What is Robotic trolling?

Robotic trolling or 'robotrolling' is the coordinated use of fake accounts on social media.

It is important to note that the divide between a bot and genuine account, by which we mean human-controlled account, is not a clear-cut one. Bots should be thought of on a spectrum, from fully automated to varying degrees of partially automated accounts.

Not all bots claim to be humans. We are especially interested in bots which impersonate people or rely on fake identities. Such accounts may be operated either by human-handlers, or by algorithm. It may be impossible to separate the two. As a result, we refer to bot-like activity, that is, activity that could be performed by a computer. Whether it is a human or an algorithm copy-pasting messages is not important for our purposes: in either case, it is an example of deception.

Do you mean most Russian Twitter users are automated? That they are in fact bot accounts?

No. We claim most Russian-language Twitter accounts that mention a specific, narrow set of terms are bots. Whether this claim extends to the rest of the Russian Twittersphere is an open question which requires more research. We intend to tackle this in the next issues of the regular product.

Who is responsible for this activity?

In the online space, attribution is difficult. It will not do to attribute all Russian-language automated activity to the Kremlin. We know there are a host of providers of bot-services in Russia. The bulk of their activity is apolitical. Kelly et al. (2012, p7) found marketing accounts dominated Russian language Twitter spam. Consequently, our preferred terminology is Russian-language bots. If the automated activity is especially politicised it may be appropriate to refer to pro-Kremlin bots. The level of coordination in some of the campaigns points to a state or other well-resourced actor, but we are not making any attributions at this point.

What do the published numbers actually mean, then?

That bots are a big problem. The results indicate that at least one of the following statements are true:

- The Russian Twittersphere as a whole is dominated by bots
- Russian language conversations about NATO activities in Eastern Europe have been systematically swamped by automated activity.

In either case, the findings are very important. Either we are dealing with a broken social media environment, or we are seeing targeted virtual manipulation, or some combination of both.

Why are you studying the NATO presence in the Baltic States and Poland?

We chose a subject that is of relevance to NATO, which is likely to be the target or robotic activity, which is clearly delimited and therefore easily identifiable by search terms. Additionally, as this is a regular product, it needs a subject matter that will remain relevant for a period of time. This thus is an appropriate subject area, given NATO deployments are scheduled to continue at least until 2021.

How did you collect the data?

We use a number of methods including public APIs (application programming interfaces) to access the raw data. Our data collection process is in line with current academic best practice.

What search keyword criteria did you use

We filtered our results based on Boolean searches. To be included in the final analysis a tweet should mention NATO and one of the following states: Estonia, Latvia, Lithuania, or Poland. We collected content for both English and Russian.

And retweets?

They were not considered for this study.

What is the impact of bot activity?

Great question. Measuring impact is a challenge, but one to which we will return in the next few issues of Robotrolling. Stay tuned!

Are there legitimate uses of automation?

Yes, absolutely. Many media outlets automatically post news content to Twitter. Other accounts are humorous and easily identifiable as bots. For instance, the Twitter account that tweets the works of Shakespeare. Many organisational accounts rely on automation to schedule or buffer tweets. But even though many bots are harmless, they should still not be confused with individuals using Twitter.

Twitter is full of accounts aiming to spread messages as widely as possible – this includes bots, media companies, and institutions – but that the number of citizens receiving and discussing this content may be much smaller than is popularly believed.

How do you identify bots?

It's a somewhat involved process. Bear with me:

Our first iteration was based on software developed by academic researchers at Indiana University. They have a publicly available service called Botometer which will estimate whether an account is automated or not. We use the language agnostic version of their service, which basically means linguistic features are excluded from the algorithm. This allows us to estimate scores for Russian language accounts.

Next, we use a combination of algorithms and human coding to iteratively get better results. The human coder is shown some examples of accounts that are likely bots and some that are likely human. If the coder is able to identify the account as either a human or a robot, this information is stored to a database.

Then we use an ensemble of supervised machine learning algorithms to identify which of a few hundred variables are most helpful to predict the scores given by the human coder. As the machine learning algorithms see more data, their predictions are more accurate. Thus, we use the computer to replicate human coding.

Additionally, we use the predictions to identify accounts that may have been misclassified – false positives or false negatives. Fixing mistakes makes the predictions more accurate.

Can your estimates beat the accuracy of human coders?

Often yes. Humans are better at identifying certain types of patterns than are computers, but there are many cases where the computer can make confident assertions where a trained coder might struggle. We give the computer examples of fake and real accounts and ask it to figure out what sets these two groups apart. If this process introduces errors, the output predictions will be flawed.

However, adopting an iterative approach, we are able to gradually remove errors in coding. Secondly, the algorithm often identifies patterns of anomalous user behaviour that the human coder might not spot. Such accounts are given a high probability of being automated, prompting the coder to revisit the examples, and to amend the classification. Thirdly, the computer is better at identifying conspicuously parallel behaviour between accounts: if two or more users, which individually seem perfectly plausible accounts, repeatedly and simultaneously perform specific actions, then at least one of these users is probably automated.

What types of bot activity are you able to identify?

The simplest bots are the easiest to identify. For instance, a user that always tweets links to a single website, or who tweets at mechanical, regular intervals is likely automated. But because only basic programming skills are needed to create bots, and there are numerous 'how-to' guides available online, there are now hundreds of different types of bots active on Twitter. Some of these are hard to identify.

That answer seems a bit evasive – can you give any more detail?

We'd love to, but it's hard to offer too much detail. Bot detection is a game of cat and mouse. Publicly exposing how bots work has at least two undesired effects. First, the botmasters running the fake accounts change their scripts so as to evade future detection. This makes our job harder. Second, knowledge about current practice gives would-be botmasters a headstart when creating their own virtual miscreants. This further pollutes the social media environment. For these reasons, we don't want to reveal too much information about which of the more sophisticated bots we can identify.

But presumably there is some bot activity you cannot yet identify?

Almost certainly. We are continuously looking at our data from new angles, and we often identify new types of suspicious activity. This tells us we can still improve. Moreover, there are probably bots that we will never detect. As a result, the estimates published in our studies may be conservative.

What are the greatest challenges in estimation?

There is always a margin of error. When working with probabilities and thousands of data points it is statistically inevitable that some of the predictions will be wrong. However, this margin goes both ways, and the errors should average out across data our size. We also do not expose individual users as bots; instead we publish aggregate statistics. Just to be safe we also emphasise that our numbers as approximate (rounded to the closest hundred). In the first issue, we say 28% or roughly 1 in 4 English language users are automated. Percentages should be considered to be accurate within a range of ±3 or 4 percentage points.

Our estimates are inexact because of the calculations used, but also because of ambiguous cases that could be classified either as human or as bot. Classifying such accounts is difficult. In practice, our algorithms make a guess about each account. The confidence of this guess is expressed as a probability. For instance, the algorithm might think an account is 75% likely to be a bot. We use a threshold to make a binary decision for the purpose of estimating the total bot population. If the threshold is set at 50%, this account would be deemed a bot. If the threshold were 90%, it would be deemed human (or at least not a bot). Scores nearer to 50% point to a high degree of uncertainty. For the first issue, we used a threshold of 66%.