CS 460 Spring 2018 Tristan Gurtler (tgurtler)
Final Project

1 Introduction

For my final project, I aimed broadly to investigate malevolent activity on Twitter. I had multiple ideas, though they largely fell into one of two categories: detecting illicit activity by bot accounts posing as people on Twitter, and detecting abusive activity by human individuals on Twitter. This was largely inspired by the, at the time, apparently numerous missteps made by Twitter themselves to handle the issue. As I will discuss in section 3, I seem to have picked a portentous time to do this, as Twitter themselves have recently made significant successes to curb such behavior. While this is good for the ecosystem of Twitter, it does have the unfortunate consequence that my work to detect malevolent behavior was made more difficult by the efforts Twitter took to punish those engaging in it.

This report will proceed as follows. In section 1, we introduce the report. In section 2, we discuss what relevance bot accounts and trolls have to Twitter as a medium. In section 3, we discuss a timeline of efforts Twitter has made to abate such malevolent behavior on its platform, including the consequences on current research. In section 4, we discuss current efforts by academics done to try to detect bot behavior on Twitter. In section 5, we discuss efforts by various groups to try to detect specifically the incidence of "Russian troll factory" accounts – accounts allegedly run by operatives of the Russian Federation for agitprop purposes. In section 6, we outline the tools suggested we would create for this project, and methodological successes and failures we had there. In section 7, we discuss the actual results we observed of the tools we did build. In section 8, we conclude.

2 Motivation

Even from before the 2016 American Presidential Election, individuals have been drawing attention to non-human activity on Twitter and the impact that it could have. As early as 2013, folks were beginning to write out about the prevalence of bots on Twitter, although at that point they were largely seen as either designed to post spam and get users to click suspicious links or to post bizarre and humorous content [1]. As 2016 drew nearer, more attention was drawn to the effect bots had on public discourse regarding the election [2]. Allegations that bots were manipulating the outcome of online social media polls about the candidates, or were being used to astroturf support abounded. In fact, tools used to track bots today show a particular interest in political events; among tweets including the hashtag "#MAGA", allegedly around two-thirds are actually posted by bot accounts [3]. Twitter bots have been shown as being parts of efforts to disseminate fake news across the site leading up to the 2016 election [4]. Today, bots can be shown to post approximately two-thirds of all links appearing on Twitter to popular news and current event sites [5]. Bots have had a large presence on Twitter for several years besides merely doing things like posting gibberish, although even this is of interest in some cases. Twitter specifically has been interested in curbing the use of bots to inflate follower counts and manipulate Twitter algorithms [6].

Meanwhile, abusive behavior on Twitter has also generated interest, whether human or not. Notably, openly identifying members of the American Nazi Party and other white supremacist groups were allowed to use the platform unperturbed until December of 2017 [7]. Actions Twitter has taken to curb abusive behavior have proven controversial at times; most notably alleged bias against conservative users in particular. Project Veritas – the group now infamous for their botched attempt to manipulate

the Washington Post into posting untrue information about Roy Moore in an attempt to discredit other individuals who spoke out against him [8] – in particular claimed to have proof of Twitter shadowbanning conservatives to censor political thought it opposed [9]. Other claims include that swearing at famous people, like Vice President Mike Pence, will get users' accounts locked [10]. Getting these things right are difficult, though clearly worth pursuing.

3 Twitter's efforts at curbing abuse

For a significant period of time, Twitter did not seem to put significant effort into curbing bot or abusive behavior. Prior to 2015 especially, Twitter had a very low track record of banning or suspending users from its platform [11]. 2016 marked an uptick, including notable banning of Milo Yiannopoulos for inciting abusive behavior towards actress Leslie Jones [12]. 2017, however, was when Twitter made itself most clear about its desire to clean up the platform of abusive behavior. Twitter CEO Jack Dorsey publicly promised "a completely new approach to abuse on Twitter"; what this ended up meaning was that tweets deemed abusive would be more difficult to see, hidden behind extra layers of buttons to click [13]. How tweets were deemed abusive was never announced though, a trend Twitter has stated was in order to avoid malevolent users attempting to game the system. Some users claimed that obscene language, like swearing at other users [10] and words generally deemed rude (e.g. "retard") were getting them punished, though other efforts to confirm this seemed to indicate that single acts of such behavior were insufficient to find oneself punished [14]. Twitter notably faced some severe hiccups with this technology, though, banning one Japanese man for making a death threat to a mosquito [15], as well as for sexually harrassing Tony the Tiger [16]. The man banned in the Tony the Tiger case said of the situation, "America has this back-ass-wards Calvinist streak where calling for the expulsion and genocide of non-white races is just a difference of opinion, but making a sex joke at a corporate mascot who paid money to advertise to you is cause for censure." His implication, it was clear, was that Twitter wasn't focusing its attention on places that actually mattered. While another rule change coming late in 2017 ended up more directly addressing this concern [7], Twitter had certainly not had the best track record to that point.

