PLOS ONE

RESEARCH ARTICLE

Bots, disinformation, and the first impeachment of U.S. President Donald Trump

Michael Rossetti[1,2☯]*, **Tauhid Zaman**[3☯]

1 Operations and Information Management, Georgetown University, Washington, District of Columbia, United States of America, **2** Technology, Operations, and Statistics, New York University, New York, New York, United States of America, **3** Operations Research, Yale University, New Haven, Connecticut, United States of America

☯ These authors contributed equally to this work.
* mjr300@georgetown.edu

Citation: Rossetti M, Zaman T (2023) Bots, disinformation, and the first impeachment of U.S. President Donald Trump. PLoS ONE 18(5): e0283971. https://doi.org/10.1371/journal.pone.0283971

Editor: Alexandre Bovet, University of Zurich, SWITZERLAND

Received: June 24, 2022

Accepted: March 21, 2023

Published: May 8, 2023

Data Availability Statement: We have uploaded the data to a public GitHub repository (https://github.com/s2t2/tweet-data-2020). We have included the Tweet IDs for researchers who wish to recollect the data from Twitter. We also have included some user analysis results, but have anonymized the user identifiers and excluded their names and screen names, to respect user privacy.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Automated social media accounts, known as bots, have been shown to spread disinformation and manipulate online discussions. We study the behavior of retweet bots on Twitter during the first impeachment of U.S. President Donald Trump. We collect over 67.7 million impeachment related tweets from 3.6 million users, along with their 53.6 million edge follower network. We find although bots represent 1% of all users, they generate over 31% of all impeachment related tweets. We also find bots share more disinformation, but use less toxic language than other users. Among supporters of the Qanon conspiracy theory, a popular disinformation campaign, bots have a prevalence near 10%. The follower network of Qanon supporters exhibits a hierarchical structure, with bots acting as central hubs surrounded by isolated humans. We quantify bot impact using the generalized harmonic influence centrality measure. We find there are a greater number of pro-Trump bots, but on a per bot basis, anti-Trump and pro-Trump bots have similar impact, while Qanon bots have less impact. This lower impact is due to the homophily of the Qanon follower network, suggesting this disinformation is spread mostly within online echo-chambers.

Introduction

On December 18, 2019, the United States House of Representatives voted to approve articles of impeachment against President Donald Trump. The resulting trial in the Senate concluded on February 5, 2020 when the Senate voted to acquit the president. During this period, online social media platforms became a battlefield for information warfare between supporters and opponents of the president [1].

While much of the activity originated from users engaging in genuine political debate, a significant proportion came from accounts known as *bots*. Social bots are automated social media accounts programmed to share certain content and interact with other users [2]. Different kinds of bots are programmed for different purposes, including traditional spam bots which aggregate news content or distribute links [2], financial bots which advertise commercial products or attempt to influence financial markets [2], "astroturf" bots which promote political

figures and their policies [2, 3], and fake follower accounts used for inflating a user's follower network [2].

Due to their automation, social bots have the potential to manipulate social media discussions [2, 4, 5]. The use of bots to spread targeted political messages online, known as computational propaganda [6], has been growing in recent years. Studies have chronicled efforts by bots to manipulate online discussions surrounding U.S. elections and political events since 2016 [7–13].

Recent trends show bots amplifying and spreading false or misleading news stories and conspiracy theories (collectively known as *disinformation*). The phenomenon of bots spreading disinformation has been observed in online discussions ranging from U.S. elections [13] to public health issues such as vaccines [14, 15] and the COVID-19 pandemic [16]. The introduction of disinformation makes the risk posed by bots much greater, as it allows bots to create false narratives that take hold with a large population, resulting in dangerous outcomes. For example: on December 4, 2016, an armed believer of the "Pizzagate" conspiracy theory fired his weapon in a D.C. pizza parlor [17–20]; and on January 6, 2021 a mob of Trump supporters influenced by the "Qanon" conspiracy theory would storm the U.S. Capitol building in an effort to forcibly overturn the results of the 2020 U.S. presidential election [21–23]. Due to the unique threat bots pose, it is important to be able to identify the bots and quantify their impact in spreading disinformation.

There is a large body of research on bot detection. Many bot detection methods rely on analyzing characteristics and behaviors of individual accounts. Some methods in this category focus on the temporal behavior of accounts [24, 25]. Others focus on the text posted by the accounts [26–28]. Many approaches combine a variety of features of an account and use them as input to a machine learning classifier [29–32]. The most prominent method in this class is the Botometer (formerly BotOrNot) [33] which uses a machine learning model applied to all data of an account (tweets, profile, followers). Botometer has been used in a variety of academic studies of social media bots [4, 11, 12, 15, 34, 35].

It has been observed that methods which look at accounts individually may not be able to detect evidence of coordination between groups of accounts [31, 36, 37]. For this reason, other methods have been developed which focus on analyzing collective behaviors and similarities among groups of accounts [36–40], or including such analysis in a hybrid "tiered" approach [41]. It has been found that some of these collective behavior based methods outperform methods which look at accounts individually [40].

In terms of attempts to quantify bot impact, existing studies provide simple statistics such as the number of bots or volume of content they share [11, 12]. However, these statistics do not incorporate the interaction of social network structure, user activity levels, and user sentiment. Some studies have looked individually at the positioning of bots within the social network [42] or the bot retweet response time [4], but not the interaction of these factors. Recent work has presented a novel network centrality measure known as *generalized harmonic influence centrality* that combines all of these factors to assess bot impact in online political discussions [40, 43]. This centrality measure provides a better measure of bot impact than individually examining each factor.

In this study, we focus on social bots discussing the first impeachment of U.S. President Donald Trump on Twitter. We find that bots are 66 times more active than normal human users, producing nearly one third of all impeachment-related content, despite representing less than 1% of all accounts discussing the impeachment. Bots tend to share news from lower quality sources (including disinformation) than their co-partisan human counterparts. Bots have an unusually high prevalence among Qanon conspiracy supporters, indicating that efforts are being made to artificially amplify this disinformation. Using generalized harmonic

influence centrality, we show that pro-Trump and anti-Trump bots have a similar level of impact per bot, but Qanon bots have a lower per bot impact. Analysis of the Qanon follower network suggests this lower impact is due to Qanon users existing in an online echo-chamber with a high amount of ideological homogeneity.

Results

Our analysis contains multiple steps, which are shown in Fig 1. We begin by collecting social media data related to the impeachment. Then we apply a variety of methods to classify the accounts by partisanship, bot status, and Qanon support. We perform a variety of analyses comparing the different types of accounts in terms of their posted content and network structure. Finally, we quantify the impact of the bots using generalized harmonic influence centrality.

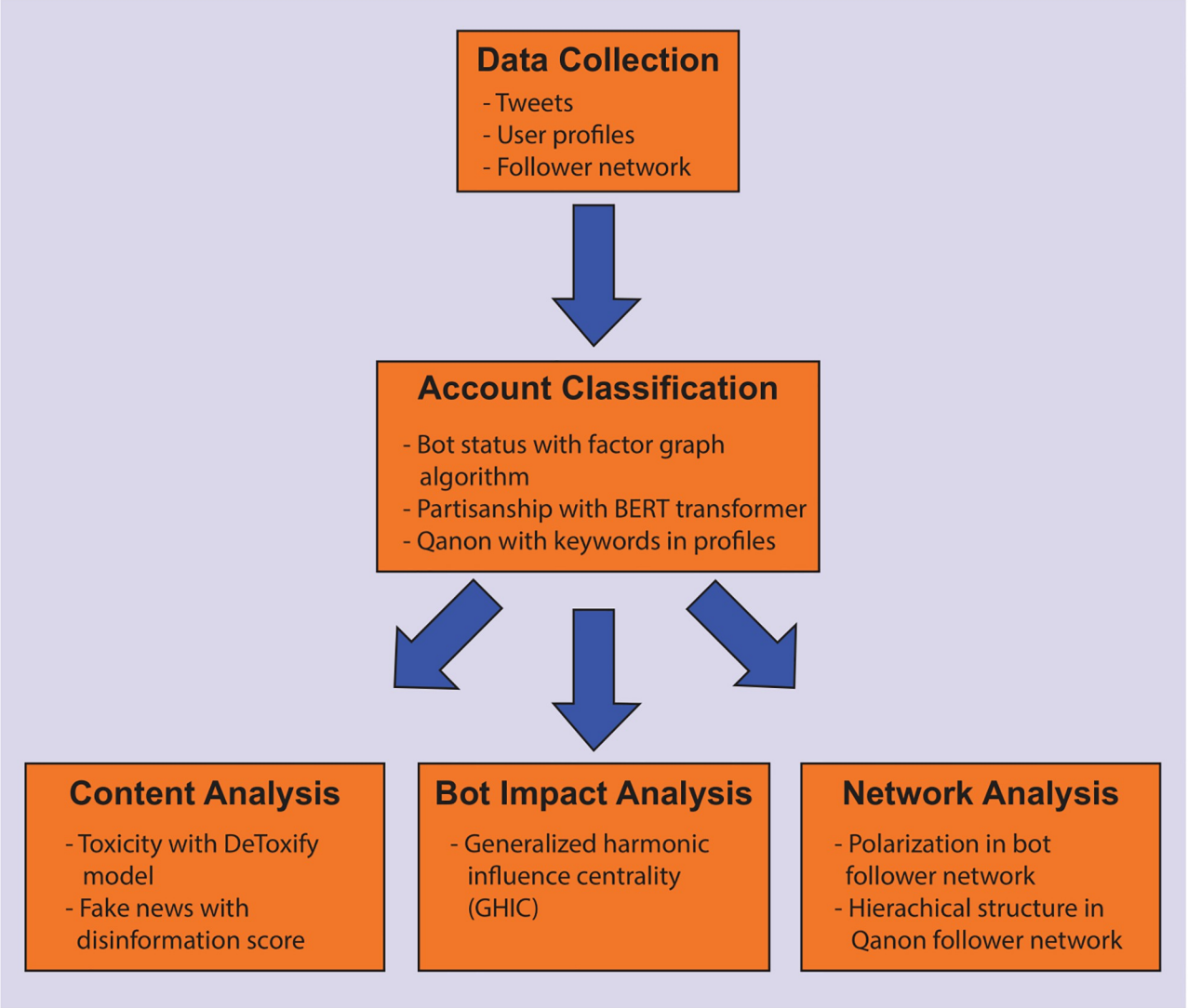

Fig 1. Diagram of the different steps taken in our analysis of bots and disinformation in the Twitter discussion surrounding Donald Trump's first impeachment.

https://doi.org/10.1371/journal.pone.0283971.g001

Table 1. Data collection results, by account type.

Partisanship	Bot status	Qanon status	Accounts	Tweets
Anti-Trump	Human	Normal	2,273,831	25,103,340
Pro-Trump	Human	Normal	1,279,638	18,717,915
Pro-Trump	Human	Q-anon	22,926	2,845,521
Pro-Trump	Bot	Normal	11,571	9,880,481
Anti-Trump	Bot	Normal	10,145	9,220,258
Pro-Trump	Bot	Q-anon	2,434	1,899,042
			3,600,545	**67,666,557**

Statistics of account types and tweets in Twitter dataset for the first impeachment of U.S. President Donald Trump.

https://doi.org/10.1371/journal.pone.0283971.t001

Data collection

We constructed a list of keywords and hashtags related to the first impeachment of President Donald Trump. This list included partisan terms advocating for each side in the discussion, and non-partisan terms related to developments in the impeachment (see Methods). From December 12, 2019 to March 24, 2020, we collected tweets which contained at least one of these keywords. This collection process resulted in a dataset of 67.7 million tweets posted by 3.6 million unique Twitter accounts. We also collected the user profiles of these accounts and their 53.6 million edge follower network (see Methods). We provide summary statistics of this dataset in Table 1.

Account classification

We classified the Twitter accounts into different groups which characterized different aspects of their preferences and behavior. These included political partisanship, support for the Qanon movement, and whether or not the account was an automated bot. Each group label was assigned using a different method. A summary of the group statistics is provided in Table 1.

First, we considered the political partisanship of the accounts. We trained a bidirectional encoder representations from transformers (BERT) model to measure the political partisanship of any given tweet text [44]. The model was trained using a subset of the impeachment tweets for which we were able to assign ground-truth labels (see Methods). A label of zero represents strong anti-Trump sentiment and a label of one represents strong pro-Trump sentiment. We used the trained model to measure the sentiment of all tweets in our dataset. Then, we assigned a partisanship score to each Twitter account equal to the mean value of the partisanship score of their tweets. The accounts were labeled anti- (pro-) Trump if their partisanship score was less than or equal to (greater than) 0.5. In total, our dataset had 2.3 million anti-Trump accounts and 1.3 million pro-Trump accounts. This left leaning bias in the Twitter conversation aligns with findings from previous studies [40].

