to journalists or institutions with fake feedback. In early 2017 we observed almost no accounts of this type. They emerged in the summer of 2017 and have been a mainstay of Russian-language messaging ever since.

The final category, 'Other' contains hybrid accounts controlled in part by algorithm and in part by humans, as well as other hard-to-label bots. These accounts are quite sophisticated from a technical perspective. It is noteworthy that the proportion of messaging from such accounts has been decreasing, meaning Twitter has been successful in removing a range of fake account types.

Overall, however, the reduction in fake activity observed in 2017 and 2018 has slowed. In the current quarter the proportion of messages from Russian-language bots has in fact increased. ■

Measuring the impact of Russian Internet Research Agency Comments in Reddit Conversations
Injected messages reduced the quality of subsequent conversations



Case study: Impact of Internet Research Agency operations

For the last decade, people worldwide have been targeted with information operations conducted across social media, many originating from the 'troll farm' run by the Russian Internet Research Agency (Russian IRA). This agency aims to influence online conversations about regional, national, and international issues affecting Russian foreign and domestic policy interests. As these campaigns grow steadily in number and scale, they are also gaining international attention. Their tactics have inspired similar campaigns from other nations (e.g. Iran, China), as well as non-state groups (e.g. Daesh, other far-right groups).

While these activities have become more common, measuring their effect in online conversations remains problematic, because it is difficult to predict how the information environment would have looked in their absence. Researchers can count the number of reactions to each Russian IRA story or post and measure the attention attracted by the inauthentic social media accounts, but it is more difficult to measure the subtle influence they may have on promoting Russian interests, engaging with users to sway opinion, and fuelling both sides of controversial online debates.

Recent evidence published in Defence Strategic Communications looks to address this gap. Using natural language programming, statistical modelling, and insights from social psychology, the authors analysed the effect of >16 000 Reddit posts attributed to the Russian IRA. Using a technique called 'causal impact modelling', they estimated the causal effect of an injected comment from the Russian IRA on specific measures of conversation quality. This method predicts how the quality of a conversation would have evolved if the intervention had never occurred and compares it to what actually happened.

The impact of Russian IRA activity on the conversational quality was tested using three measures of conversation quality:

- Cognitive Complexity—how nuanced and multi-dimensional is the conversation? Low scores reflect simplistic thinking or a tendency toward one-dimensional 'echo-chambers'.
- Identity Attack—does the conversation contain negative or hateful comments targeting people due to their race, group membership, or social background?
- Toxicity—how rude, disrespectful, or likely to offend is the conversation?

The results, visualized in Figure 5, show that Russian IRA accounts were able to reduce Cognitive Complexity in subsequent conversations between genuine users. In manipulated conversations fewer viewpoints were expressed compared to unmanipulated ones, and the reasoning became more simplistic. Russian IRA manipulation also caused a short-term increase in the number of identity attacks by other users and a longer-term increase in the toxicity of conversations that persisted over 100 comments after the manipulation. These effects may drive opposing groups farther apart, while a lack of civility will prevent constructive discussion, especially on controversial topics, increasing social polarisation.

These findings highlight the importance of investigating how hostile actors are able to change the nature of online discussions. Whether these effects have repercussions in the offline world remains unknown, but as social media use grows ever more popular, they may have a significant impact on social cohesion. ■

Prepared by Dr. Rolf Fredheim, Belen Carrasco Rodriguez and John Gallagher published by
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO StratCom Centre of Excellence, based in Latvia, is a Multinational, Cross-sector Organization which provides Comprehensive analyses, Advice and Practical Support to the Alliance and Allied Nations.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org