While they were busy attempting to put out the fire of abusive behavior, Twitter seemed not to pay much attention to bots. In early 2018 was when Twitter began to take more drastic efforts to curb bot abuse, specifically making changes to its API to attempt to make it more difficult for spam bots and bots designed to propagate information widely across multiple accounts to operate [17]. The timing of this announcement, shortly after Twitter confirmed that over 50,000 bots with ties to Russia appeared to have attempted to interfere with the 2016 Election, seems to acknowledge the large problem Twitter has had with bots. This too, however, has not been an effort without issues. Very recently, Twitter had a run of users negatively impacted by an algorithm designed to detect spam tweets; specifically, users who used the word "thanks" or "thank you" in tweets found their accounts limited until they proved they were human [18].

While there are significant methodological issues for a student project on automatically identifying abusive behavior (that will be covered in section 6), another notable problem that I observed was that it actually was difficult as a normal user to find abusive content. While it certainly still existed, the efforts Twitter has taken to "limit its reach" do legitimately mean that bare effort attempts to, for example, find Nazi-affiliated accounts are next to impossible (this would seem to make sense, as if they weren't, Twitter would likely have banned them already as part of the late 2017 policy changes). This made my work significantly more difficult than I anticipated when I initially set out on this project.

4 Current techniques to detect bots

In academic circles, between detecting bots and detecting abusive or manipulative human behavior, it appears that most work has focused on detecting bots. This seems relatively unsurprising, as unlike essentially performing some kind of "truth detector" like ritual on an account to determine its sincerity, bots could unwittingly leave traces of their inhumanity that could be tracked. I personally found three relatively large (numerous publications in larger conferences/journals) projects, and will discuss their main methodologies here.

4.1 Botometer

The most comprehensive project, by far, is Botometer [19]. Mostly conducted out of Indiana University, the project aims to detect bot behavior with a claimed 1,150 features fed into a machine learning amalgamation. These features came from six different broad domains. Four of these are language-independent: User features (features revolving around a single account, like if it has a default profile image, or how frequently it posts), "Friends" features (features revolving around other users the account in question retweets/mentions and is retweeted/mentioned by), Network features (largely, applying PageRank algorithms to the networks of retweeting/mentioning a particular account finds itself in to determine the topology that the account usually posts to), and Timing features (features revolving around the length of the time intervals between when a user posts or retweets things). The other two domains are largely English specific: Content features (features revolving grammaticality and understandability of posts made by an account) and Sentiment features (features revolving around things like how positive or negative a tweet was). This veritable deluge of features allows Botometer to claim accuracy rates as high as 89% (suspiciously, higher than their own inter-annotator agreement) [20]. I attempted to replicate this tool (as seen in section 6); in doing so, I noticed two particular things. First, although at [19] it appears that the tool can largely be used standalone on any account, many of the features it claims to use actually track accounts over time (for example, watching to see how many follows it has over time, to try to detect if the account churns through accounts it follows quickly or grows a group of people to follow more naturally); this would indicate that either the tool has a massive database it keeps updated with information about a huge number of Twitter accounts, or that its release form features a more relaxed version with fewer features. Due to this, in my replication, I attempted to replicate the publicly available tool, instead of the original research tool (and thus had lower accuracy, around 80%). Second, Twitter rate-limiting is particularly odious and, well, limiting – even working with a limited pool of accounts to keep track of, it would be a wonder for their tool to keep up with as many accounts as they claim to; merely doing one iteration on every account would take (at the rates available to apps that don't pay Twitter, at least) approximately two days, at minimum, due to certain API calls it makes being particularly limited. As such, I worked with a significantly smaller, random pool of accounts, also leading to my lower accuracy.

4.2 DeBot

The next substantial project, out of the University of New Mexico, is DeBot [21]. Their main technique is to observe that bots at some times seem to directly copy the activity of other accounts. Two seemingly unrelated accounts will at times post the exact same concept; doing this for long enough is indicative of automated behavior. Although [20] claims that a small number of bots will behave this way, DeBot claims to have found over 730,000 distinct bots since February 2015. It remains to be seen if their technique to detect bots will withstand the API changes Twitter made in early 2018 that more directly prohibit the copying activity between multiple accounts DeBot relies on to discover bots [17].