We next identified accounts who were supporters of the Qanon movement, based on those who promoted Qanon hashtags used in their profile descriptions (see Methods). Using this process, we identified 25,360 Qanon accounts.

To identify bots we used an algorithm based on a factor graph model [40]. This algorithm simultaneously detects multiple bots among a set of Twitter accounts, using only their retweet network (see Methods). We chose this algorithm because it had minimal data requirements compared to other algorithms [33] (as we are easily able to construct the retweet network from

the collected tweets), and also because it exhibited performance similar to or better than other algorithms [40]. We applied the bot detection algorithm to daily retweet networks constructed from tweets posted each day. This allowed us to identify daily sets of active bots. The algorithm we use is designed to identify bots that engage in excessive retweeting. These retweet bots produce a disproportionate volume of social media content that can distort online discussions, so we focus our analysis on this type of bot. Our detection algorithm will not identify bots which exhibit different behavioral patterns that do not consist of excessive retweeting. Therefore, when we classify an account as a bot, we specifically mean a bot retweeting at an unusually high rate. We note that one limitation of our bot detection algorithm is that it relies upon the retweet network, and so can only detect bots who retweet someone. However, since our focus is on retweet bots that amplify certain voices, this limitation is not a major issue.

Our final set of bots consisted of the union of these daily bot sets. In total we found 24,150 bots, of which 10,145 were anti-Trump and 14,005 were pro-Trump. This ratio of bots to humans aligns with other studies of U.S. politics on Twitter [13, 40].

We summarize the prevalence of bots in different groups in Fig 2. We see that bots have a slightly higher prevalence among pro-Trump accounts than anti-Trump accounts (p-value $< 10^{-6}$). However, the bot prevalence among Qanon supporters is nearly an order of magnitude larger than it is among normal accounts (p-value $<10^{-6}$). This suggests malicious actors may be attempting to use artificial accounts to amplify Qanon content. On one hand, this may be reassuring, as it suggests there aren't as many real Qanon supporters on social media as there may appear. On the other hand, significant bot presence means that Qanon content is being spread at a higher rate and with a potentially higher reach than could be achieved with humans alone. This high bot fraction is even more concerning when looking at the tweet rate of the accounts in Fig 2, as Qanon human supporters tweet approximately ten times more frequently than regular humans (p-value $< 10^{-6}$), and bots tweet approximately

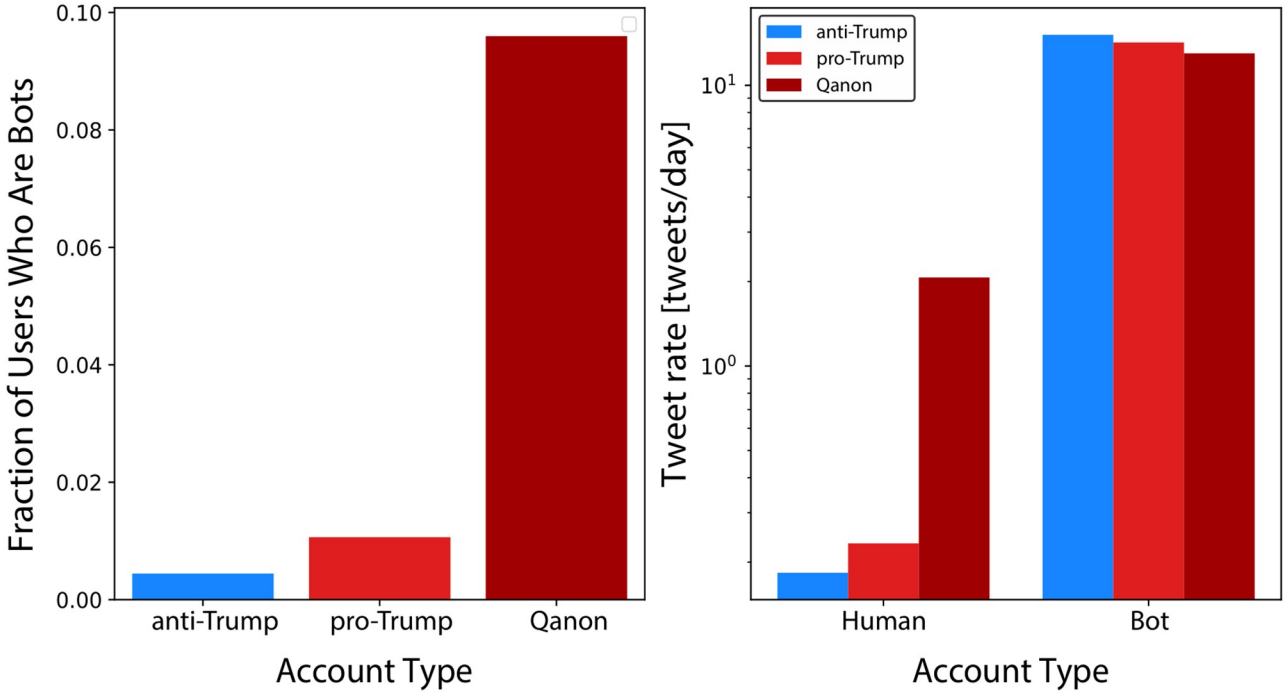

Fig 2. Bot prevalence and activity rates. (left) Fraction of bots and (right) average tweet rate of each account type category.

https://doi.org/10.1371/journal.pone.0283971.g002

one hundred times more frequently than regular humans (p-value $< 10^{-6}$). Having a large number of Qanon bots can lead to an unusually large amount of Qanon content being spread through Twitter. This potentially enhanced reach increases the risk posed by an already dangerous ideology.

Content analysis

There are differences in the nature of content posted by the various account types. We first considered the quality of news stories shared by the accounts. We leveraged a previously published set of 60 news sites (20 mainstream, 20 hyper-partisan, and 20 fake news, with liberal and conservative leaning sites in each category) whose trustworthiness had been rated by eight professional fact-checkers [45]. We followed the approach used in prior work [46, 47] and calculated a media quality score for each user by averaging the trustworthiness ratings of any of their impeachment related tweets that contained links to any of those 60 sites. There was a link of some sort in 11% of the tweets, and of those with links, 12% had a link from one of the 60 news sites. In total, we were able to calculate a media quality score for 217,692 users. The media quality score ranges from one to five, with higher values indicating higher trustworthiness. Like other researchers in this space [46–49], we use source trustworthiness as a proxy for article accuracy, because it is not as feasible to rate the accuracy of every shared link. We note that our analysis of news quality is based on accounts in our dataset who share news stories, so care must be taken when generalizing to a broader set of accounts. Our conclusions apply only to the subset of accounts of each type who choose to share news stories.

Fig 3 shows the average media quality score of the different account types. The first striking observation is that the media quality score is much higher for anti-Trump accounts than pro-

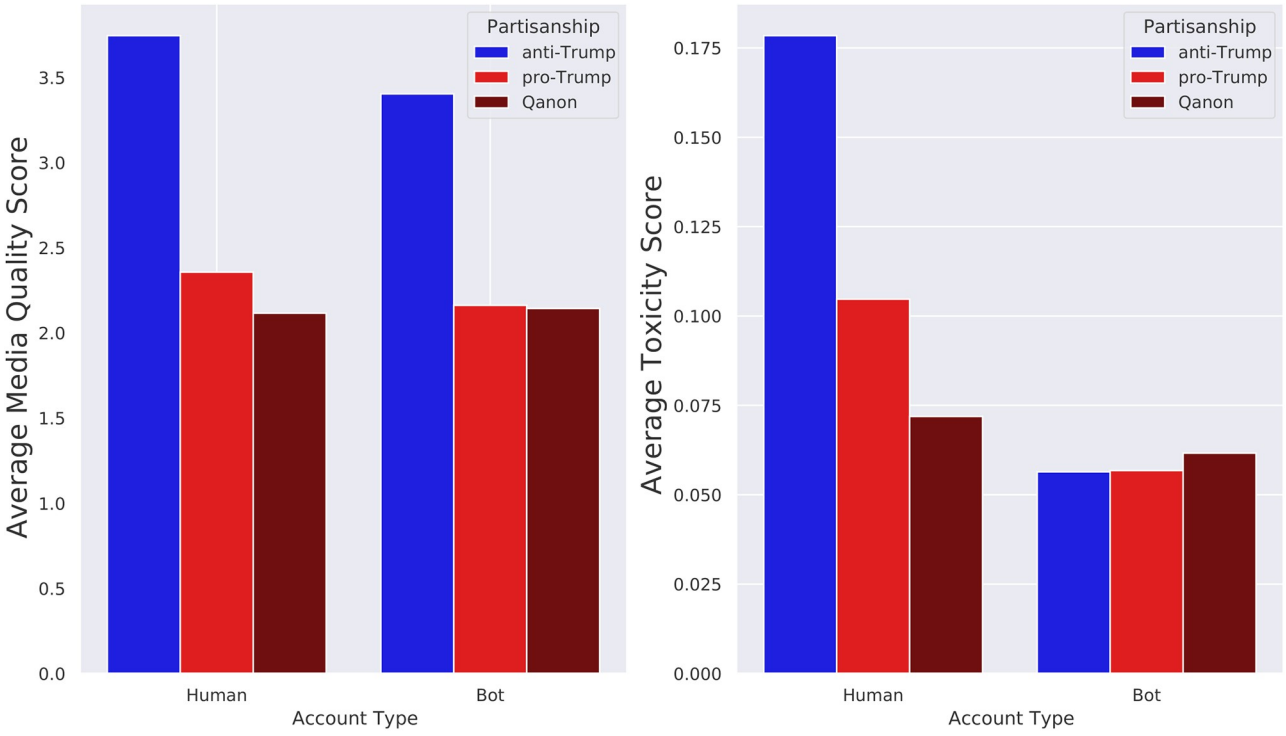

Fig 3. Bot language and content analysis. (left) Average media quality score and (right) average toxicity score of tweets posted by different account types.

https://doi.org/10.1371/journal.pone.0283971.g003

Trump accounts. This is consistent with past work which found that pro-Trump (Republican) users were much more likely to share news from untrustworthy news sites than anti-Trump (Democratic) users [49]. We also find that within each partisan group, the media quality score of bots is lower than that of humans (p-value $< 10^{-6}$ for bots versus humans in each partisan group). Qanon humans have a lower media quality score than both normal pro-Trump humans (p-value $< 10^{-6}$) and normal pro-Trump bots (p-value = 0.0002). However, we find no statistical difference between the average media quality score for Qanon humans and Qanon bots (p-value = 0.29).

In addition to what news media the accounts are sharing, we also investigate the tone and sentiment of the tweets they post. One measure of this is known as toxicity, which captures how harmful or unpleasant a tweet is. We measured the toxicity of all tweets in our dataset using Detoxify [50], a neural network model for toxicity detection that has achieved similar performance to top scorers in multiple Kaggle competitions related to detecting toxic language in web-based content [51]. Detoxify has been used in several studies for profiling hate speech spreaders [52], detecting cyberbullying in online forums [53], establishing toxicity thresholds for content [54], manipulating language models [55], video-text retrieval systems [56], and measuring the safety of conversational neural network models [57].

Despite its high performance, the model is sensitive to insults and profanity, which may result in higher toxicity scores regardless of the intent or tone of the message. The model produces toxicity scores ranging from zero to one, with higher values indicating higher toxicity. We applied Detoxify directly to the tweet text. If the tweet is a retweet, then Detoxify is applied to the text of the retweet. We do not open any links in a tweet or measure the toxicity of the destination website. We feel this is sufficient as a link does not have the same impact on an online conversation as toxic text in a tweet.

The average toxicity score for each account type is shown in Fig 3. We note that the distribution of toxicity scores for each group is characterized by two clusters located near zero and one. The cluster near one contains the toxic users which set the average toxicity of the group. Further analysis of the distribution of these scores is provided in the Materials and Methods. One striking observation in Fig 3 is that anti-Trump humans have the highest average toxicity score by a wide margin. The next highest toxicity group are pro-Trump humans. Qanon humans and bots have lower toxicity than normal humans (p-value $< 10^{-6}$). Bots of each partisanship group have lower toxicity levels than their co-partisan human counterparts (p-value $< 10^{-6}$). The high toxicity levels of anti-Trump humans may be due to their outrage about Trump's actions, and their disagreement with his acquittal. Conversely, low toxicity levels among pro-Trump users may reflect their attempts to post positive content in a show of support for the president.

