RoBhat Labs Follow
Using Data Science and Machine Learning to provide new insights across disciplines.
www.robhat.com
Oct 30, 2017 · 6 min read

Identifying Propaganda Bots on Twitter

Methodology

What is a Propaganda Bot? What are Bots in General?

Bots on Twitter are semi-automated or automated programs that use the normal functions of Twitter, such as tweeting, re-tweeting, and posting content. Unlike other social media websites such as LinkedIn, Twitter allows the use of bots on their platform. For example, there is a bot on Twitter that tweeted every word in the English language and completed the task back in 2014. There are third-party applications that help you get set up with a bot yourself, and Twitter's API allows for automated posts. For the most part, the bots are harmless and publicly reveal themselves as bots.

However, there is a group of political Twitter accounts, commonly referred to as political propaganda bots, that are different. These are accounts who identify themselves as real humans and whose tweets are politically polarizing. They often retweet content instead of actually creating their own. We have seen cases where these accounts are re-tweeting falsified information or advocating for violence or civil unrest. Furthermore, these accounts all seem to be part of the same networks of other political propaganda bots, which allow them to promote content very quickly to real humans on the network. But they try really hard to look like real people. In our analysis of these accounts, we have seen that some of these propaganda bots start off as human-like accounts, suggesting that some of these bots could be from hacked or sold human accounts.

Identifying Bots on Twitter

Inherently, the difficulty of identifying bots on a social media platform like Twitter is the fact there is no way of fully knowing what a bot looks like. Unlike academic datasets, there is no ground truth or labels for these accounts. This is a chicken-and-egg problem: you can't identify bots if you don't know what bots look like, and you don't know what bots look like if you can't identify bots.

In order to get around this dilemma, we identified accounts with certain suspicious behavior as *high-confidence bot accounts*. Behavior such as tweeting every few minutes in a full day, endorsing polarizing political propaganda (including fake news), obtaining a large follower account in a relatively small time span, and constant retweeting/promoting other *high-confidence bot accounts* are all traits that lead to *high-confidence bot accounts*. These are the accounts that we aim to classify and bring to the attention of the Twitter community.

The heuristics above help us identify these *high-confidence bot accounts* in some cases, but this simple rule-based system is bound to have exceptions. For example, a celebrity who recently created a Twitter account will gain a high number of followers in a short period of time. As another example, an avid college basketball fan may tweet about highlight plays every few minutes during the hype of March Madness.

If a rule-based identification system is insufficient, how do we get a machine to identify highly suspicious accounts and bots on Twitter? This is where machine learning comes in.

What is Machine Learning and How Does It Work?

Machine Learning is form of Artificial Intelligence, and it is useful to get a machine to discover trends in data. For example, we can use machine learning to teach a machine how to read human handwriting by looking at thousands of examples of handwriting. The idea behind this is by allowing a machine to look at multiple handwritten examples, the machine learns hidden traits (loops, straight lines, sharp corners) that fundamentally make up the symbols in handwritten text. This is what makes machine learning so powerful; with some clever tweaking, a machine can find some incredible trends in the data without "hardcoding" or "pre-defining" any rules!

In our case, we hope that a machine learns to recognize the differences between regular human profiles and these politically-charged accounts that exhibit bot-like behavior. By feeding a machine different instances of the *high-confidence bot accounts* and normal twitter users, we can hope that the machine understands some internal pattern among the accounts in the training set, and then we can classify any new account.

But how do we generate the examples for the machine to learn in the first place?

Generating a Training Set

One of the drawbacks of machine learning is that it requires lots of examples in order to be effective. Therefore, we needed more data than just the hand-classified *high-confidence bot accounts*. We noticed that a large proportion of followers of these accounts were also *high-confidence bot accounts* as well, presumably to help their network grow.

In order to get more data (without putting in hours on hours on manual human labor classifying each individual account), we added these accounts' followers of these *high-confidence bot accounts* (and applied a few simple heuristics as a filter) in our bot training set.

This allowed us to generate tens of thousands of examples very quickly. The process that we used to generate the training set is not a classifier in itself, as we cannot apply them an general account. By using machine learning on this dataset, we can learn very powerful trends that generalize well and as a result, we can identify the presence of bot behavior of any Twitter account.

After creating a bot dataset, we needed a dataset of non-political, non-bot like accounts in order to train our classifier. We populated this dataset by using verified Twitter profile accounts, since Twitter had already done its due diligence to recognize these accounts as human. In addition, verified profiles have the benefit of representing the wide span of disciplines of user-interests on twitter.

The Classifier

There are hundreds of inputs for each profile that help the model classify a profile as exhibiting political bot-like behavior. Join date, follower count, tweeting rate, retweeting rate, and tweet text are just a handful of traits that model looks at. These are features that can be found on a user's public Twitter profile.

Validating the Classifier

How do we know if the model we build works? We have to validate it so we can have more confidence in how it would perform in a user's hands. It is especially important to minimize false-positives. A classifier like this does not help if it classifies every user who tweets anything about Donald Trump or Hillary Clinton as a bot-like user!

We generated our own validation data that were handpicked by humans to ensure the quality of the validation set. Our classier achieved a 93.5% accuracy on this dataset. This means when the algorithm is queried with a *high-confidence bot account*, we are able to identify this account 93.5% of the time.

As of Monday, November 13th, the false positive rate we're currently seeing cumulatively since launch is 2%. What does this mean?

This number means nothing without understanding the distribution of propaganda bots vs humans during regular use of botcheck.me. As of writing this on Monday, November 13th, we've noticed that of the 20,000+ classifications we've seen since launch, we've been accurate a little over 94.5% of the time.

Why do we find this false positive rate acceptable? Based on the last 2 weeks of use, when botcheck.me has predicted an account in the wild as having Bot-like behavior it has been right over 90% of the time. In other words—given 10 accounts classified as *high confidence propaganda bot behavior* our algorithm will classify around 9 correctly on average.

However, these are the numbers since launch, and we're actively improving our models so that we can improve our accuracy and false positive rates.

Why release the classifier?

We want users of Twitter to query any suspicious accounts they find. Using this classifier that we have developed, we were able to monitor the activity of these accounts and find some interesting patterns in how they behave and interact with the rest of the Twitter community. We described our findings here.

In that post, we describe how these bots spread fake news or falsified information that create great political divide. In addition, these bots create social networks that are capable of creating posts that reach large audiences by retweeting each other. A user who sees these bots in their feed might be susceptible to confirmation bias of their own political beliefs or be deceived into believing false stereotypes of the opposite political spectrum.

Ultimately, we are releasing this classifier for the benefit of the Twitter consumer. We use social networks like Twitter to engage and connect with others. Tools like ours help you be more confident that you're engaging with real, live people.

Sincerely,
Ash Bhat, Rohan Phadte, and the team at Robhat Labs

A special thanks to Joseph Gonzalez, Canzhi Ye, Romi Phadte, Nathan Malkin, Zack Baker, Ananya Krishnaswamy & CS Department at UC Berkeley