4.3 NYU Bot Detection

There also appears to be a group operating out of NYU working on bot detection on Twitter [22]. It is unclear exactly what techniques they use to detect bots; only that their focus was specifically on two periods of time relevant to Russian politics (most of 2014 and most of 2015, separately). Based on what they report, they may have used similar techniques to Botometer, as they directly claim that the single most powerful predictor for an account being a bot is if it posted via web instead of via a mobile phone. What features besides this (and retweet rate and geo-location being turned on) they actually may have used, though, are not clear. It is also notable that they confirm that bots seem to have a large hand in spreading news stories and potentially manipulating which news stories have attention drawn to them (similar to [4]).

5 Current techniques to detect trolls

Although Russia isn't the only country to use social media agitprop for its own devices [23], it certainly has the most attention in the United States. Twitter identified 2,752 accounts tied to Russia's "Internet Research Agency" troll farm [24], and more still appear to exist [25]. While we can observe their behavior (trying to blend in as American citizens, local news agencies, and accounts tied to local political groups), and watch as they do much the same as we've observed in bots (focus on distributing specific information, rather than directly inventing it) [26], this doesn't give us many pointers on how to actually detect them. We do get some clues; posting is significantly more frequent on these types of accounts during normal business hours in Moscow instead of during times Americans might expect to be awake, but this by itself is not proof. Some groups have given advice for things to watch out for (grammatical mistakes around words like "a" and "the", which do not have equivalents in Russian, or regarding question word order, which differs between Russian and English), but these aren't sufficient either [27]. Probably the best advice is just to watch out for accounts that, if you go back through their history, in particular seem to follow along with the narrative that the Kremlin is trying to establish for particular news stories. However, identifying this automatically seems very difficult to do (how can you direct a computer to identify somebody holding to a specific narrative about a news story?), and even for humans this takes an inordinate amount of time and research every time you encounter somebody you might have reason to check up on. At large, although the problem of detecting bots automatically seems to have been reasonably well-studied, detecting troll behavior automatically seems considerably more difficult.

6 Tools described at project outset

In my project proposal, I outlined four tools that I was interested in building in regards to the problems observed across this paper. I wanted to replicate and improve on current techniques of identifying bots on Twitter, to replicate and improve on current techniques of identifying trolls on Twitter, to detect brigading on Twitter, and to detect abusive behavior beyond just swearing on Twitter. In the end, I was only able to partially accomplish the first of these; I will now discuss why that is.

6.1 Replication/Improvement of bot detection

As discussed in section 4, technological means for detecting bots is actually fairly mature. Botometer [19] in particular is very well fleshed out with relatively high success rates. I spent a good chunk of my coding time just understanding what this system actually even did to make its measurements and build its ML system. I noticed two particular things about the system, as mentioned earlier. The tool

that is publicly advertised seems able to be used standalone on any account, even though many of the features it claims to use actually track accounts over time. This would indicate that either the tool has a massive database it keeps updated with information about a huge number of Twitter accounts, or that its release form features a more relaxed version with fewer features. Due to this, in my replication, I attempted to replicate the publicly available tool, instead of the original research tool. Second, Twitter rate-limiting is particularly limiting, and as such, I worked with a significantly smaller pool of accounts. Even so, collecting data took several hours. Another note I had was that some of its features depended on language specific features; this too would have required its own corpus of "valid Twitter speech"; I largely did not replicate this, as they did not outline very clearly where they obtained such a corpus. This replication can be found in my repo.

6.2 Replication/Improvement of troll detection

At the outset of this project, I did not realize anywhere near how daunting a task identifying the behavior of trolls is. We could broaden this to detect "bad faith" actors (i.e. individuals who seem to be arguing in bad faith with another individual or otherwise intending to aggravate others), which would avoid the inherent magic of attribution that we would otherwise need to implement, but even then we need some form of gold standard from which we can decide whether somebody is acting in bad faith or not. To do this, we would need a pool of accounts to investigate and human annotators; in order to get a sufficient number to work with, we'd likely need a good number of annotators to do that work. Generating this dataset quickly goes out of scope from what an undergrad with a budget of $0 can do to something that actually needs monetary support to be able to accomplish (since, to my knowledge, no such corpus already exists, unlike with the bot detection systems). I did consider even more significant relaxations of the question at hand (specifically, can I detect the account of a Nazi), but ran into the problem that if an account belonged obviously enough to somebody who identified as a Nazi (e.g. swastikas in their profile pic), they would have certainly already been banned. Though this doesn't mean that there are no Nazis on Twitter, finding them would then be subject to much the same problem as the "bad faith" detection. I decided that it was unfeasible to attempt this portion of the project.