When we focus on the bot accounts, we find two general patterns. Within each partisan group, bots share lower quality media than humans and post less toxic content than humans. From this we can deduce that bots mainly share low quality news stories (relative to their human counterparts within their partisan group), but they tend not to amplify negative messages which use aggressive language. This suggests that the bots are more focused on spreading information and not on agitating users with toxic posts.

Finally, we study bot retweet patterns. A retweet occurs when one account re-posts the content of another, thus sharing the original content with all of their followers. High retweet counts indicate high levels of popularity for the original tweet's content and author. Tables 2 and 3 show the content most retweeted by anti- and pro-Trump bot accounts, respectively, and the number of bots who retweeted each message. These top retweeted messages are primarily authored by elected officials involved in the impeachment proceedings.

Table 2. Statuses most retweeted by anti-Trump bots.

Retweeted status text	Bot count
RT @SpeakerPelosi: The House cannot choose our impeachment managers until we know what sort of trial the Senate will conduct. President Trump blocked his own witnesses and documents from the House, and from the American people, on phony complaints about the House process. What is his excuse now?	4,983
RT @RepAdamSchiff: Lt. Col. Vindman did his job. As a soldier in Iraq, he received a Purple Heart. Then he displayed another rare form of bravery—moral courage. He complied with a subpoena and told the truth. He upheld his oath when others would not. Right matters to him. And to us.	4,225
RT @RepAdamSchiff: Impeachment of a president is a serious undertaking. The Senate's role is to act as an impartial jury and provide a fair trial. Fair to the President and to the American people. That means seeing all the evidence, documents and witnesses. What is McConnell afraid of?	4,212
RT @SpeakerPelosi: In the Clinton impeachment process, 66 witnesses were allowed to testify including 3 in the Senate trial, and 90,000 pages of documents were turned over. Trump was too afraid to let any of his top aides testify & covered up every single document. The Senate must #EndTheCoverUp	3,977
RT @SpeakerPelosi: The President & Sen. McConnell have run out of excuses. They must allow key witnesses to testify, and produce the documents Trump has blocked, so Americans can see the facts for themselves. The Senate cannot be complicit in the President's cover-up. #DefendOurDemocracy	3,955

The messages retweeted by the greatest number of anti-Trump bots, and the number of anti-Trump bots who retweeted each.

https://doi.org/10.1371/journal.pone.0283971.t002

Fig 4 shows the accounts most retweeted by anti- and pro-Trump bot accounts, respectively, and the number of retweets each account received from bots. This allows us to examine which accounts benefited most from bot activity.

Here we see a clear difference in the structure of the bot retweet distributions. On the anti-Trump side, the bot retweets are distributed in a rather uniform manner over the top ten accounts, with retweet counts ranging from 240,336 to 92,233. However, bot retweet counts on the pro-Trump side range from 765,512 to 113,261, with a very concentrated distribution in which Donald Trump receives the most bot retweets by far, earning more than twice the amount of retweets than the second most retweeted account. This suggests Donald Trump is a singular figure in terms of bot retweets.

It is also interesting to note that the top ten bot retweeted accounts on the anti-Trump side are all political pundits. There are no elected or government officials. In contrast, among the

Table 3. Statuses most retweeted by pro-Trump bots.

Retweeted status text	Bot count
RT @realDonaldTrump: I was very surprised & disappointed that Senator Joe Manchin of West Virginia voted against me on the Democrat's totally partisan Impeachment Hoax. No President has done more for the great people of West Virginia than me (Pensions), and that will. . ..	8,265
RT @realDonaldTrump: "Nancy Pelosi said, it's not a question of proof, it's a question of allegations! Oh really?" @JudgeJeanine @FoxNews What a disgrace this Impeachment Scam is for our great Country!	8,134
RT @realDonaldTrump: As hard as I work, & as successful as our Country has become with our Economy, our Military & everything else, it is ashame that the Democrats make us spend so much time & money on this ridiculous Impeachment Lite Hoax. I should be able to devote all of my time to the REAL USA!	8,059
RT @realDonaldTrump: Crazy Nancy Pelosi should spend more time in her decaying city and less time on the Impeachment Hoax!	7,645
RT @realDonaldTrump: Many believe that by the Senate giving credence to a trial based on the no evidence, no crime, read the transcripts, "no pressure" Impeachment Hoax, rather than an outright dismissal, it gives the partisan Democrat Witch Hunt credibility that it otherwise does not have. I agree!	7,473

The messages retweeted by the greatest number of pro-Trump bots, and the number of pro-Trump bots who retweeted each.

https://doi.org/10.1371/journal.pone.0283971.t003

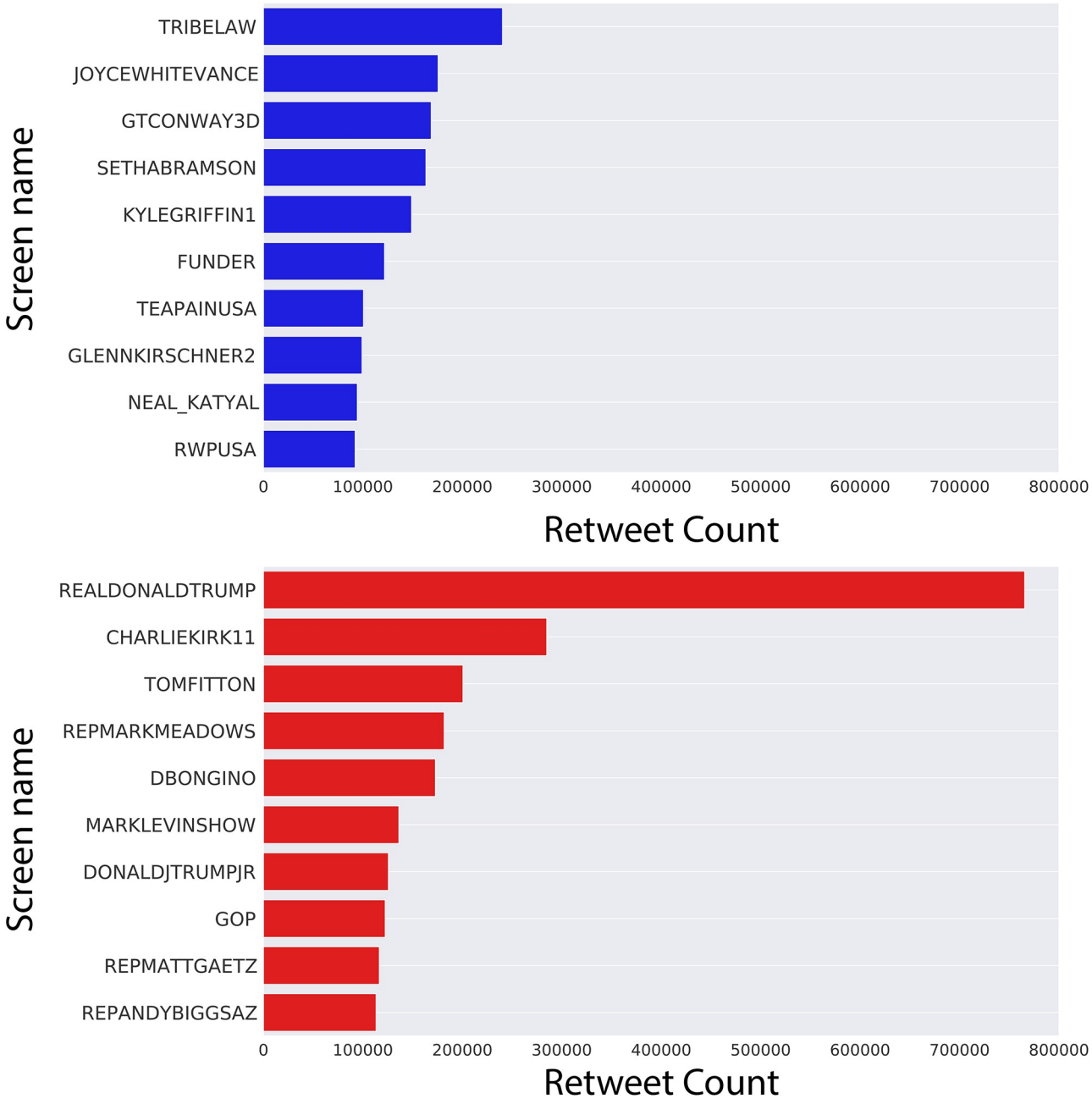

Fig 4. Bot retweet targets. Number of retweets by (top) anti-Trump bots and (bottom) pro-Trump bots for top retweeted Twitter accounts.

https://doi.org/10.1371/journal.pone.0283971.g004

top ten accounts on the pro-Trump side, there are three elected officials (President Donald Trump, Congressman Matt Gaetz, Congressman Andy Biggs) and one cabinet official (Chief of Staff Mark Meadows). Finally, we note that the official Twitter account of the republican party (GOP) is among the top ten accounts retweeted by bots, while the account of the democrat party (DNC) is not. This analysis suggests that pro-Trump bots are more actively amplifying officials with political or government power than anti-Trump bots.

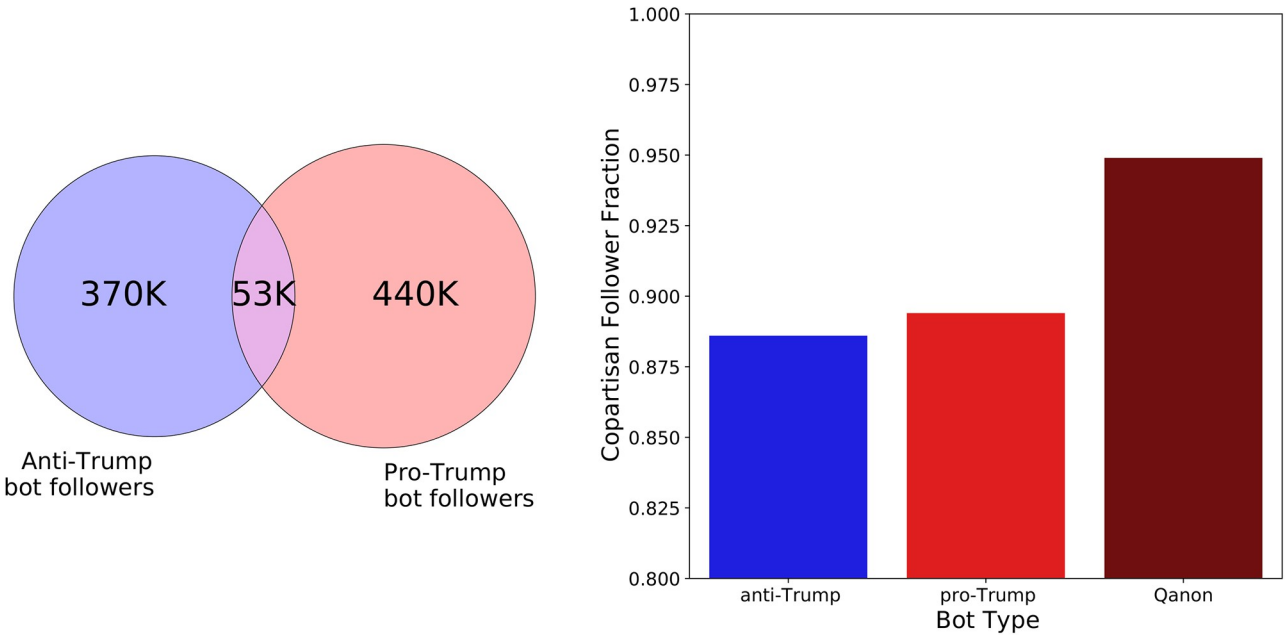

Fig 5. Bot follower network reach and composition. (left) Venn diagram of users who follow anti-Trump and pro-Trump bots. (right) Fraction of co-partisan followers for different bot types.

https://doi.org/10.1371/journal.pone.0283971.g005

Network analysis

Bots and Qanon supporters have networks that exhibit distinct properties. We first consider the bot follower network. Fig 5 shows the number of unique accounts who follow at least one bot, and the partisanship of the bots they follow. We find that pro-Trump bots have more followers than anti-Trump bots, which may be due to the pro-Trump bots' greater numbers. Less than 6% of the bot followers follow bots in both partisan groups, suggesting that the bot followers' network is highly polarized. This polarization is very visible in the bots' follower network, as shown in Fig 6. Here we see that the two partisan groups of bots are almost totally disconnected. To obtain a more quantitative measure of this polarization, we calculate the fraction of followers of each bot type who are co-partisan (i.e. those who share the same political affiliation). A higher value for this measure indicates greater ideological homogeneity in the followers of the bots. Fig 5 shows the co-partisan fractions for each bot type. We find the values are quite high, being above 0.87 for all bot types. There is a 1% difference in the co-partisan follower fraction between the anti-Trump and non-Qanon pro-Trump bots (p-value $< 10^{-6}$). However, the Qanon bots have have a co-partisan fraction that is 6% to 7% greater than the non-Qanon bots (p-value $< 10^{-6}$). This suggests that Qanon bots have an audience that is even more partisan than a standard bot.

