Twitter Thread by Yoel Roth

Yoel Roth
@yoyoel

We’ve recently seen research about so-called “bots” and misinformation on Twitter and wanted to share our perspective on why findings that might seem remarkable at first are likely inaccurate. We’re working on a more detailed explanation, but some comments for now.

We continue to be excited by the research opportunities that Twitter data provides. Our service is the largest source of real-time social media data, and we make this data available to the public for free through our public API. No other major service does this.

Many researchers, academics, and journalists use our public API — a set of tools for programmatically accessing information on Twitter. We make all public Twitter content available via our APIs. You can learn more about them here: https://t.co/QJQ0USRvI2

The basic issue with much of the research based on our public APIs is simple: The APIs don't provide insight into our defensive actions to protect Twitter from manipulation, including bots.

Because of this, API-based research can't distinguish between accounts we've already identified as bad (and hidden or removed) and real, authentic ones.

This means that our primary actions here — challenging, filtering, and removing bad actors before they have a chance to disrupt people's experience on Twitter — are not reflected.

Why not include this data? Because doing so would make it easier for bad actors to get around our defenses. https://t.co/Q5yweOXc1x

Let’s take a step back and look at the issue of “bots” in general. Even among researchers, there’s little agreement about what “bot” means. It’s a term used to refer to everything from accounts that post automatically to spammers to real people that Tweet something controversial.

The lack of understanding of what a “bot” is and is not contributes to fear, uncertainty, and distrust — in short, unhealthy conversations.

The same way we sometimes see people dismissing facts as "fake news," we also see real people labeling each other as bots rather than engaging with each other — to the detriment of the public conversation.

We've also seen bot detectors and dashboards created by commercial entities, which claim conversations are full of bots, seemingly in an effort to boost their own business models.

When we talk about bots, we mean accounts engaged in platform manipulation and spam. Even then, identifying bots using only public data is very difficult.

Since nobody other than Twitter can see non-public, internal account data, third parties using Twitter data to identify bots are doing so based on probability, not certainty.

One of the most common signals used to predict if someone is a bot is how often they Tweet, or how many times they Retweet. The obvious problem there is, people who are passionate about politics, or sports, or music also Tweet a lot.

Some people only Retweet. There are lots of different ways to use Twitter, and labeling certain uses “bot-like” is unhelpful. Other signals, like political views, the presence of a profile photo, frequency of Retweets, or number of followers seem obvious, but are not clearcut.

These behaviors differ globally, across age groups, language usage, and people’s individual choices about their own privacy and self-expression online.

Many of the common “bot detectors” or “troll hunters” use machine learning techniques to return a “bot score.” What does this actually mean? The answer is very little.

In order to train a machine learning model, you have to start with a training set of users you “know” are bots, so the model can predict whether other users are similar to or different from them.

These tools and approaches are deeply flawed. In our experience, most people aren’t very good at identifying bots from public information alone.

The end result is a staggering margin of error, and one that builds in preconceptions and biases about Tweet volume, political beliefs, and user behavior. These issues rarely make it into media reports, but are often the reasons why some numbers are surprisingly large.

Much of what is being presented as categorical findings is in fact an extrapolated guess and not even close to being accurate. There isn’t really a bot behind every flag. This concern was articulated by one leading researcher in this Buzzfeed piece: https://t.co/WqydQjiYIE

We continue to be committed to enabling academic research, at scale, using Twitter data. Our policies are written to support this work — including when the results are unflattering to Twitter.

However, we believe that to protect our efforts promoting healthy public conversations, there’s a need to speak up here — a lot of this “bot research” is not peer reviewed and not reflective of the facts on any level.

These types of studies, that are covered widely in the media, do not stand up to scrutiny and undermine healthy public conversation, our singular mission as a company.

Oh, and if you see a suspicious account, use our new reporting feature and let us know. It helps our work to make this place better for everyone. Thanks for reading. https://t.co/kypOkCyWk9

> Activity that attempts to manipulate or disrupt Twitter\u2019s service is not allowed. We remove this when we see it.
>
> You can now specify what type of spam you're seeing when you report, including fake accounts. pic.twitter.com/GN9NKw2Qyn
>
> — Twitter Safety (@TwitterSafety) October 31, 2018