6.3 Brigade detection

While it is notable that Twitter publicly states that "the number of reports we receive does not impact whether or not something will be removed" [28], Twitter does not release any way for non-Twitter-affiliated individuals to be made aware or to discover when people report tweets or accounts. That is, to design a tool that detects when people fraudulently report an account in order to flag it for abuse, you need to know when people are reporting an account. Since this is not something Twitter makes available, there is no way to implement this tool as it stands.

6.4 Abusive behavior detection outside of swearing

This proposal involves a catch 22 I did not realize when I submitted it. In order to determine if something is "abusive behavior", I need to define what "abusive behavior" is. Twitter, of course, is notoriously vague as to what that means, and determining it for myself is non-trivial. If I attempt to use Twitter to determine this (that is, going with my original idea of detecting when accounts were flagged for being abusive and determining if it was just based on recent posts), then the first problem is that the problematic behavior will not be possible to be viewed, as by definition when Twitter flags accounts in this way none of their posts can be seen. This is, of course, ignoring the fact that determining when an

account has been flagged is itself non-trivial (it does not appear to be a part of the public API to see when an account is suspended or banned or such). Thus, I had to additionally give up on this project.

7 Results of implementing tools

In my attempt to replicate Botometer [19], I used a much smaller sampling of accounts. 64 accounts belonged to human users, and 48 accounts belonged to bots. I also used a much smaller feature set: only focusing on the User, Timing, Friend, and Content features, and at that only focusing on the features that could be collected in one sitting, to replicate the standalone tool. Opposed to the 89% accuracy found by the study correlated to Botometer [20], I obtained an average accuracy of 80% (and more specifically, a precision of 79.5% and a recall of 78.6%) when using 10-fold cross-validation; considering my extremely small dataset and limited feature set, I am very pleased that my replication was able to achieve so close to the very high figure of the original paper. I also wish that the original paper explained its choice of features more clearly, as it was very nonobvious what their Network features actually measured, and no explanation or code was available pretty much anywhere (humorously, there is a github repo associated with the project, that only includes code to query their own API, behind which they hide their actual implementation).

8 Conclusion

Although there is good reason to be interested in the state of Twitter and abusive behavior/bot abuse, particularly regarding more recent studies observing the very high rates of bot participation on the network [5], Twitter has in the past six months especially taken drastic steps to curb bot abuse and abusive behavior, to the point that finding abusive behavior in the wild in particular is actually noticeable difficult. It seems that there is some ways to go for Twitter, in particular regarding bot detection (most of the bots identified by these external services are still happily plugging away on Twitter, although some are beginning to go missing), but steps are being made in the right direction. It is a pleasant surprise to see that a social media site is making successful bona fide steps to clean its platform up.

References

[1] R. Dubbin, "The rise of Twitter bots," *New Yorker*, November 2013. [Online]. Available: https://www.newyorker.com/tech/elements/the-rise-of-twitter-bots

[2] D. Guilbeault and S. Woolley, "How Twitter bots are shaping the election," *Atlantic*, November 2016. [Online]. Available: https://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/

[3] C. Daileda, "Botcheck.me will tell you whether that Twitter account is fake," *Mashable*, November 2017. [Online]. Available: https://mashable.com/2017/11/02/botcheck-california-students-twitter-fake-accounts-bots/#DBtlL4DOaOqU

[4] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Anatomy of an online misinformation network," *PLOS One*, April 2018. [Online]. Available: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196087

[5] S. Wojcik, S. Messing, A. Smith, L. Rainie, and P. Hitlin, "Bots in the Twittersphere," April 2018. [Online]. Available: http://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/

[6] N. Confessore, G. J. Dance, and R. Harris, "Twitter followers vanish amid inquiries into fake accounts," *New York Times*, January 2018. [Online]. Available: https://www.nytimes.com/interactive/2018/01/31/technology/social-media-bots-investigations.html

[7] "Twitter suspends Britain First leaders," *BBC News*, December 2017. [Online]. Available: http://www.bbc.com/news/technology-42402570