The Qanon follower network, comprised of Qanon bots and humans, has an interesting structure. We find that this network contains a core of bots which are connected to each other, surrounded by a periphery of humans. Interestingly, these humans are only connected to the bots, and not to each other. We show this network in Fig 6, where the nodes are laid out with bots in the center to highlight this structural property. The connectivity among bots and lack of connectivity among humans suggests there is a hierarchical structure within the Qanon community. The bots act as sources of content for the humans. The humans appear not to

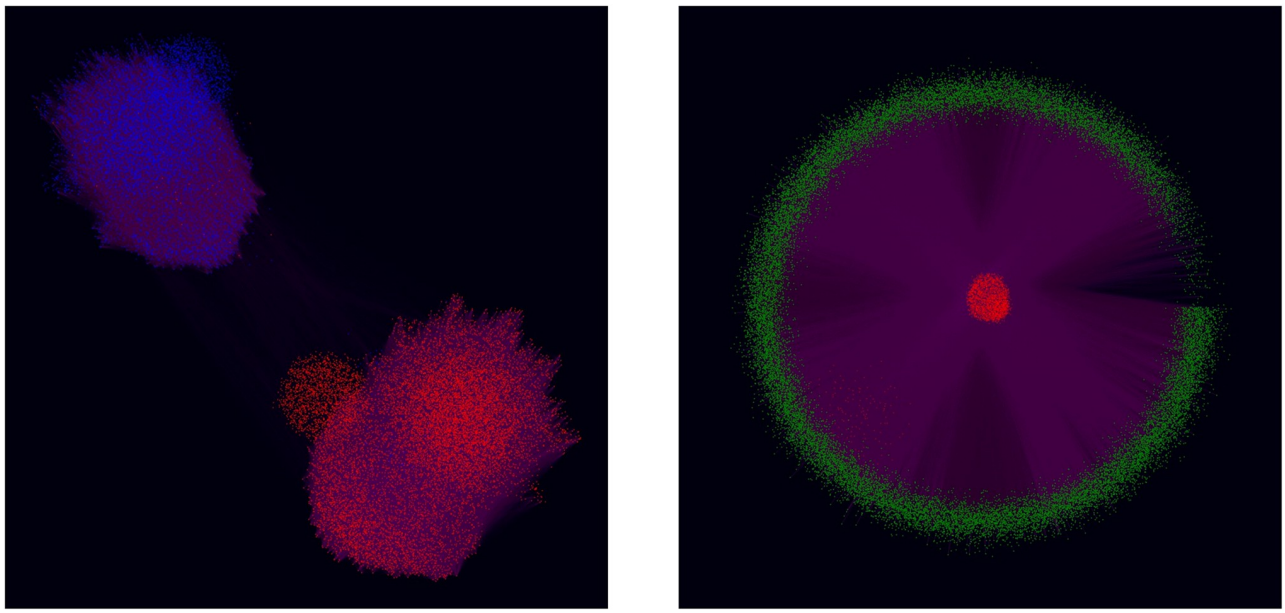

Fig 6. Bot follower network structure. (left) Follower network of bot accounts colored by partisanship, where anti-Trump bots are blue, and pro-Trump bots are red. (right) Follower network of Qanon accounts, where bots are colored red and humans are colored green.

https://doi.org/10.1371/journal.pone.0283971.g006

engage with each other, but rather mainly consume content from the bots. We will see later that this network structure has implications for the impact of Qanon bots.

Bot impact analysis

Recall that our focus is on bots that retweet excessively. We next provide a quantitative measure of the impact of these retweet bots on the impeachment discussion. Thus far we have presented a detailed analysis of the activity, sentiment, and network structure of the bots. Impact is a combination of all of these factors. We would expect more active bots to have more impact, as they post more content. Bots with larger network reach should also have greater impact. Finally, bots whose followers are not strong co-partisans would have more impact because they can persuade these followers to their side. In contrast, strong co-partisan followers who already agree with the bots likely cannot be persuaded further. A measure that combines these factors is known as harmonic influence centrality [58]. This is a network centrality based on a classic model for opinion dynamics [59] which incorporates stubborn users with immutable opinions [60]. Harmonic influence centrality measures how much a set of nodes shifts the average equilibrium opinion in the network. The centrality naturally incorporates activity, network reach, and opinions to measure the impact of nodes in a network because it is based on an opinion dynamics model which utilizes these factors. In its original form, harmonic influence centrality used only the network structure and activity level. A more recent version, known as generalized harmonic influence centrality (GHIC), incorporates additional data, such as the node sentiment, making it more appropriate for real social networks. GHIC has been used to quantify the impact of bots in Twitter networks discussing various geo-political events [40]. Because our data is quite similar in nature, we use this generalized version of the measure to quantify the impact of the bots.

To apply GHIC, one first must define the network. We want to calculate a daily impact measure for the bots to see how their impact evolves over time. Therefore, the networks we use

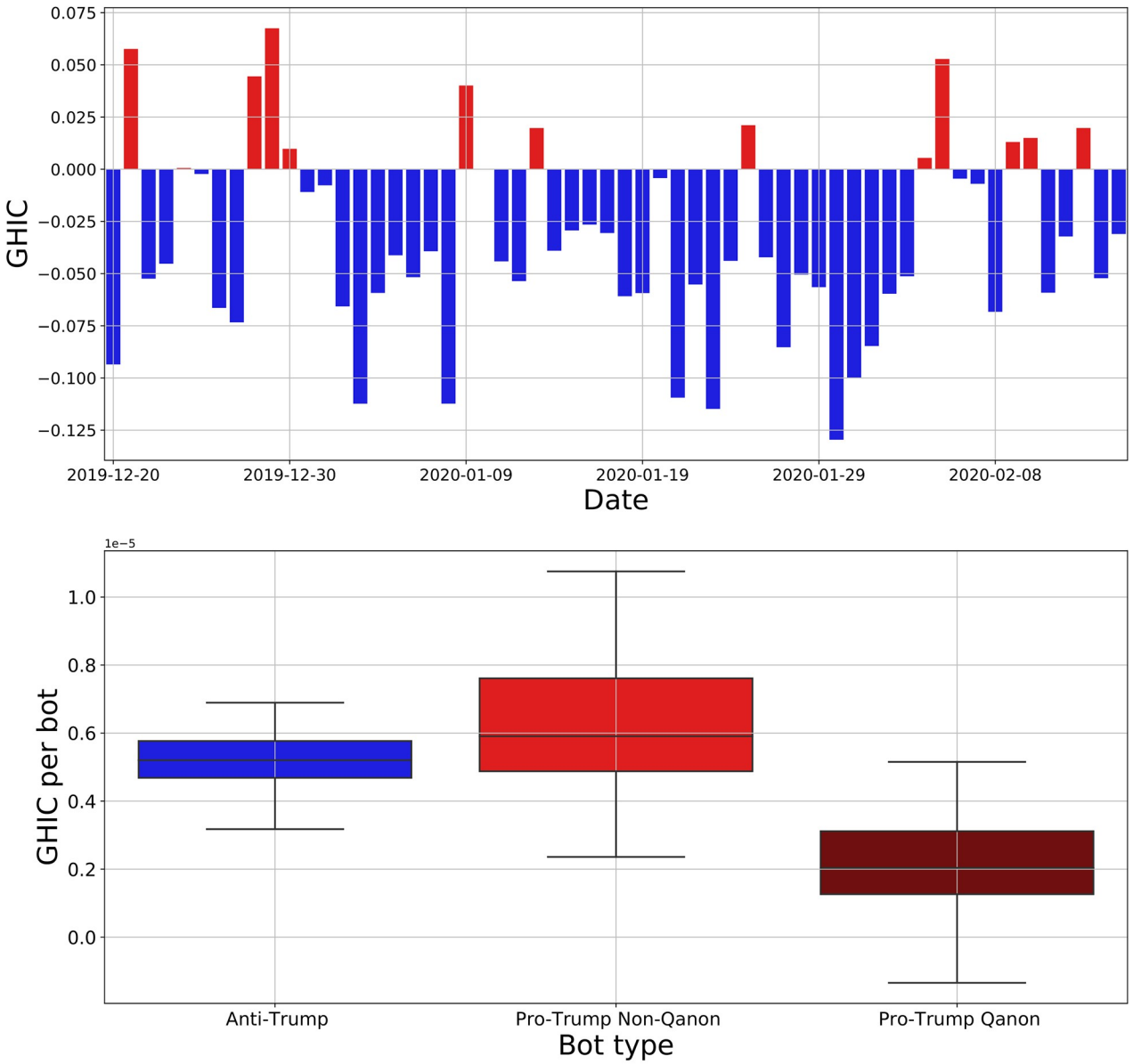

Fig 7. Bot impact analysis. (top) Daily generalized harmonic influence centrality (GHIC) score for all bots versus date. (bottom) Boxplot of the daily GHIC per bot for different bot types.

https://doi.org/10.1371/journal.pone.0283971.g007

are daily active follower networks. For a given day, the daily active follower network is the subnetwork of the entire Twitter follower network induced by accounts which are active (post at least one tweet) that day. We follow the approach in [40] to calculate the GHIC of different groups of bots on these networks (see Methods).

We first look at the GHIC of all bots. The daily GHIC of all bots is shown in Fig 7, where positive values indicate a shift toward pro-Trump opinions, and negative values indicate a shift toward anti-Trump opinions. This shows on a given day which partisan side of bots had greater impact. We observe that for most days, the anti-Trump bots have greater impact. However, there are days when the pro-Trump bots have greater impact. Upon closer investigation,

we find that on these days there was a news story which excited pro-Trump users on Twitter. For instance, on December 29, 2019 Donald Trump posted a controversial tweet referring to U.S. Speaker of the House of Representatives Nancy Pelosi as "Crazy Nancy". We see that the GHIC had a large positive value that day, indicating that the pro-Trump bots had greater impact. In general, we find that the GHIC achieves a large magnitude on days with polarizing events.

In addition to the daily GHIC, we are also interested in the efficiency of each group of bots. By efficiency, we mean the GHIC per bot for a group of bots. This measure would identify which group of bots are more impactful, on average. We consider three groups of bots: anti-Trump, pro-Trump non-Qanon, and pro-Trump Qanon. For each group, we calculate their daily GHIC per bot for all days. The boxplot in Fig 7 shows the distribution of the daily GHIC per bot for each group. We see that both groups of non-Qanon bots have very similar GHIC per bot distributions, though the pro-Trump bots have a slightly higher mean value. However, statistical tests do not provide strong evidence that the means of these two groups are different (p-value = 0.05). In contrast, we find strong evidence that Qanon bots have a lower mean GHIC per bot than non-Qanon bots (p-value $< 10^{-6}$). To understand why Qanon bots are less efficient, we look at Fig 6. We saw there that Qanon bots have a higher fraction of co-partisan followers than non-Qanon bots, meaning that Qanon bots are more ideologically aligned with their followers. Because of this, these bots cannot impact their followers' opinions as much, which lowers their GHIC efficiency. Our finding suggests that Qanon bots are not as effective at persuasion. Rather, they likely preach to a converted audience, which diminishes their impact compared to non-Qanon bots.

Discussion

Our study of the Twitter discussion surrounding the first impeachment of Donald Trump found that a small number of bots generated a disproportionately large amount of content. In addition, the combined follower reach of these bots is extensive. A primary bot activity is to spread news, and we found that bots generally spread lower quality news than humans. However, the language used in bot tweets is generally less toxic than that of humans. Using the GHIC measure we are able to quantify the daily impact of the bots. We found that bot impact is highest on days with politically charged events. Overall, anti-Trump bots have a greater impact, but their per bot impact is similar to the pro-Trump bots. Qanon bots have a lower per bot impact than the other bots. This is likely due to their high fraction of co-partisan followers, which limits their persuasive ability.

The excessive reach and activity level of bots, combined with their propensity to share news from low quality sources, are cause for concern. A small number of bots can amplify certain stories or narratives, causing them to reach a large audience. Bots seem to have the greatest impact on days when there is a large amount of partisan agitation, suggesting that bots may be increasing online polarization. However, one encouraging finding is that the Qanon bots, who spread a particularly dangerous form of disinformation, exist within strong echo-chambers, and as a result have less impact than normal bots.

Materials and methods

Data collection

The keywords and hashtags used as search criterion for tweets related to the first impeachment of President Donald Trump are shown in Table 4. Some of these words were added to the collection list as news stories developed. The table shows the date each term was added. The Twitter Streaming API was used to collect in real time any tweets containing at least one of these

Table 4. Tweet collection terms.

Keyword	Date added
#FactsMatter	2019-12-12
#IGHearing	2019-12-12
#IGReport	2019-12-12
#ImpeachAndConvict	2019-12-12
#ImpeachAndConvictTrump	2019-12-12
#SenateHearing	2019-12-12
#TrumpImpeachment	2019-12-12
impeach	2019-12-12
impeached	2019-12-12
impeachment	2019-12-12
Trump to Pelosi	2019-12-12
#25thAmendmentNow	2019-12-18
#ImpeachAndRemove	2019-12-18
#ImpeachmentEve	2019-12-18
#ImpeachmentRally	2019-12-18
#NotAboveTheLaw	2019-12-18
#trumpletter	2019-12-18
#GOPCoverup	2020-01-22
#ShamTrial	2020-01-22
#AquittedForever	2020-02-06
#CountryOverParty	2020-02-06
#CoverUpGOP	2020-02-06
#MitchMcCoverup	2020-02-06
#MoscowMitch	2020-02-06

Keywords used to collect impeachment related tweets, and the date each was added to the collection.

https://doi.org/10.1371/journal.pone.0283971.t004

terms. After an initial trial run from December 12 to 18, 2019, we then ran this collection process continuously from December 20, 2019 to March 24, 2020. Over this entire time period, we were able to collect 67.6 million tweets, posted by 3.6 million unique Twitter users.

As a part of this tweet collection process, we also collected the Twitter profile of each user. The profile included information such as the name of the user, location (if provided), and a short description provided by the user.

We later used the Twitter Search API to collect user follower networks in a separate collection process. Our network convention was to have follower edges point from a user to the person that followed them. This way the follower edges point in the direction of information flow, as tweets from a user appear in the timeline of their followers. To build the follower network for the users in our dataset, we used a customized web crawler to collect a list of *followings* for each user (i.e. the users they follow). We chose to collect the followings rather than the followers for each user because it reduced our data collection burden. We observed that the follower count can be much larger than the following count for a Twitter user, especially for the more popular users. Therefore, to more easily collect all edges in the follower network, we collected the users followings. To be able to collect the follower network in reasonable time, we collected a maximum of 2,000 followings per user. This value was sufficient for our data collection purposes as 85% of the following counts were below this value. In total we obtained 53.4 million edges in this follower network.

The data and code needed to reproduce our results can be found in the GitHub repositories located at https://github.com/s2t2/tweet-data-2020 and https://github.com/s2t2/tweet-analysis-2020, respectively. We have included in the data repository the tweet identifiers if one wishes to recollect the data from Twitter. All data was collected and is shared in accordance with Twitter's Terms of Service.

Partisanship classification model

The partisanship classifier we used is a bidirectional encoder representations from transformers (BERT) language representation model [44]. Transformers are a neural network architecture that have shown incredible success in language modeling [61]. BERT is a bidirectional transformer pre-trained using a combination of a masked language modeling objective and a next sentence prediction objective on the Toronto Book Corpus (800 million words) [62] and Wikipedia (2.5 million words). The BERT model provides a sentence embedding that can be used for many natural language processing tasks such as sentiment classification. The tweet text is fed into the BERT model and mapped into an embedding representation. We use the base BERT model which produces a 768 dimensional embedding. This representation is then fed to a fully connected single layer of 768 neurons with linear activation. The linear layer has two outputs which correspond to pro-Trump and anti-Trump sentiment. Our architecture follows the standard use of BERT for sentiment classification. Details on the architecture can be found in [44]. To obtain the sentiment of the tweet we use the value from the pro-Trump output so strong anti-Trump and pro-Trump sentiment are equal to zero and one, respectively.

We created training data for the model using strongly partisan users. These users were identified by the content of their Twitter profile descriptions. If a user's description contained any of the words in Table 5 and none of the keywords in Table 6, the user was given a label of

Table 5. Keywords used to identify anti-Trump users.

Hashtag	Description
#BIDEN2020	
#BLM	"Black Lives Matter"—a movement for racial equality
#BLUEWAVE	
#BLUEWAVE2020	
#DEMCAST	A left-leaning media outlet
#FBR	"Follow Black Resistance"
#IMPEACH	
#IMPEACHANDREMOVE	
#IMPEACHTRUMP	
#IMPEACHTRUMPNOW	
#IMPOTUS	"Impeached POTUS"
#METOO	A movement for gender equality
#NOTMYPRESIDENT	
#RESIST	
#RESISTANCE	
#RESISTER	
#THERESISTANCE	
#VOTEBLUE	
#VOTEBLUE2020	
#VOTEBLUENOMATTERWHO	
#WTP2020	"We The People 2020"

https://doi.org/10.1371/journal.pone.0283971.t005

Table 6. Keywords used to identify pro-Trump users.

Hashtag	Description
#1A	The First Amendment
#2A	The Second Amendment
#AMERICAFIRST	A Trump campaign slogan
#BUILDKATESWALL	
#BUILDTHEWALL	A Trump campaign slogan
#CODEOFVETS	
#CONSERVATIVE	
#DEPLORABLE	Refers to a Hillary Clinton quote from the 2016 election
#DRAINTHESWAMP	A Trump campaign slogan
#KAG	"Keep America Great"—a Trump campaign slogan
#MAGA	"Make America Great Again"—a Trump campaign slogan
#NRA	The National Rifle Association
#PATRIOT	
#POTUS45	45th President (Trump)
#QANON	Related to Qanon conspiracy theory
#THEGREATAWAKENING	Related to Qanon conspiracy theory
#TRUMP	
#TRUMP2020	
#TRUMPTRAIN	
#VETERAN	
#WALKAWAY	
#WWG1WGA	Related to Qanon conspiracy theory

https://doi.org/10.1371/journal.pone.0283971.t006

zero, indicating anti-Trump sentiment. The opposite was done to identify pro-Trump users, who were given a label of one. The labels of these strongly partisan users were assigned to their impeachment related tweets in our dataset. This process created a labeled training set of over 14 million tweets.

We used 637,672 of the labeled tweets to train the BERT sentiment classifier The tweets were chosen so that 50% were anti-Trump and 50% were pro-Trump, creating a balanced set of labels. We did not use all of the labeled tweets because the size of this dataset made the training process very slow. We found that using a smaller number of tweets resulted in a much faster training process while still producing a highly accurate classifier. To prevent over-fitting during training we use a dropout layer on the BERT output with a dropout probability of 0.3. The classifier was trained for ten epochs over the data using the Adam optimizer [63] and a cross-entropy loss function. The data was split into 80% for training, 10% for validation, and 10% for testing in a stratified manner.

The trained classifier is quite effective at measuring opinions about the impeachment. On the held out testing data it achieved a 96.3% accuracy score. We show the confusion matrix on the test data in Fig 8. As can be seen, the classifier achieves high true positive and negative rates while maintaining low false positive and negative rates. Also, the rates are nearly equal for both classes. We provide some random samples of tweets from each end of the sentiment spectrum and their partisan sentiment measured by the classifier in Table 7. We plot in Fig 9 a histogram of the partisanship scores of all users in our dataset as measured by the classifier. The user scores are obtained by averaging the partisanship score of each user's tweets.

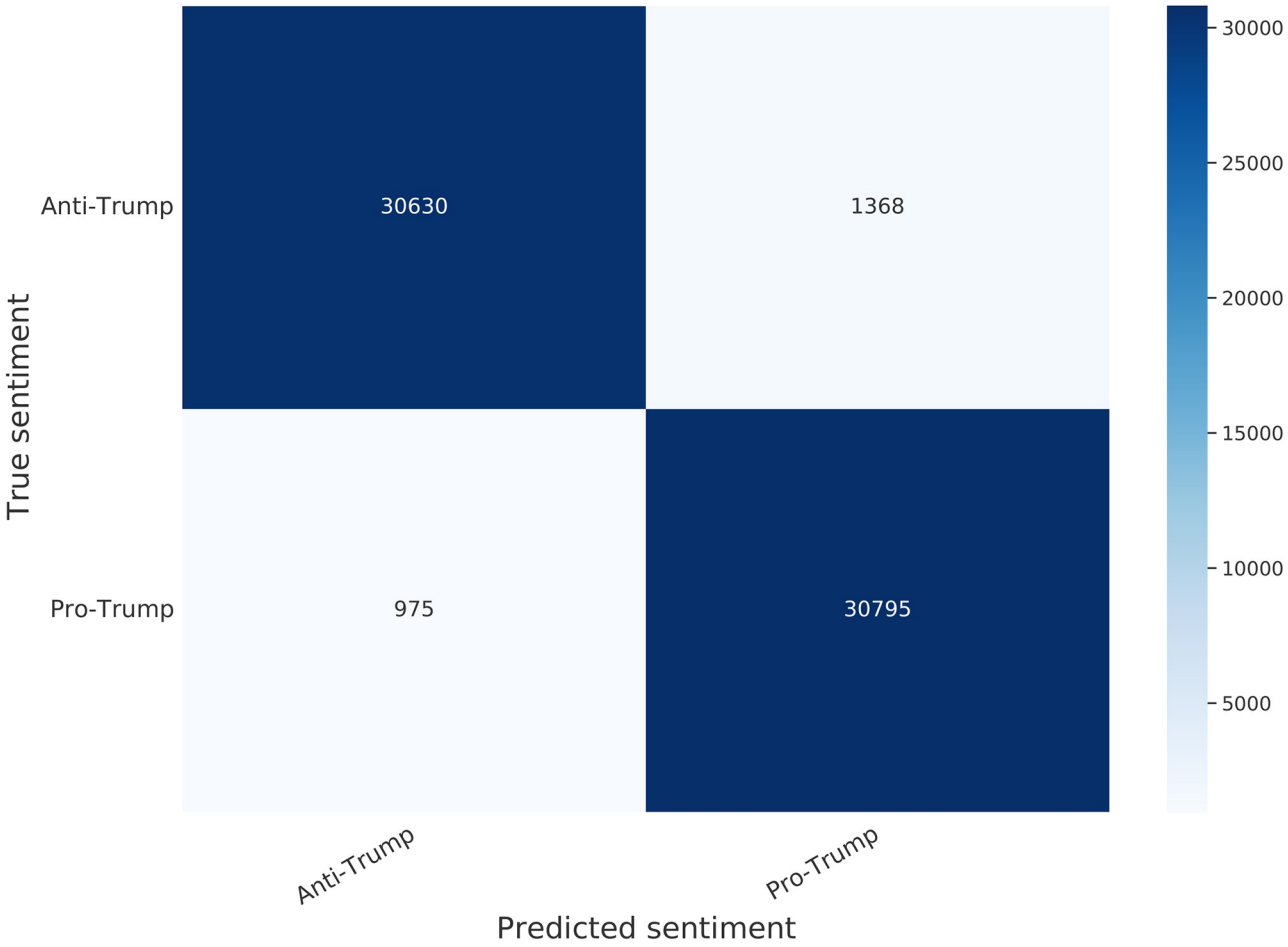

Fig 8. Confusion matrix of trained partisanship classifier applied to test data.

https://doi.org/10.1371/journal.pone.0283971.g008

Qanon classification

To identify Qanon supporters, we utilized a list of terms commonly used by Qanon supporters at the time, which are shown in Table 8. The terms on this list are consistent with those discussed by other work in this field [23, 64]. Other researchers have classified Qanon supporters based on hashtags used in their tweets or retweets [23], however our method is based on

Table 7. Example partisanship classification scores.

Tweet text	Opinion
Mark my words. . . Trump is starting a war to distract from the Impeachment.	0.105
LA Times joins growing list of papers calling for Trump's impeachment	0.17
Flynn sentencing—Jan 28 State of the Union -Feb 4 Stone sentencing—Feb 6 Impeachment—Forever	0.26
Americans have officially lost any belief in the Democrats' partisan impeachment sham.	0.96
It's time for the Senate to end this partisan, impeachment sham once and for all.	0.97
Pelosi and Dems blew it! Impeachment Sham is Backfiring!	0.98

Tweets from testing dataset and their opinion scores assigned by the BERT sentiment classifier. The tweets are ordered by opinion score. An opinion of zero is anti-Trump and an opinion of one is pro-Trump.

https://doi.org/10.1371/journal.pone.0283971.t007

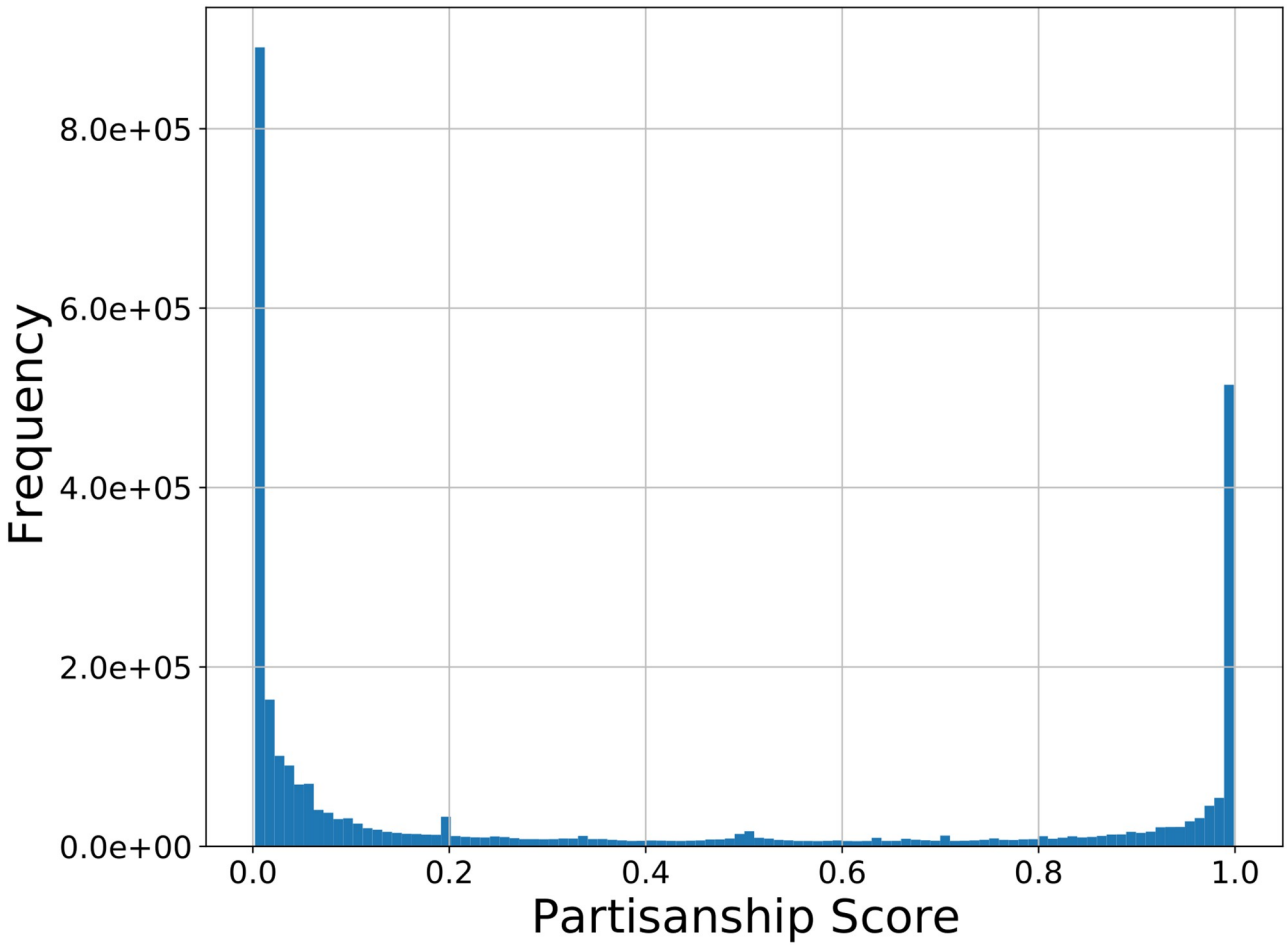

Fig 9. Histogram of the partisanship scores of all users.

https://doi.org/10.1371/journal.pone.0283971.g009

hashtags used in profile descriptions. Any user account that included at least one of these terms in their profile description was labeled as Qanon (after excluding any anti-Trump accounts or accounts which tweeted using anti-Trump hashtags). A limitation of this approach is that our list of hashtags is likely not comprehensive of all Qanon hashtags, especially as they may evolve over time.

Table 8. Q-anon user profile terms.

Keyword
#QANON
#WWG1WGA
#GREATAWAKENING
#WAKEUPAMERICA
#WEARETHENEWSNOW

Keywords used to identify Qanon supporters.

https://doi.org/10.1371/journal.pone.0283971.t008

Bot detection

Bots exhibit certain traits and behaviors that allow them to be identified. Some of these include excessive retweeting and never posting original tweets. These behaviors are likely due to the bots' automated nature. Many algorithms have been developed for bot detection in Twitter and each has its own strengths. The algorithm of [40] uses a factor graph model that allows for the simultaneous detection of multiple bots based on the collective retweeting behavior of users discussing a specific topic. The data required by this algorithm is the set of tweets, or more precisely, the set of retweets, about the topic. Another popular algorithm is the machine learning based Botometer (formerly BotOrNot) [33] which utilizes a large amount of data about an individual account, including followers, friends, tweets, and profile, in order to determine if it is a bot. Botometer is a good algorithm to use when one wants to determine if an individual account is a bot. However, when one wants to find bots among a large set of users discussing a topic, as is the case in our impeachment analysis, the factor graph algorithm is more convenient. We can identify bots with this algorithm without having to collect any additional data, which is important given how large our daily active user sets are. In addition, it has been found that the factor graph algorithm has slightly better performance than Botometer [40].

The factor graph algorithm first constructs a retweet network based on the retweets in a given set of tweets. In this retweet network the nodes are the users who tweet and an edge (u, v) pointing from user u to user v means that v retweets u. The edge (u, v) is given a weight w_{uv} which equals the number of times v retweets u. Like with the follower network, the edge direction in the retweet network indicates the flow of information. The algorithm uses the retweet network structure to calculate a bot probability for each user. This probability is based on the empirical observations that bots are likely to retweet humans, but unlikely to retweet bots, and humans are likely to retweet humans and less likely to retweet bots [40]. This homophily for the humans and heterophily for the bots is utilized by the algorithm to determine bot probabilities from the structure of the retweet network.

We used the factor graph algorithm to identify bots active each day. To apply the algorithm, we first constructed a retweet network from the tweets posted on a given day. Once the retweet network is constructed, we apply the algorithm to the network using the parameters specified in [40] to simultaneously obtain the joint bot probability for all users. We show one example of the bot probability distribution for a single day in Fig 10. As can be seen, the bulk of users have bot probabilities near 0.5, which is the algorithm saying it cannot determine one way or another what the user is. In the upper probability range we see that the histogram decreases up to approximately 0.8, and then increases afterwards. This suggests there is a cluster of nodes with bot probabilities in the interval [0.8, 1.0]. We used the lower probability bound of this cluster as the bot probability threshold so that this cluster of nodes is identified as bots.

We found a similar behavior in the bot probability distribution across all days. Therefore, we used the same bot probability threshold of 0.8 for each day. Any user who had a bot probability over the threshold in at least one day in our dataset was declared to be a bot. We use this approach because due to automation, it is easy for a bot to act like a human (retweet less frequently). However, it is harder for a human to behave like a bot (retweet at an extremely high rate). Also, bots may intentionally behave in this non-constant manner. A bot may be extremely active on certain days when it is amplifying certain users, and quiet or only slightly active on others. Humans showing elevated retweet rates may have difficulty being as active as a bot. Under these assumptions, exhibiting bot behavior on at least one day would indicate the account is a bot, and humans would not show bot-like behavior any day. Across all days we found 24,150 bots among 3.6 million active accounts. As a terminology note, in this paper

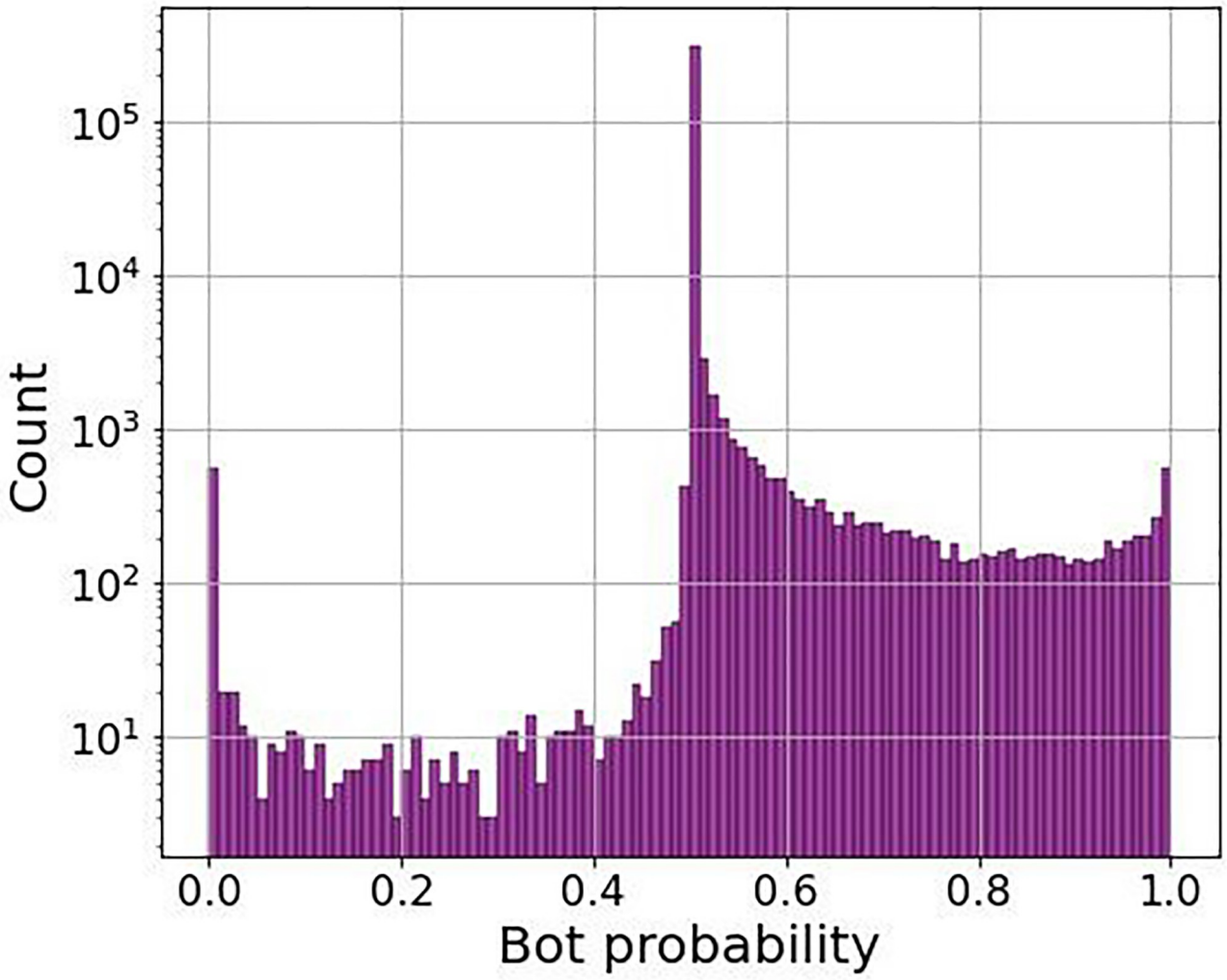

Fig 10. Bot probability classifications on an example day. Histogram of the bot probabilities calculated by the factor graph algorithm [40] based on the impeachment retweet network on February 1, 2020 (an example day).

https://doi.org/10.1371/journal.pone.0283971.g010

when we refer to the word "bot" we mean the accounts likely to be retweet bots, and when we refer to the word "human" we mean accounts not likely to be retweet bots (i.e. non-bots).

To explore the validity of our bot classifications, we took a random sample of around 7500 accounts with a roughly 60/40 human to bot ratio, and compared them against bot scores obtained from the Botometer API [3]. The Botometer API provides scores from 0 to 1 representing the likelihood that a given account is a bot (i.e. "overall" and "cap" scores), as well as separate scores for a number of bot sub-types (e.g. "astroturf", "financial", etc.). We construct receiver operating characteristic (ROC) curves for different Botometer scores and our labels as a ground-truth. We use area under the curve (AUC) score to measure the agreement between our bot classifications and each kind of score provided by Botometer. We find that the AUC score for the "overall" bot scores is around 0.79, while the AUC score for the "astroturf" scores is much higher, around 0.93 (see Fig 11). This suggests that our bot detection algorithm, when used within the context of a political discussion, performs generally well at identifying bots overall, and performs even better at identifying hyper active political bots.

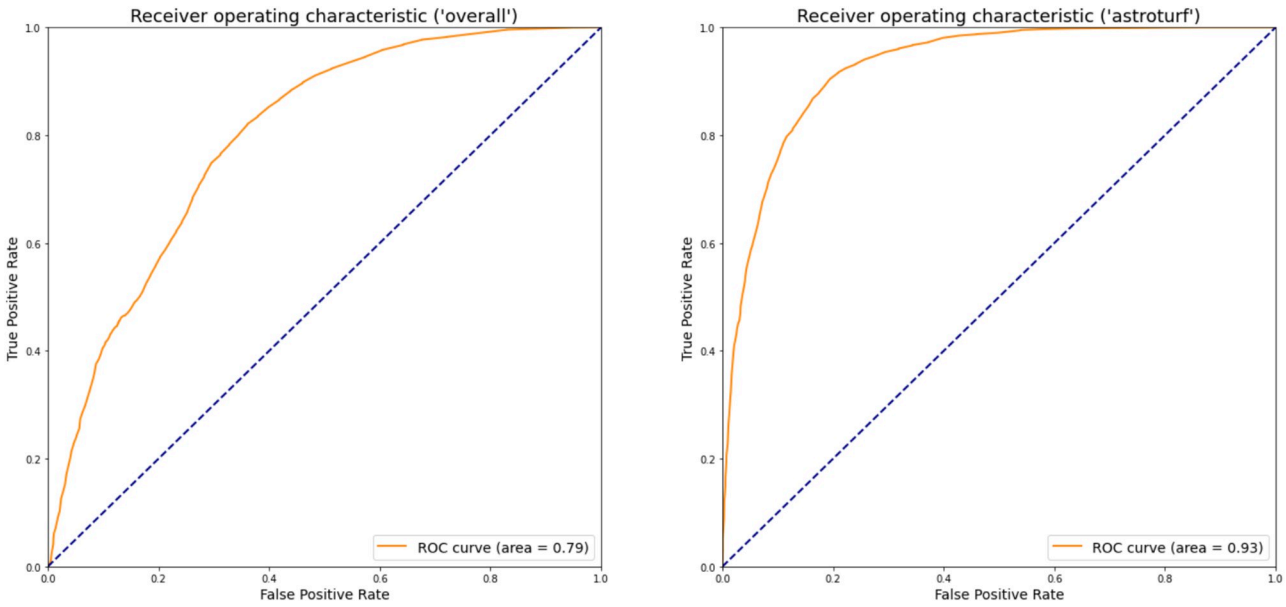

Fig 11. ROC curves for (left) "overall" Botometer scores and (right) "astroturf" Botometer scores using our bot classification results as ground truth.

https://doi.org/10.1371/journal.pone.0283971.g011

Distribution of toxicity scores

Fig 12 shows a violin plot of the mean user toxicity score for the different user types. From this figure we see that there are clusters of humans with very high toxicity scores. The bots do not have such a toxic cluster. To further quantify these clusters, let us define a toxic user as one whose average toxicity score is greater than or equal to 0.9. With this definition, we find that 2% of Republican humans are toxic, while 6% of Democrat humans are toxic. There are zero toxic bots under this definition.

Generalized harmonic influence centrality calculation

To calculate the generalized harmonic influence centrality (GHIC) of a set of nodes we need several pieces of information. First, we need the follower network of the nodes. We use the daily active follower network for this purpose, which is the follower network induced by users who tweet about the impeachment on a given day.

Second, we need to characterize the activity level of the users. We define this to be the tweet rate of the users. We measure the tweet rate as the total number of impeachment tweets posted divided by the duration of our data collection. This gives a stable measure of the posting rate of a user. We use this same posting rate for each daily GHIC calculation as it provides an accurate measure of the general activity level of a user.

Third, we need to know the political sentiment of each user. We obtain this by taking the average sentiment of a user's tweets as measured by our BERT partisanship classifier.

Fourth, we need to identify a subset of users in the daily active follower network as stubborn. GHIC assumes that the stubborn users are not persuadable and their sentiment does not change. We first set all bots to be stubborn, as they are automated accounts that do not respond to persuasion. We identify stubborn users among the humans based on their political sentiment or opinion. Studies have shown that stubborn users have extreme opinions [65]. To operationalize this principle, we follow the approach used in [40, 43] to define extreme

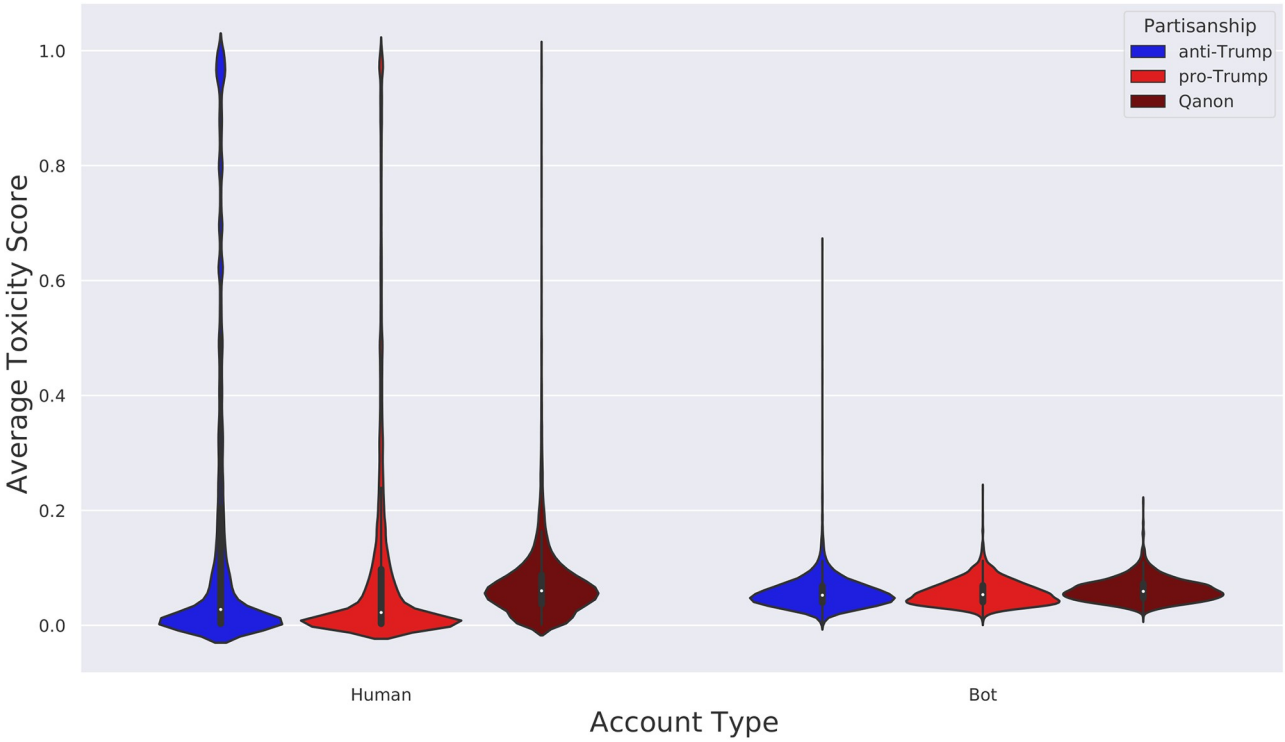

Fig 12. Violin plot of the mean user toxicity score for different user types.

https://doi.org/10.1371/journal.pone.0283971.g012

opinions as those below the 10th percentile and above the 90th percentile of the opinions of all users in our dataset. It has been shown that the GHIC is robust to the precise value of these thresholds [40].

Once the requisite information has been obtained, the GHIC of a set of nodes can be calculated. We present the steps for GHIC calculation here, drawing from the presentation in [40]. We are given a follower network $G = (V, E)$ with node set V (the Twitter users) and edge set E (the follower edges). We assume each node follows a set of users that we define as their *following*. Each node $v \in V$ posts content (tweets) at a rate λ_v. We define the stubborn users as the set $V_0 \subset V$ and the non-stubborn users as $V_1 \subset V$. We define $\mathbf{\Psi}$ as the vector of stubborn opinions and θ as the vector of non-stubborn opinions. We are given the stubborn opinions Ψ and we obtain the non-stubborn opinions θ by solving

$$\mathbf{G}\theta = \mathbf{F}\mathbf{\Psi}, \tag{1}$$

where the matrix $\mathbf{G}$ is given by the equation:

$$\mathbf{G}_{ij} = \begin{cases} -\sum_{k \in \text{following of } i} \lambda_k & i = j, i \in V_1 \\ \lambda_j & i \neq j, (j,i) \in E, i, j \in V_1 \\ 0 & \text{else,} \end{cases}$$

and the matrix **F** is given by

$$\mathbf{F}_{ij} = \begin{cases} -\lambda_j & (j,i) \in E, i \in V_1, j \in V_0 \\ 0 & \text{else.} \end{cases}$$

Eq (1) is the equilibrium opinions in the opinion dynamics model upon which GHIC is based.

Next we choose a set of nodes $S \subset V$ for which we want to calculate the GHIC. We define a new network $G' = (V', E')$ where $V' = V/S$ and G' is the sub-network of G induced by V'. G' is the network G but with the nodes in S removed. Let θ and θ' be the solution of equation (1) assuming the underlying network is G and G', respectively. The GHIC of S is then given by

$$\text{GHIC}(S) \quad = \frac{1}{|V_1/S|} \sum_{i \in V_1/S} \theta_i - \theta'_i.$$

We see from this that the GHIC of S is the change in mean non-stubborn equilibrium opinion caused by the presence of the S nodes in the network.

Robustness of generalized harmonic influence centrality

We perform checks to demonstrate the robustness of the generalized harmonic influence centrality (GHIC) with respect to the bot probability threshold. Bots are identified as accounts whose bot probability, as determined by our algorithm, exceeds a threshold of 0.8. We test different values for this threshold and see how the GHIC is affected. We chose to test the values 0.72 and 0.88, which are 10% above and below the baseline threshold. The results of this analysis are shown in Fig 13. Changes in the bot probability threshold do not cause substantial changes in the GHIC, as can be seen from the narrow error bars. We find that the median absolute shift of the GHIC with respect to the bot threshold is 6%, suggesting that the GHIC is robust to the bot probability threshold.

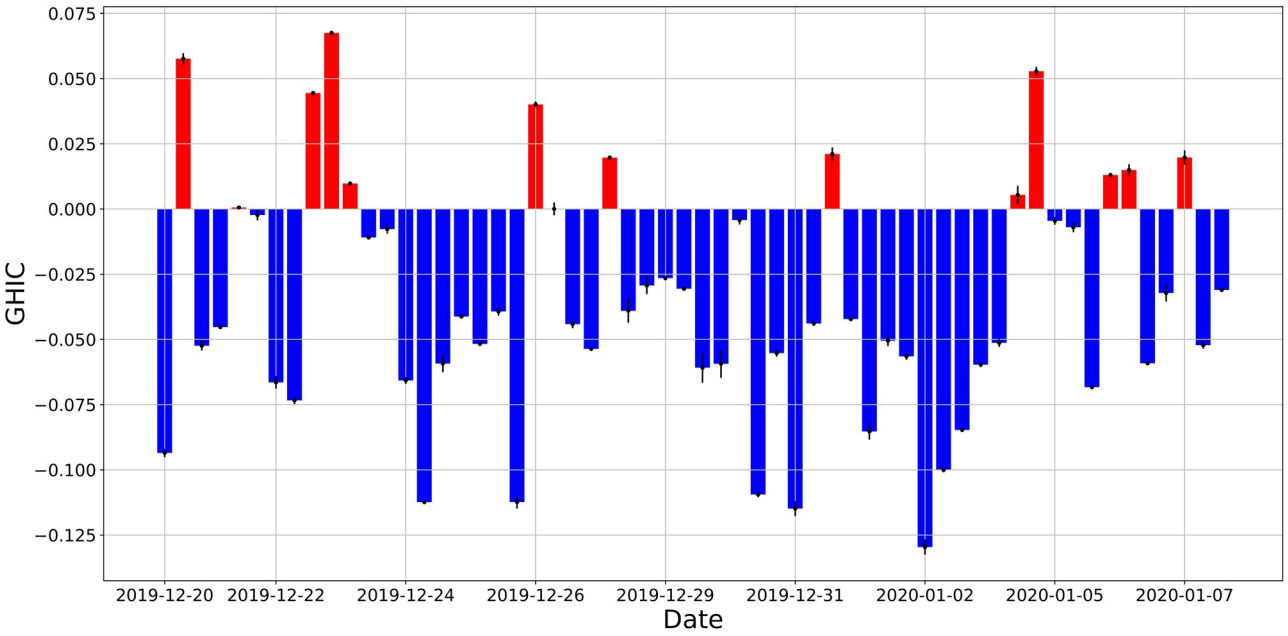

Fig 13. Daily generalized harmonic influence centrality (GHIC) score for all bots versus date. The error bars correspond to the range of GHIC values for bot probability thresholds of 0.72, 0.80, and 0.88.

https://doi.org/10.1371/journal.pone.0283971.g013

Author Contributions

Conceptualization: Tauhid Zaman.

Data curation: Michael Rossetti, Tauhid Zaman.

Formal analysis: Michael Rossetti, Tauhid Zaman.

Investigation: Michael Rossetti, Tauhid Zaman.

Methodology: Michael Rossetti, Tauhid Zaman.

Project administration: Michael Rossetti, Tauhid Zaman.

Resources: Tauhid Zaman.

Software: Michael Rossetti, Tauhid Zaman.

Supervision: Tauhid Zaman.

Validation: Michael Rossetti, Tauhid Zaman.

Visualization: Michael Rossetti, Tauhid Zaman.

Writing – original draft: Michael Rossetti, Tauhid Zaman.

Writing – review & editing: Michael Rossetti, Tauhid Zaman.

References

1. Roose K. Brace Yourself for the Internet Impeachment. The New York Times. 2019;.
2. Ferrara E, Varol O, Davis C, Menczer F, Flammini A. The rise of social bots. Communications of the ACM. 2016; 59(7):96–104. https://doi.org/10.1145/2818717
3. Sayyadiharikandeh M, Varol O, Yang KC, Flammini A, Menczer F. Detection of novel social bots by ensembles of specialized classifiers. In: Proceedings of the 29th ACM international conference on information & knowledge management; 2020. p. 2725–2732.
4. Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, Menczer F. The spread of low-credibility content by social bots. Nature communications. 2018; 9(1):1–9. https://doi.org/10.1038/s41467-018-06930-7 PMID: 30459415
5. Varol O, Ferrara E, Davis C, Menczer F, Flammini A. Online human-bot interactions: Detection, estimation, and characterization; 2017.
6. Woolley SC, Howard PN. Computational propaganda: Political parties, politicians, and political manipulation on social media. Oxford University Press; 2018.
7. Parlapiano A, Lee JC. The propaganda tools used by Russians to influence the 2016 election. The New York Times. 2018;.
8. Shane S. The fake Americans Russia created to influence the election. The New York Times. 2017;.
9. Guilbeault D, Woolley S. How Twitter bots are shaping the election. The Atlantic. 2016;.
10. Byrnes N. How the bot-y politic influenced this election. Technology Rev. 2016;.
11. Bessi A, Ferrara E. Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday. 2017; 22(8-7).
12. Bessi A, Ferrara E. Social bots distort the 2016 US Presidential election online discussion. First Monday. 2016; 21(11-7).
13. Ferrara E, Chang H, Chen E, Muric G, Patel J. Characterizing social media manipulation in the 2020 US presidential election. First Monday. 2020; 25(11-2).
14. Walter D, Ophir Y, Jamieson KH. Russian Twitter accounts and the partisan polarization of vaccine discourse, 2015–2017. American Journal of Public Health. 2020; 110(5):718–724. https://doi.org/10.2105/AJPH.2019.305564 PMID: 32191516
15. Broniatowski DA, Jamison AM, Qi S, AlKulaib L, Chen T, Benton A, et al. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American journal of public health. 2018; 108(10):1378–1384. https://doi.org/10.2105/AJPH.2018.304567 PMID: 30138075

16. Ferrara E. What types of COVID-19 conspiracies are populated by Twitter bots? First Monday. 2020; 25 (6-1).
17. Zuckerman E. QAnon and the Emergence of the Unreal. Journal of Design and Science. 2019; 6:1–15.
18. Miller ME. Pizzagate's Violent Legacy. Washington Post. 2021;.
19. Robb A. Anatomy of a Fake News Scandal. Rolling Stone. 2017;.
20. Kang C, Goldman A. In Washington pizzeria attack, fake news brought real guns. New York Times. 2016;.
21. Roose K. What is QAnon, the viral pro-Trump conspiracy theory. The New York Times. 2021;.
22. Tollefson J. How Trump turned conspiracy theory research upside down. Nature. 2021; 590:192–193. PMID: 33542489
23. Xu W, Sasahara K, et al. A network-based approach to QAnon user dynamics and topic diversity during the COVID-19 infodemic. APSIPA Transactions on Signal and Information Processing. 2022; 11(2). https://doi.org/10.1561/116.00000055
24. Ferraz Costa A, Yamaguchi Y, Juci Machado Traina A, Traina Jr C, Faloutsos C. Rsc: Mining and modeling temporal activity in social media. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining; 2015. p. 269–278.
25. Zhang CM, Paxson V. Detecting and analyzing automated activity on twitter. In: Passive and Active Measurement: 12th International Conference, PAM 2011, Atlanta, GA, USA, March 20-22, 2011. Proceedings 12. Springer; 2011. p. 102–111.
26. Kudugunta S, Ferrara E. Deep neural networks for bot detection. Information Sciences. 2018; 467:312–322. https://doi.org/10.1016/j.ins.2018.08.019
27. Igawa RA, Barbon S Jr, Paulo KCS, Kido GS, Guido RC, Júnior MLP, et al. Account classification in online social networks with LBCA and wavelets. Information Sciences. 2016; 332:72–83. https://doi.org/10.1016/j.ins.2015.10.039
28. Clark EM, Williams JR, Jones CA, Galbraith RA, Danforth CM, Dodds PS. Sifting robotic from organic text: a natural language approach for detecting automation on Twitter. Journal of computational science. 2016; 16:1–7. https://doi.org/10.1016/j.jocs.2015.11.002
29. Morstatter F, Wu L, Nazer TH, Carley KM, Liu H. A new approach to bot detection: striking the balance between precision and recall. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE; 2016. p. 533–540.
30. Gilani Z, Kochmar E, Crowcroft J. Classification of twitter accounts into automated agents and human users. In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017; 2017. p. 489–496.
31. Yang KC, Varol O, Hui PM, Menczer F. Scalable and generalizable social bot detection through data selection; 2020.
32. Dickerson JP, Kagan V, Subrahmanian V. Using sentiment to detect bots on twitter: Are humans more opinionated than bots? In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). IEEE; 2014. p. 620–627.
33. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F. Botornot: A system to evaluate social bots. In: Proceedings of the 25th international conference companion on world wide web; 2016. p. 273–274.
34. Luceri L, Deb A, Badawy A, Ferrara E. Red bots do it better: Comparative analysis of social bot partisan behavior. In: Companion proceedings of the 2019 world wide web conference; 2019. p. 1007–1012.
35. Pozzana I, Ferrara E. Measuring bot and human behavioral dynamics. Frontiers in Physics. 2020; p. 125. https://doi.org/10.3389/fphy.2020.00125
36. Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: Proceedings of the 26th international conference on world wide web companion; 2017. p. 963–972.
37. Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In: Proceedings of the 10th ACM conference on web science; 2019. p. 183–192.
38. Cresci S, Di Pietro R, Petrocchi M, Spognardi A, Tesconi M. DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems. 2016; 31(5):58–64. https://doi.org/10.1109/MIS.2016.29
39. Vo N, Lee K, Cao C, Tran T, Choi H. Revealing and detecting malicious retweeter groups. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017; 2017. p. 363–368.
40. des Mesnards NG, Hunter DS, el Hjouji Z, Zaman T. Detecting bots and assessing their impact in social networks. Operations Research. 2022; 70(1):1–22. https://doi.org/10.1287/opre.2021.2118

41. Beskow DM, Carley KM. Its all in a name: detecting and labeling bots by their name. Computational and mathematical organization theory. 2019; 25:24–35. https://doi.org/10.1007/s10588-018-09290-1

42. Shao C, Hui PM, Wang L, Jiang X, Flammini A, Menczer F, et al. Anatomy of an online misinformation network. Plos one. 2018; 13(4):e0196087. https://doi.org/10.1371/journal.pone.0196087 PMID: 29702657

43. Hunter DS, Zaman T. Optimizing opinions with stubborn agents under time-varying dynamics. arXiv preprint arXiv:180611253. 2018;.

44. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;.

45. Pennycook G, Rand DG. Fighting misinformation on social media using crowdsourced judgments of news source quality. Proceedings of the National Academy of Sciences. 2019; 116(7):2521–2526. https://doi.org/10.1073/pnas.1806781116 PMID: 30692252

46. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG. Shifting attention to accuracy can reduce misinformation online. Nature. 2021; 592(7855):590–595. https://doi.org/10.1038/s41586-021-03344-2 PMID: 33731933

47. Mosleh M, Martel C, Eckles D, Rand D. Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a Twitter field experiment. In: proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–13.

48. Guess A, Nagler J, Tucker J. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. Science advances. 2019; 5(1):eaau4586. https://doi.org/10.1126/sciadv.aau4586 PMID: 30662946

49. Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D. Fake news on Twitter during the 2016 US presidential election. Science. 2019; 363(6425):374–378. https://doi.org/10.1126/science.aau2706 PMID: 30679368

50. Hanu L, Thewlis J, Haco S. How AI is learning to identify toxic online content. Scientific American. 2021;.

51. Hanu L, contributors. Detoxify; 2020. https://github.com/unitaryai/detoxify.

52. Huertas-García Á, Huertas-Tato J, Martín A, Camacho D. Profiling Hate Speech Spreaders on Twitter: Transformers and mixed pooling. CLEF (Working Notes). 2021;2021.

53. Vo HHP, Tran HT, Luu ST. Automatically Detecting Cyberbullying Comments on Online Game Forums. In: 2021 RIVF International Conference on Computing and Communication Technologies (RIVF). IEEE; 2021. p. 1–5.

54. Iqbal W, Arshad MH, Tyson G, Castro I. Exploring Crowdsourced Content Moderation Through Lens of Reddit during COVID-19. In: Proceedings of the 17th Asian Internet Engineering Conference; 2022. p. 26–35.

55. Bagdasaryan E, Shmatikov V. Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures. In: 2022 IEEE Symposium on Security and Privacy (SP). IEEE; 2022. p. 769–786.

56. Hanu L, Thewlis J, Asano YM, Rupprecht C. VTC: Improving Video-Text Retrieval with User Comments. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. Springer; 2022. p. 616–633.

57. Sun H, Xu G, Deng J, Cheng J, Zheng C, Zhou H, et al. On the safety of conversational models: Taxonomy, dataset, and benchmark. arXiv preprint arXiv:211008466. 2021;.

58. Vassio L, Fagnani F, Frasca P, Ozdaglar A. Message passing optimization of harmonic influence centrality. IEEE transactions on control of network systems. 2014; 1(1):109–120. https://doi.org/10.1109/TCNS.2014.2304870

59. DeGroot MH. Reaching a consensus. Journal of the American Statistical association. 1974; 69 (345):118–121. https://doi.org/10.1080/01621459.1974.10480137

60. Mobilia M. Does a single zealot affect an infinite group of voters? Physical review letters. 2003; 91 (2):028701. https://doi.org/10.1103/PhysRevLett.91.028701 PMID: 12906515

61. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017; 30.

62. Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 19–27.

63. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

64. O'Connor C, Gatewood C, et al. The Boom Before the Ban: QAnon and Facebook; 2020.

65. Moussaid M, Kammer JE, Analytis PP, Neth H. Social influence and the collective dynamics of opinion formation. PloS one. 2013; 8(11):e78433. https://doi.org/10.1371/journal.pone.0078433 PMID: 24223805