[8] S. Boburg, A. C. Davis, and A. Crites, "A woman approached The Post with dramatic — and false — tale about Roy Moore. She appears to be part of undercover sting operation." *Washington Post*, November 2017. [Online]. Available: https://www.washingtonpost.com/investigations/a-woman-approached-the-post-with-dramatic--and-false--tale-about-roy-moore-sje-appears-to-be-part-of-underco 2017/11/27/0c2e335a-cfb6-11e7-9d3a-bcbe2af58c3a_story.html?utm_term=.f02fb35337df

[9] T. Hains, "'Project Veritas' hidden camera: Twitter and Reddit use 'shadow ban' algorithms to censor political opinions," *Real Clear Politics*, January 2018. [Online]. Available: https://www.realclearpolitics.com/video/2018/01/12/project_veritas_hidden_camera_twitter_and_reddit_do_shadow_ban_certain_political_opinions.html

[10] R. Brandom and C. Newton, "Twitter is locking accounts that swear at famous people," *The Verge*, February 2017. [Online]. Available: https://www.theverge.com/2017/2/24/14719828/twitter-account-lock-ban-swearing-abuse-moderation

[11] "Twitter suspensions." [Online]. Available: https://en.wikipedia.org/wiki/Twitter_suspensions

[12] M. Isaac, "Twitter bars Milo Yiannopoulos in wake of Leslie Jones's reports of abuse," *New York Times*, July 2016. [Online]. Available: https://www.nytimes.com/2016/07/20/technology/twitter-bars-milo-yiannopoulos-in-crackdown-on-abusive-comments.html

[13] C. Newton, "Twitter begins filtering abusive tweets out of your replies," *The Verge*, February 2017. [Online]. Available: https://www.theverge.com/2017/2/7/14527104/twitter-reply-filters-safe-search-abuse-harassment

[14] K. Flynn, "Just swearing at Trump (probably) won't get you in Twitter timeout," *Mashable*, February 2017. [Online]. Available: https://mashable.com/2017/02/24/twitter-timeout-trump/#HEIb5rHAliqN

[15] P. Harrison, "Man banned from Twitter over mosquito death threat," *BBC News*, August 2017. [Online]. Available: http://www.bbc.com/news/blogs-trending-41097947://www.bbc.com/news/technology-42402570

[16] R. Broderick, "This guy was suspended from Twitter after he sexually harassed Tony the Tiger," *Buzzfeed News*, August 2017. [Online]. Available: https://www.buzzfeed.com/ryanhatesthis/you-can-release-your-great-all-inside-me-daddy?utm_term=.kp9qqDp9X#.odXGGZDd4

[17] J. Russell, "Twitter is (finally) cracking down on bots," *TechCrunch*, February 2018. [Online]. Available: https://techcrunch.com/2018/02/22/twitter-is-finally-cracking-down-on-bots/

[18] J. Hathaway, "Is Twitter banning users for saying 'thank you'?" *Daily Dot*, April 2018. [Online]. Available: https://www.dailydot.com/unclick/twitter-ban-thank-you/

[19] Indiana University Network Science Institute and Center for Complex Networks and Systems Research, "Botometer." [Online]. Available: https://botometer.iuni.iu.edu/#!/

[20] *Online Human-Bot Interactions: Detection, Estimation, and Characterization*, May 2017. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

[21] N. Chavoshi, H. Hamooni, and A. Mueen, "Debot." [Online]. Available: http://www.cs.unm.edu/~chavoshi/debot/

[22] J. Devitt, "How to detect russian bots on twitter." [Online]. Available: https://www.futurity.org/russian-bots-twitter-1633842/

[23] P. Vogt and A. Goldman, "The prophet," December 2017. [Online]. Available: https://www.gimletmedia.com/reply-all/112-the-prophet

[24] D. Frommer, "Twitter's list of 2,752 Russian trolls," *Recode*, November 2017. [Online]. Available: https://www.recode.net/2017/11/2/16598312/russia-twitter-trump-twitter-deactivated-handle-list

[25] Alliance for Securing Democracy, "Hamilton 68." [Online]. Available: https://dashboard.securingdemocracy.org/

[26] W. Lyon, "The story behind Russian Twitter trolls: How they got away with looking human – and how to catch them in the future," March 2018. [Online]. Available: https://neo4j.com/blog/story-behind-russian-twitter-trolls/

[27] Atlantic Council, "#TrollTracker: How to spot Russian trolls," March 2018. [Online]. Available: https://medium.com/dfrlab/trolltracker-how-to-spot-russian-trolls-2f6d3d287eaa

[28] Twitter, "Hateful conduct policy." [Online]. Available: https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy