Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

You Are a Bot! - Studying the Development of Bot Accusations on Twitter

Dennis Assenmacher[1*], **Leon Fröhling**[1*], **Claudia Wagner**[1,2]

[1]GESIS - Leibniz Institute for the Social Sciences, Cologne, Germany
[2] RWTH Aachen, Aachen, Germany
{dennis.assenmacher, leon.froehling, claudia.wagner}@gesis.org

Abstract

The characterization and detection of bots with their presumed ability to manipulate society on social media platforms have been subject to many research endeavors over the last decade. In the absence of ground truth data (i.e., accounts that are labeled as bots by experts or self-declare their automated nature), researchers interested in the characterization and detection of bots may want to tap into the wisdom of the crowd. But how many people need to accuse another user as a bot before we can assume that the account is most likely automated? And more importantly, are bot accusations on social media at all a valid signal for the detection of bots? Our research presents the first large-scale study of bot accusations on Twitter and shows how the term bot became an instrument of dehumanization in social media conversations since it is predominantly used to deny the humanness of conversation partners. Consequently, bot accusations on social media should not be naively used as a signal to train or test bot detection models.

Introduction

Automated accounts on social media platforms (bots) have gained a lot of attention from a broader public in recent times, as tech entrepreneur Elon Musk, in a turbulent year for the social media platform Twitter (𝕏)[1], started bringing up his quest to combat bots among the main motivations to acquire the platform in spring 2022. He then did not agree with Twitter's estimates of the prevalence of bots on the platform and tried to use this disagreement as justification for terminating the agreed-upon deal. After ending up acquiring the platform anyway, Musk made *fighting the bots* a top priority for the platform's remaining resources. While the academic discourse on bots usually receives a lot less attention than during the months of the Twitter takeover, the questions of *whether*, *how many*, and *what types of bots* act on online discussion platforms have a tradition of being heavily discussed. The issues of who is considered a bot and what their prevalence and influence really are did not just arise in 2022, but researchers have been investigating the phenomenon of social bots for years. Bots have been found (partly) responsible for many electoral surprises, like the outcome of the Brexit referendum (Howard and Kollanyi 2016) and the election of Donald Trump as U.S. president in 2016 (Bessi and Ferrara 2016), or large-scale disinformation campaigns, for instance around many aspects of COVID-19 (Himelein-Wachowiak et al. 2021; Ferrara 2020). These studies focus on identifying and measuring the influence that bots, through the propagation of certain opinions and sentiments in online networks, have on human users, allegedly allowing the bot operators to not only steer the online discourse but also impact important offline events.

While most research on social bots is focused on the development of detection methods and the characterization of suspected bot operations (Yan and Yang 2022), in this exploratory study, we switch perspectives and approach the issue of bots on social media platforms in general and Twitter in particular through the user's perspective. We present the first large scale study on bot accusations on Twitter and explore to what extent those accusations may be useful for training automated bot detection systems. Further, we test the assumption that "bot" is used as a pejorative term to indicate disagreement and discredit (Wischnewski et al. 2022; Halperin 2021) on a larger scale.

Unlike previous research we do not rely on settings where participants are faced with constructed bot accounts in an experimental setup and later surveyed for their experiences (Yan et al. 2021; Wischnewski et al. 2021), but follow an empirical approach and analyze inter-user communications, in particular situations in which one Twitter user *accuses* another one of being a bot.[2] This allows us to not only explore the characteristics of the accounts frequently accused as bots by other Twitter users, but also gives us insights into the topical contexts as well as the motivation and reasoning provided in the accusations, as they often contain justification for the verdict. Leveraging data from Twitter's inception in 2007, we explore the context and meaning of the bot accusations from different perspectives and track their evolution over the long term.

[*]These authors contributed equally.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1]In this work we will refer to the platform as Twitter, which was its name at the time our data was collected.

[2]We refer to a *user* as a Twitter account that could be potentially controlled by either a real human or an automated software.

Concretely, we follow three distinct research questions:

- RQ1: How did bot accusations change over time?
- RQ2: What are the contexts in which Twitter users accuse others of being a bot?
- RQ3: Do the definitions of bots internalized by Twitter users align with the definitions used in popular bot detection methods?

The implications of our research are twofold. Firstly, our research has implications for scholars and practitioners interested in developing automated bot detection systems since our results show that bot accusations on social media should not be naively used as a signal to train bot detection models or as ground-truth data (which is rarely available) to test such models. Secondly, our research provides empirical support for Haslam's theoretical argument that dehumanization is not only an important phenomenon in the intergroup context but also in the interpersonal context (Haslam 2006). Our results highlight that the term bot became an instrument of dehumanization since it is predominantly used to deny the humanness of conversation partners.

With this exploratory analysis, we publish the underlying "You're a bot!" datasets, containing the accusation situations that we analyze throughout the following pages.[3] These datasets allow for a number of follow-up studies, including an investigation of the campaigns and types of bots recognized by Twitter users and a more in-depth, qualitative study of the Tweet patterns and user profile features that might have triggered the bot accusation. As we hope to motivate in the following study, we would like to see research efforts being redirected towards studying the effects of bot accusations on individual users and the impacts of the dismissal of human opinions as "bot chatter" on the public discourse.

Related Work

Only very recently, bot research has begun to move from working on detection possibilities to focusing on the user perspective, asking how platform users interact with and are affected by bots. Wischnewski et al. study the perceptions of political social bots on Twitter in an experimental study and investigate their participants' ability to distinguish explicitly partisan bots they created from human Twitter users. Their *motivated reasoning* hypothesis assumes that *opinion-incongruent* accounts, i.e., accounts of different political beliefs, are more likely to be perceived as bots. While they have to reject this hypothesis for the main set of participants, they find that it holds true for the more experienced Twitter users, leading them to speculate that this effect stems from a different usage of the term (social) bot between the two groups. According to the authors, *"participants with prior knowledge of social bots and participants who spend more time on social media might apply the term social bot as a pejorative term to indicate disagreement and discredit accounts"* (Wischnewski et al. 2021). While they speculate that the desire to "show disagreement by labeling accounts as social bots (expressive disagreement)" amplifies the effect of motivated reasoning, they do not find other evidence for this claim than a single blog post from 2019[4] and a general reference to "popular media".

In a follow-up study from the same research group, using a similar experimental design, Wischnewski et al. investigate the willingness of Twitter users to interact with *social bots*. They hypothesize and subsequently show that users are more likely to interact with *bots* that are more human-like, and that they are generally more likely to interact with *opinion-congruent* bots, i.e., those bots perceived to be of the same partisanship as the user (Wischnewski et al. 2022). While their hypothesis that affirmative actions (following and retweeting) are more likely to be directed towards opinion-congruent accounts holds, their assumption that *motivated reasoning* does not apply to more ambiguous actions (quote-tweeting and replying) is rejected. Their hypotheses that those actions are thus equally likely to be directed towards *opinion-congruent* and *-incongruent* accounts do not hold. They thus find that Twitter users are more likely to interact with other accounts they perceive to be of the same partisanship, no matter whether in actions signaling agreement or being of a more ambiguous nature.

As an exception from the purely experimental studies presented before, Halperin investigates the user discourse about automated manipulation on Facebook pages in the context of the lead-up to Israel's national elections in 2019 in a data-driven way. The author qualitatively assesses 525 user comments originating from Israeli politicians' posts in which bots are explicitly mentioned. He finds that users widely differ in their understanding of the construct and use it to push their political agenda. Moreover, Halperin finds that partisan commenters *"strip the term of its original technical connotations and recast it as a pejorative concept used to belittle and delegitimize human right-wingers"* (Halperin 2021). Halperin concludes that this new direction in the discourse on automated accounts leads to new possibilities to attack opponents in an online conversation and may even increase the division between political camps. While the study provides interesting insights into discussions and accusations around bots, it's anecdotal nature (focus on a small sample of comments in context of a concrete political event) is mentioned as a limitation by the author, who emphasizes the need for larger studies across the social media sphere with additional textual analyses.

Törnberg develops a model for the increased polarization of society, rejecting the echo chamber hypothesis (selective exposure and isolation from opposing views drive polarization) in favor of one based on affective polarization, where the interactions outside of the local bubble, facilitated by social media, drive polarization (Törnberg 2022). Affective polarization literature finds that opposing partisans have grown to "dislike, even loathe" each other (Iyengar, Sood, and Lelkes 2012). The rejection of the echo chamber hypothesis by Törnberg is supported by empirical evidence, finding

[3] All analysis scripts and a list of Tweet IDs can be found in the following repository: https://github.com/Dennis1989/YaB.

[4] https://saoornik.medium.com/everybody-i-dont-agree-with-is-a-russian-bot-or-how-it-is-easier-to-believe-an-evil-mastermind-ca02391055cb

that social media users of opposing political worldviews do indeed interact with each other (Barberá et al. 2015). However, Törnberg describes these interactions as "contentious and conflictual" rather than "rational arguments and deliberations" held in good faith. In conclusion, Törnberg describes the affective polarization arising on social media as characterized by "difference, distrust, and disdain for one's political opponent".

Bot Definitions

Providing a clear definition of what a bot is remains one of the main challenges for any research on bots in a social media context. Researchers often ground their definition of the concept bot in past publications, citing a range of studies that report having found bots acting in different contexts (e.g., manipulating the discourse on a certain topic or influencing voters and impacting elections). Grimme et al. provides an overview of the changing understandings of the concept of "bot" in academia. In the early years, bots occurred only as chat-bots, built for a specific topic and deployed in one-to-one communication settings (Grimme et al. 2017). This changed when spam-bots were initially developed, specialized in one-to-many-communication, and therefore key to quickly and widely amplifying content. Finally, a more recent definition that goes beyond the aspects of automation and one-to-many-communication comes from Ferrara et al., who define a bot as a program *"that automatically produces content and interacts with humans on social media"* (Ferrara et al. 2016). This definition is often additionally extended by the notion that these bots attempt to mimic human users (Abokhodair, Yoo, and McDonald 2015; Stieglitz et al. 2017). In political contexts, a specific objective is frequently added to the definition; smearing opposing candidates (Metaxas and Mustafaraj 2012), drowning political discussions (Thomas, Grier, and Paxson 2012), or interfering with important political events (Woolley and Howard 2016; Bastos and Mercea 2019; Bessi and Ferrara 2016). Once the first hurdle of providing a definition is cleared, researchers need to operationalize that definition to differentiate between bots and non-bots in a given dataset. A frequent reason for a deficit in validity in social media research is a mismatch between the definition for a specific construct (e.g., bot) and an operationalization that is measuring a different construct (Sen et al. 2021). While some researchers use account characteristics (username, profile description, profile image), activity (frequency of posting, share of retweets), or content (amount of hashtags, amount of URLs) as features to construct rules upon (Kollanyi, Howard, and Woolley 2016), which are then used to classify accounts as bots, many others rely on the data-driven nature of supervised Machine Learning (ML) methods. The most prominent example of this approach is *Botometer* (Davis et al. 2016). Since its publication, *Botometer* has been used to detect bots in several studies in different contexts, with its API and the corresponding ease of use contributing strongly to its popularity. Supervised ML approaches do not require hand-craft rules based on which accounts are classified but leverage the ability of ML to identify patterns in a large number of features and thereby learn classification rules directly from the provided training data. The operationalization in these approaches is then not stated explicitly but induced by the training data and represented by the trained model. This makes it even more complicated to align definition and operationalization, requiring researchers to check whether the characteristics of the instances labeled as bots in the training data are the same as those implied by definition.

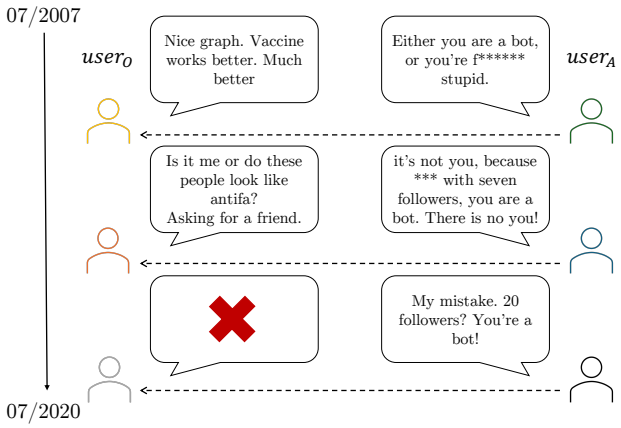

Figure 1: Dataset structure with different accusation situations over time. Because of the data collection strategy, the objects $user_A$ and $tweet_A$ are always present in any accusation situation, whereas the objects $user_O$ and $tweet_O$ could be missing since the tweet and/or the user account have been deleted and are consequently no longer available through the Twitter API.

Accusation Datasets

We are interested in accusation situations; communication situations in which one Twitter user accuses another user of being a bot. We collect these situations by searching for bot accusations in the set of all Reply-Tweets that contain the keyword bot. Therefore, the accusation situations we collect always follow the same pattern. One Twitter user ($user_O$) posts a Tweet ($tweet_O$ - we refer to this Tweet as the *original* Tweet, since it is the first Tweet in our accusation situation, however, it is not necessarily the first Tweet in the full conversation) of any type or content, to which another Twitter user ($user_A$) replies with a Tweet containing a bot accusation ($tweet_A$) (see Figure 1). To construct a dataset of situations in which one Twitter user accuses another Twitter user of *being a bot*, we split the dataset creation process into two phases. During the first phase, we collect as many *candidate* accusation situations as possible, accepting an increased number of false positives while aiming for a high recall. In the second phase, we filter out instances deemed irrelevant to our research design, hoping to reduce the number of false positives while ideally maintaining high precision. In the following, we describe the data collection (first phase) and data processing (second phase), the resulting datasets, and provide some descriptive statistics of them.

Data Collection

For data collection, we used the academic access to the Twitter v2 API in order to obtain access to its full-archive search endpoint.[5] This endpoint returns all Tweets matching a given query. It is important to note that the endpoint employs a token-based strategy for matching the keyword specified in the query, meaning that only Tweets containing the keyword either as a freestanding word or preceded and/or followed by a punctuation mark are matched by the API. Together with the other known but inevitable limitation of Tweets that have been removed or deleted from the platform not being available for retrospective collection through the API, this collection strategy allows us to collect all Tweets matching our query since the inception of Twitter in 2007 [6].

We utilized the query *"bot is:reply lang:en"* to match all Tweets containing the key-token "bot" that are sent as a reply to another Tweet and that are classified as written in English by the Twitter API. Our choice of query and particularly the focus on Reply-Tweets serves as the first necessary step to collect what we are calling accusation situations, i.e., situations in which one user accuses another user of *being a bot*. By restricting our data collection to English Tweets, we avoid complications associated with the processing and analysis of textual data in different languages. As English is the main language used on Twitter and the default language used by international and professional audiences, we are confident that even with this restriction, we cover an interesting share of the bot accusations on Twitter. However, any findings will, by design, only apply to English-speaking user communities on Twitter.

After the data retrieval for the whole available period, beginning April 2007 and ending December 2022, we were left with a dataset containing 22,275,139 Tweets[7]; replies that contained the keyword bot and that we thus consider *potential* accusation situations (bot_{all}). As visualized in Figure 1, each accusation situation ideally consists of four different objects (and their associated metadata). However, since we are matching any potential accusation situation from the Twitter API solely based on $tweet_A$, we can only be sure that $tweet_A$ and $user_A$, i.e., the *accusing* Tweet and the user who sent it, are actually available in our dataset. This is because both objects must have been available at the moment of data collection, as otherwise they would not have been matched via the API. $tweet_O$ and $user_O$, on the other hand, were only included in the dataset if they have not been deleted on Twitter and consequently were still available at the moment of data collection. Otherwise those instances have been marked as missing in our dataset.

[5]https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

[6]Please be informed that the academic API was discontinued in April 2023.

[7]This is already filtered down from the 39,896,323 Tweets returned from the API queries. We removed Tweets that had the bot token only in the Tweet's mentions (and not in the remaining *main* text), as these Tweets might be directed towards an account with *bot* as part of the username without mentioning the term bot in the main message, e.g. "@bot123 how is your day?"

Data Processing

Based on our data collection strategy, it is evident that not all of the retrieved tweets are relevant, as they may not contain bot accusations and need to be filtered accordingly. To address this issue, we decided to implement a two-phase filtering approach. In the first phase, we fine-tuned a language model for detecting bot accusations. Two expert annotators labelled 2,000 potential accusations that were randomly sampled from the bot_{all} dataset to fine-tune the model (training details are provided in the Methods Section). The annotation task was framed as a binary classification problem, determining if a Tweet posted by $user_A$ is *"accusing another Twitter user of being a bot, irrespective of whether the accusation is implicit or explicit."*. By training on this randomly sampled, annotated subset, the final model learned a wide range of different accusation patterns and was able to exclude irrelevant Tweets, such as those containing self-accusations like "I am a bot." Using the fine-tuned model, we filtered the original dataset down to more than 9 million accusation situations ($bot_{general}$), which were used to investigate the evolution of accusations (RQ1) and the topical accusation contexts (RQ2). However, for the investigation of the agreement between accusations and bot scores (RQ3), we had to introduce an additional filtering step. While our model is able to identify accusations, it is often not clear if $user_A$ is truly accusing $user_O$, for example, in Tweets like *"@User1 @User2 @User3 Clearly a bot! stupid moron."* Since RQ3 is based on the assumption of direct accusations, we introduced a second filtering step to exclude any ambiguous tweets.

To increase the precision of the dataset used to answer RQ3, we filtered the situations, retaining only the ones in which $tweet_A$, the *accusing* Tweet, contained the regular expression *"you are a [a-z]*bot|you're a [a-z]*bot"* (bot_{direct}). While this inevitably lowered the recall of our dataset by dropping situations in which the bot accusation does not explicitly follow our template (e.g., "You look like a bot to me.") or is more verbose in its use of the template (e.g., "You are a stupid bot!"), we deliberately decided to trade off some of the high recall of the $bot_{general}$ dataset for a very high precision in the bot_{direct} dataset. Using this rather strict template, we are confident that the resulting dataset almost exclusively contains situations in which $user_O$ is directly accused of being a bot.

To summarize, our prepossessing steps resulted in three datasets of different granularity that help us to investigate our research questions:

- bot_{all} (all Reply-Tweets containing "bot")
- $bot_{general}$ ($user_A$ accuses some other user)
- bot_{direct} ($user_A$ accuses $user_O$)

The distributions of all datasets can be seen in Figure 2.

Dataset Characteristics

In each of the datasets in Figure 2, the distribution of instances over time hints at a shift that has happened in the prevalence of bot accusations around 2017. Prior to 2017, the average yearly share of $bot_{general}$ accusations in the

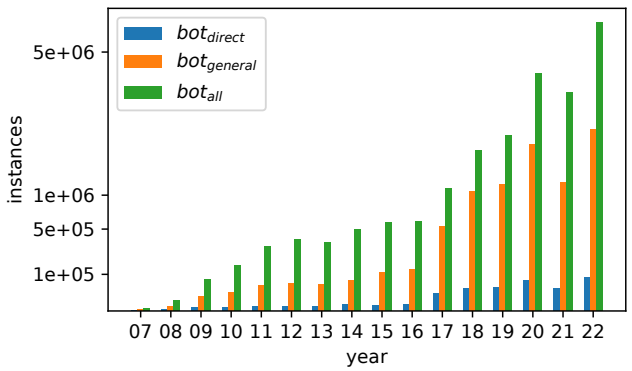

Figure 2: Number of instances in the different data subsets. While the sizes of the different datasets differ, their developments follow the same trends, like the steep increase after 2017.

	$bot_{general}$	bot_{direct}
all accusation pairs	9,069,182	309,107
complete accusation pairs	65,5%	62,6%
unique $user_O$	2,683,510	171,215
unique $user_A$	2,391,749	185,254

Table 1: Composition of the different accusation datasets.

number of bot_{all} Tweets collected was at 17.4%. In the years since 2017, the share of $bot_{general}$ accusations increased drastically, to an average yearly share of 46.3%.

As discussed above, it was not possible to collect the full accusation situation as depicted in Figure 1 for all instances in the datasets due to Tweet or user account deletions. Table 1 shows the composition of the two accusation datasets that we use for our analysis.

In the bot_{direct} dataset for which we know that it is $user_O$ who is being accused, the $user_O$ most often accused of being a bot was accused 1,321 times, by 983 different accusers. The $user_A$ responsible for the highest number of accusations has accused 747 different $user_O$, through 1,344 Tweets. 15,748 different $user_O$ have been accused by more than one $user_A$ of being a bot, being accused on average by 3.28 different users. 24,731 $user_A$ have accused more than one $user_O$ of being a bot, accusing on average 3.67 different $user_O$ of being bots.

Methods

In the following we present the computational mixed-methods used to filter out accusations and answer the research questions posed in the Introduction.[8]

Accusation Detection with BERT

To filter out general bot accusations from all reply Tweets, we utilized BERTweet (Nguyen, Vu, and Tuan Nguyen 2020), a transformer-based BERT model that was pre-trained on a large corpus of English Tweets. We annotated 2,000 randomly selected instances from our $bot_{general}$ dataset using two expert annotators, achieving an inter-annotator agreement of $\kappa = 0.82$. During the annotation process, we encountered instances that were ambiguous and challenging to classify without additional conversational context. For these instances, we mapped them to the negative class to prioritize precision over recall. We divided the annotated data into train, validation, and test sets with a split ratio of 0.8, 0.1, and 0.1, respectively. We fine-tuned multiple models on the classification task using a hyper-parameter optimization strategy that employed random search. For this purpose, we utilized the machine learning framework Ray.[9] We fine-tuned 20 individual parameter constellations resulting from the optimization strategy for four epochs. Finally, we selected the best-performing model based on its ability to filter out all accusations in our validation set. The final model achieved a macro F1 score of 0.93, recall of 0.92, and precision of 0.94 on our hold out test data. Our experiments were conducted on a NVIDIA A100 (80GB) GPU.

Word Embeddings Over Time

Word-embeddings capture semantic similarity via word co-occurrences in vector space. Similar to previous work on shifts in the meaning of individual words over time (Garg et al. 2018), we track the evolution of bot accusations by training Word2Vec-embeddings on the $tweet_A$ in the $bot_{general}$ dataset collected for each year between 2007 and 2022 and inspecting the words that are most closely associated with the term *bot*. We approximate the distance between words through their cosine-similarity. We use Gensim's implementation of Word2Vec with negative sampling and CBOW (as we are interested in the evolution of a high-frequency word).[10] To account for the well-known nondeterministic behavior of embedding models (Hellrich and Hahn 2016) and increase the robustness of our findings, we train five different embeddings for each year and consider only *stable* neighbors, i.e. those words that are closely associated with the term *bot* in all five embedding spaces. As the first years of the data collection consist of considerably lower numbers of Tweets (see Figure 2), we aggregate the years from 2007 to 2016 into one collection and treat each following year individually.

Clustering

To understand in which topical contexts users on Twitter are accused of being bots, we identify prominent clusters in the accused users' Tweets in the $bot_{general}$ dataset through unsupervised learning. We transform their Tweets ($tweet_O$) into document embeddings using sentence transformers (Reimers and Gurevych 2019). Similar to the embedding approach, we use cosine-similarity to approximate distance in the resulting vector space and to group together documents in close proximity. We tested several thresholds and found that a threshold value of 0.7 resulted in the most

[8]The evaluation scripts can be found here: https://github.com/Dennis1989/YaB.

[9]https://www.ray.io

[10]https://radimrehurek.com/gensim/models/word2vec.html

coherent and pure clusters. The resulting clusters are projected into two-dimensional space using Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville 2018). For each identified cluster, we calculate a class-based variant of TFIDF (cTFIDF), which aggregates all cluster documents into one artificial document. This ensures that only the important tokens characterizing each cluster are highlighted.

Toxicity

To measure the toxicity of Tweets over time, we utilize the pretrained Detoxify classifier (Hanu and Unitary team 2020). The model offers a fine-grained labeling schema and is able to classify Tweets at scale. We use the pretrained base models without fine-tuning. To test the general performance of the classifier in detecting toxicity in Tweets in our specific scenario, we annotated 100 random Tweets from our $bot_{general}$ dataset manually. On this test set, Detoxify reached a balanced accuracy of 0.90. To offer an approximation of the overall toxicity on Twitter as a baseline, we collect a dataset of generic English-language replies with the same settings described in our data collection, using the query *"a is:reply lang:en"*. This baseline, however, is limited to showing overall trends and patterns in the toxicity on Twitter, and we do not use it to compare absolute levels of toxicity.

Bot Scores

We utilize the latest *Botometer* version (Sayyadiharikandeh et al. 2020) to see how far the definitions of bots internalized by Twitter users, as operationalized through their accusations, align with the definition operationalized in this popular tool for bot detection on Twitter. Botometer is a supervised ML model, with the latest version using an ensemble of specialized Random Forest classifiers that were trained on a plethora of existing *ground truth* bot datasets (different annotation strategies were used to identify bots in these datasets, such as honeypot traps, manual annotation, or self-report). Apart from calculating an overall score indicating how "bot-like" an account is, the classifier also returns multiple subscores corresponding to more specific bot definitions, such as "spammer" or "fake follower". Although the Botometer classifier is a helpful tool for understanding the degree of automation of a Twitter account, it comes with significant limitations. Several studies have previously demonstrated the limits of such classifiers due to concept drift in data and the uncertainty of the underlying ground truth (Martini et al. 2021; Rauchfleisch and Kaiser 2020). For our experiments, we utilize the method as a means to measure how the understanding of the concept bot held by Twitter users aligns with the academic construct definition operationalized by the classifier. We do not use the tool to determine how many and which accounts are bots and thus do not intend to construct a labeled ground truth dataset from our data. However, we extract a subset of accounts that are repeatedly accused of being bots by different Twitter users, for which we then collect their Botometer scores. To increase validity and tackle noise, we collect scores only for the 13,638 $user_O$ in the bot_{direct} dataset that have been accused of being bots by at least two *different* $user_A$. We collect their *bot*-scores from the Botometer API[11].

Ideology

We measure the ideology of accounts by utilizing Barberá's method for ideal point estimation (Barberá 2015). The method infers ideology by examining the political following network of Twitter accounts, assuming that social networks are homophilic. Since the method was developed using a recent set of political accounts in the United States, we restrict our analysis to users from there and to accusations made after 2016. We thus only consider accusation instances in which both $user_O$ and $user_A$ self-report as being from the United States. We use the meta-information *location*, a free-text field used by some Twitter users to indicate their location, in combination with Google's Geolocation API[12] to automatically parse the location information and retrieve the corresponding country. 76,855 of the accusation situations in our bot_{direct} dataset posted after 2016 had *location*-information for both users. To handle API restrictions, we randomly extract an approximate 10% sample, leaving us with 7,131 pairs of $user_O$ and $user_A$ for which we parse the location information with the Geolocation API. In this sample, we found 2,670 conversation pairs where both accounts are from the US, which we use as input for our ideology estimation method.

Results

In this section we answer the following three research questions:

- RQ1: How did bot accusations change over time?
- RQ2: What are the contexts in which Twitter users accuse others of being a bot?
- RQ3: Do the definitions of bots internalized by Twitter users align with the definitions used in popular bot detection methods?

RQ1: Evolution of Accusations

Figure 3 shows the nearest neighbor embeddings of the term bot over the years covered in our dataset. As the number of accusations for the early years up to 2016 is significantly lower (see Figure 2), we aggregate the accusations made in those years to get a sufficient number of observations for training the embeddings. The color coding we applied to the nearest embeddings vectors displayed in Figure 3 shows how the meaning of the bot accusations shifted over the years. In the years leading up to 2016, the term bot is closely and firmly associated with terms representing the automated behavior often used in academia to conceptualize bots. These include *software*, *script*, or *comments*. In the following years, the wording of the bot accusations shifted away from such signs of automation in favor of terms used to insult the accused user. The use of derogatory terms such as

[11] https://botometer.osome.iu.edu/api

[12] https://developers.google.com/maps/documentation/geocoding/overview

2007 - 2016	software , comments , person, user, hashtags , script , machine , database, guy, acct , program ,
2017	troll , idiot , person, probably, moron , entity, tool, real, account , human, paid,
2018	troll , idiot , person, probably, moron , supporter , entity, dolt , joke, shill , tool, writer, account , fool , human, russian , paid,
2019	definitely, troll , idiot , product, supporter , fool , human, account , probably, person, moron , parody, caricature, asset, stooge , robot, joke, tool, russian ,
2020	troll , idiot , trumper , person, foreigner , teenager, probably, supporter , parody, moron , robot, joke, tool, fool , human, russian ,
2021	definitely, troll , idiot , foreigner , fool , human, account , probably, shill , person, parody, moron , paid, robot, joke, tool, real, chinese , russian ,
2022	troll , propagandist , idiot , operative, foreigner , fool , account , simpleton , probably, shill , liar , person, parody, moron , satire, chinese ,

Figure 3: Nearest embedding vectors to the term *bot* over the years. We highlight terms associated with mechanics for automation in blue and dehumanizing/insulting/political terms in red.

moron, *stupid*, *idiot*, and *shill* implies that the accuser views the accused as less than human, often by questioning their mental capacity. This indicates that over time bot accusations became predominantly instances of "dehumanization" (Haslam and Loughnan 2014). By studying inherently interpersonal accusation situations, we find empirical support for Haslam's theoretical argument that "dehumanization is an important phenomenon in interpersonal as well as intergroup contexts" and also "occurs outside the domains of violence and conflict" (Haslam 2006).

For direct accusations to $user_O$ (bot_{direct}) we observe an increase in the degree of toxicity over time. Figure 4 shows the toxicity for $tweet_O$ and $tweet_A$ over time, as well as a baseline that serves as a proxy for how the overall toxicity in Tweet-Reply situations developed in parallel. It is evident that beginning with the year 2015, the toxicity of the $tweet_A$ increased strongly, settling at a new plateau around the year 2017. The baseline shows that this increase can not just be explained by the rise of the overall toxicity on Twitter, as the toxicity for replies outside our bot accusation context remains stable and relatively low. Similarly, the toxicity of the original Tweets leading up to the accusations is comparatively low and has a much less pronounced increase around 2016.

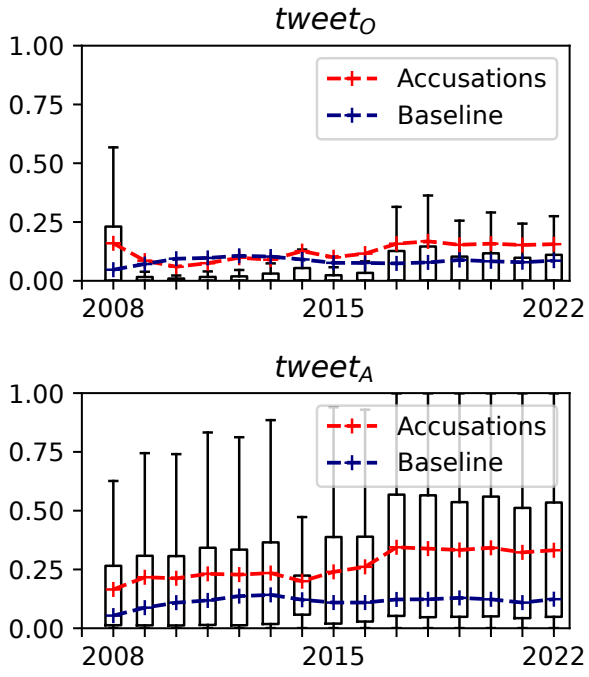

Figure 4: Development of *Detoxify* toxicity scores for $tweet_O$ and $tweet_A$ objects. Comparison with the baseline of generic Reply-Tweets indicates that the observed increase in toxicity for the accusations ($tweet_A$) after 2016 cannot be explained through an overall increase of toxicity on Twitter over time.

RQ2: Accusation Context

The shift towards dehumanization that we observed when analyzing the bot accusations over time also becomes visible in the results of our topical analyses. A manual investigation of the cluster solutions produced by the cluster algorithm discussed in our method section for the early years (2007-2016) revealed that users were frequently accused because they showed signs of automation, such as spamming repetitive content, for instance:

"follow me i follow you"
"good morning! you deserve a fantastic day today!"
"i'm a human i'm a human i'm a human".

Additionally, corporate accounts that automatically respond to public customer complaints with standardized replies like

"please dm your concern to help us assist you."

were frequently accused of being bots. We also found a considerable number of accounts that were accused because they openly discussed that they reached the limit of accounts they could follow or ran into rate limits for sending out new Tweets, such as

"oh dear, twitter says i'm not allowed to follow anymore people. what's that all about then?.

The tweet content posted by different $user_O$ that had led to an accusation situation has significantly changed over the

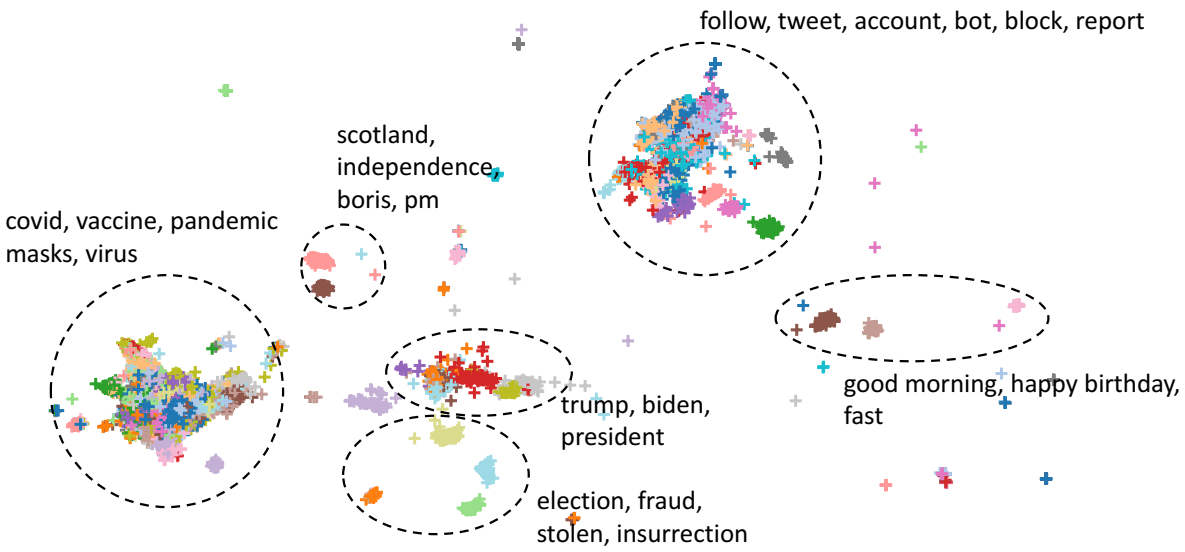

Figure 5: Top 100 clusters for the Tweets sent by $user_O$ (the user accused of being a bot) for 2021 projected onto a 2-dimensional space with UMAP. Clusters are annotated with their highest cTFIDF terms. Users are not anymore accused only in the context of automated behavior ("good morning spam") but specifically in the context of polarizing debates around topics like covid/mask/vaccine, election/Biden/Trump or Scottish independence. Additionally, we observe a large cluster indicating bot accusation loops (e.g. $user_O$*: you are a bot - blocked!* $user_A$*: i'm not a bot, but you are!*).

years. Recently (especially since 2020), most $user_O$ are accused of being bots in the context of controversial, polarizing political or politicized topics such as elections (Biden vs. Trump), the pandemic (vaccination, mask mandates), or the Brexit vote (Boris Johnson). Figure 5 displays the top cluster distribution for the year 2021. Of the twenty biggest clusters (in terms of cluster size) in 2021, thirteen were concerned with political debates, and only the remaining seven focused on repetitive automation behavior (most of them indicating accusation loops). We found that specifically $user_O$ and $user_A$ communication pairs that directly accuse each other in the context of these political topics are found on opposing sides of the ideological spectrum, e.g., users from the left of the ideological spectrum accusing users from the right of the ideological spectrum in the context of the topic of Trump vs. Biden (see colored observations in Figure 6). Interestingly, most accusers are more left-leaning, while accounts that are accused are more right-leaning (most observations are in the bottom-right quadrant in Figure 6). When it comes to intra-ideology conversation pairs, it is evident that accounts on the political Right tend to not accuse each other (upper right quadrant), while accounts that are associated with the political Left accuse each other more frequently (lower left quadrant), especially in the context of political debates.

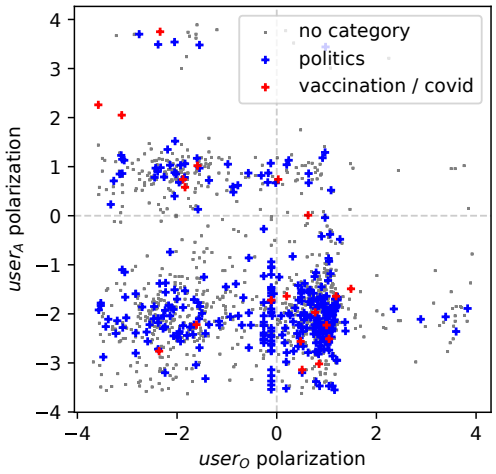

Figure 6: Ideology scores for $user_O$ and $user_A$ according to Barberá (Barberá 2015). Negative scores are associated with Left-leaning political positions, while positive scores are associated with Right-leaning political positions. One can see that most accusations occur in polarizing topics and are directed from left-leaning towards right-leaning users (lower right quadrant). Users on the political Right tend to not accuse each other, while users on the political Left accuse each other more frequently.

RQ3: Accusations vs. Bot Scores

We investigate the alignment between bot accusations and the bot probabilities assigned by popular detection mechanisms by inspecting the Botometer scores for the $user_O$ in our bot_{direct} dataset. Figure 7 displays the distribution of

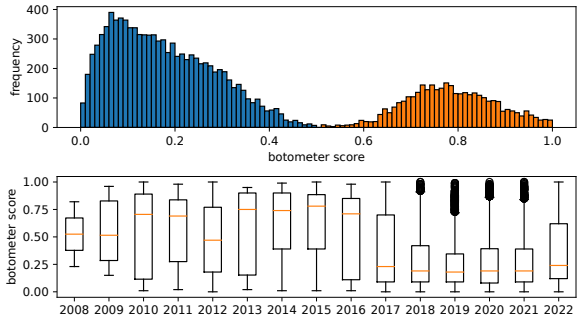

Figure 7: Distribution and development over time of Botometer scores for $user_O$ (users *accused* of being bots). Between 2008 and 2016 accusations exhibited high Bot scores, while this abruptly changed beginning of 2017.

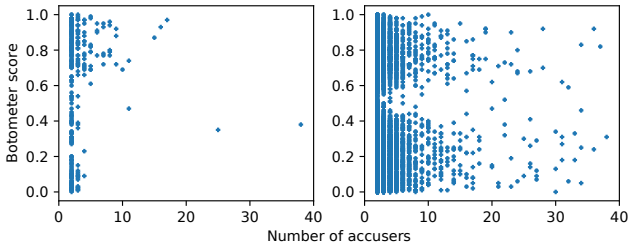

Figure 8: Scatter plot showing the relationship between the number of accusers and the Botometer score for the accused users. While high Botometer scores were associated with higher numbers of accusations that an user received before 2017 (left subfigure), this association does not hold anymore after 2017 (right subfigure).

the bot probabilities for the accused accounts as calculated by the Botometer model. The scores follow a bimodal distribution, indicating that accounts are quite confidently assigned into one of the two classes *human* or *bot*, with a larger fraction of accounts having lower scores and thus tending towards the *human* category. Consequently, there seems to be a discrepancy between the operationalization of the bot construct employed by Botometer and the definition internalized by the Twitter users. In a subsequent analysis, we explored how bot scores developed over time. The results are depicted in the bottom part of Figure 7. Interestingly, the accounts accused between 2008 and 2016 exhibit significantly higher bot scores than those between 2017 and 2022. This decrease in bot scores is well in line with the shift in the meaning of the accusations that we found before, with accusations not being exclusively used for actual automation behavior anymore. Apparently, in the early years, the Botometer scores were aligned with the accusations made by the Twitter users, as the accused users, $user_O$, also have high Botometer scores. These were the years in which the majority of $user_O$ were accused because of their automation behavior (repetitive tasks such as (re-)tweeting or large-scale following). With the observed transformation of bot accusations into a dehumanizing insult (starting in 2017), the average bot scores dropped significantly. While this is not necessarily an indicator of fewer bot occurrences (which might also be due to potential model errors), it is clearly showing that bot accusations experienced a concept drift.

Additionally, we investigated how Botometer's bot scores are correlated to the number of unique accusers for each $user_O$, expecting to find that the number of accusers increases with the bot score if the definitions internalized by the accusing users are well aligned with the definition operationalized in Botometer. Similar to our previous analyses, we again differentiate between the years before 2017 and from 2017 onwards. Calculating Spearman's rank correlation coefficient, we find a negligible positive correlation for the years after 2017 ($\rho = 0.08$, p-value < 0.0001) and only a weak positive correlation for the early years before 2017 ($\rho = 0.23$, p-value $= 0.0001$). However, a visual inspection of the scatter plots in Figure 8 clearly reveals that Botometer assigns only small bot scores for accounts with few accusations during the early years, whereas for all $user_O$ with more than three accusations, the scores tend to be in the higher regions (between 0.6 and 1). On the contrary, there is no clear pattern for the years from 2017 onwards (right sub-plot). Even with an increasing number of accusations, both low and high Botometer scores are present, indicating that the hypothesized and limitedly shown correlation between both variables does not hold anymore during the later years.

Discussion & Limitations

Our results indicate that bot accusations on the Twitter platform significantly changed over the last decade (RQ 1). Before 2017, platform accounts were mainly accused when they exhibited some explicit automation behavior. Over time and since the aftermath of the US elections in 2016, resulting in an increased interest for bots in media and academia, users increasingly accused other accounts, intending to dehumanize and insult them, questioning their intelligence and denying their right to express their opinion. In accordance with this observation, we find that bot accusations from 2017 onwards primarily occur in the context of controversial, polarizing debates like elections or covid/vaccination (RQ 2). In other words, the term "bot" has transformed into a political term that is used as an insult (or more precisely an instance of dehumanization). This starkly contrasts with the ways in which bot accounts are currently discussed in academia, focussing more on the means of automation to distribute (often) malicious content and their supposed impact on the public discourse rather than the actual usage and meaning of the term bot, as exhibited in our bot accusations. Indeed, we found that in the early years, the bot scores (calculated by Botometer) of accused accounts were significantly higher than in later years, when accusations mainly occurred as an insult to dehumanize other "people" in the network. This finding supports the theoretical argument that dehumanization is an important phenomenon in interpersonal context (Haslam 2006) and provides empirical evidence for the initial assumptions that "bot" is used as a pejorative term to in-

dicate disagreement and discredit (Wischnewski et al. 2022; Halperin 2021) on a larger scale. Our findings also have practical implications for researchers interested in bot detection since we show that bot accusations should not be naively used as a signal for automated bot detection methods or as ground-truth data.

In addition to the investigation of how automated accounts influence users on social media platforms, future research should be concerned with the impact these accusations have on individuals. Additionally, it should be investigated how bot accusations could be countered efficiently. The structure of our dataset paves the way for a plethora of follow-up research in these directions. With additional data collection efforts, the accusation situations examined in this work could be augmented to cover the whole conversation around the bot accusations, including the exchange leading up to it, as well as the reaction afterward.

Our study is not without limitations. Even though we tried to balance the issues associated with precision and recall by using different data collection strategies, the inherent problem of not being able to achieve perfect performance in selecting accusations remains. With our data collection strategies we might, for example, have missed bot accusations that refer to bots using a synonym or the same word in a different language. Also, we highlight that our ideology analysis only considered accusation situations in which both users were located in the United States. In order to use the method on accounts outside the US, the model must be trained again on a new initial set of accounts. While our methodology can be adapted to other platforms and languages, our empirical results are limited to one platform (Twitter) and one language (English). Future endeavors will show whether our findings also apply to bot accusations made on different platforms and in other linguistic and cultural settings.

Ethics Statement

This study uses publicly-accessible user-generated content online as its data source. We follow established research practice in not asking for platform users' consent before collecting this data, however, we pseudonymize all user-generated content before analyzing it. This type of data still carries risks, particularly in the form of misclassifications, as it potentially contains *false accusations*, situations in which actual human users are accused of being a bot. We stress that these accusations represent subjective opinions and should under no circumstances be used to infer the degree of automation of any account. This is especially true in light of our findings that there currently is a discrepancy between the construct definition and the Twitter users' understanding of it, and that these accusations tend to happen in an increasingly toxic and polarized environment. To address these concerns, only the IDs of the Tweet- and User-objects used in this study will be publicly released. To still allow for complete reproducibility, we invite other researchers to contact us for collaboration.

Acknowledgments

This work was created in context of the project: Digital Dehumanization: Measurements, Exposure and Prevalence (DeHum), funded by the Leibniz Association Competition (P101/2020).

References

Abokhodair, N.; Yoo, D.; and McDonald, D. W. 2015. Dissecting a social botnet: Growth, content and influence in Twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, 839–851.

Barberá, P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis*, 23(1): 76–91.

Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10): 1531–1542.

Bastos, M. T.; and Mercea, D. 2019. The Brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1): 38–54.

Bessi, A.; and Ferrara, E. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11-7).

Davis, C. A.; Varol, O.; Ferrara, E.; Flammini, A.; and Menczer, F. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, 273–274.

Ferrara, E. 2020. What Types of COVID-19 Conspiracies are Populated by Twitter Bots? *First Monday*. ArXiv:2004.09531 [physics].

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The rise of social bots. *Communications of the ACM*, 59(7): 96–104.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grimme, C.; Preuss, M.; Adam, L.; and Trautmann, H. 2017. Social Bots: Human-Like by Means of Human Control? *Big Data*, 5(4): 279–293.

Halperin, Y. 2021. When Bots and Users Meet: Automated Manipulation and the New Culture of Online Suspicion. *Global Perspectives*, 2(1): 24955.

Hanu, L.; and Unitary team. 2020. Detoxify. https://github.com/unitaryai/detoxify. Accessed: 2024-04-04.

Haslam, N. 2006. Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, 10(3): 252–264. PMID: 16859440.

Haslam, N.; and Loughnan, S. 2014. Dehumanization and infrahumanization. *Annual review of psychology*, 65: 399–423.

Hellrich, J.; and Hahn, U. 2016. Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2785–2796. Osaka, Japan: The COLING 2016 Organizing Committee.

Himelein-Wachowiak, M.; Giorgi, S.; Devoto, A.; Rahman, M.; Ungar, L.; Schwartz, H. A.; Epstein, D. H.; Leggio, L.; and Curtis, B. 2021. Bots and Misinformation Spread on Social Media: Implications for COVID-19. *Journal of Medical Internet Research*, 23(5).

Howard, P. N.; and Kollanyi, B. 2016. Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum.

Iyengar, S.; Sood, G.; and Lelkes, Y. 2012. Affect, not ideology: A social identity perspective on polarization. *Public opinion quarterly*, 76(3): 405–431.

Kollanyi, B.; Howard, P. N.; and Woolley, S. C. 2016. Bots and automation over Twitter during the first US presidential debate. *Comprop data memo*, 1: 1–4.

Martini, F.; Samula, P.; Keller, T. R.; and Klinger, U. 2021. Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society*, 8(2): 20539517211033566.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Metaxas, P. T.; and Mustafaraj, E. 2012. Social media and the elections. *Science*, 338(6106): 472–473.

Nguyen, D. Q.; Vu, T.; and Tuan Nguyen, A. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14. Online: Association for Computational Linguistics.

Rauchfleisch, A.; and Kaiser, J. 2020. The false positive problem of automatic bot detection in social science research. *PloS one*, 15(10): e0241045.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, 2725–2732. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6859-9.

Sen, I.; Flöck, F.; Weller, K.; Weiß, B.; and Wagner, C. 2021. A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1): 399–422.

Stieglitz, S.; Brachten, F.; Ross, B.; and Jung, A.-K. 2017. Do social bots dream of electric sheep? A categorisation of social media bot accounts. *arXiv preprint arXiv:1710.04044*.

Thomas, K.; Grier, C.; and Paxson, V. 2012. Adapting social spam infrastructure for political censorship. In *5th USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 12)*.

Törnberg, P. 2022. How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42).

Wischnewski, M.; Bernemann, R.; Ngo, T.; and Krämer, N. 2021. Disagree? You Must be a Bot! How Beliefs Shape Twitter Profile Perceptions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8096-6.

Wischnewski, M.; Ngo, T.; Bernemann, R.; Jansen, M.; and Krämer, N. 2022. "I agree with you, bot!" How users (dis)engage with social bots on Twitter. *New Media & Society*, 0(0): 14614448211072307.

Woolley, S. C.; and Howard, P. N. 2016. Political communication, computational propaganda, and autonomous agents: Introduction. *International journal of Communication*, 10.

Yan, H. Y.; and Yang, K.-C. 2022. The landscape of social bot research: a critical appraisal.

Yan, H. Y.; Yang, K.-C.; Menczer, F.; and Shanahan, J. 2021. Asymmetrical perceptions of partisan political bots. *New Media & Society*, 23(10): 3016–3037. Publisher: SAGE Publications.

Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. We are studying a phenomenon isolated on a social media platform, where the actual identities of the participating individuals are not of interest. Instead, we are trying to shed light on the observed practice of dismissing platform users' humanity.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. We propose in abstract and introduction to explore the nature and patterns of bot accusations on Twitter. After conducting different analyses to study the phenomenon from different angles throughout the paper, we end with a discussion of its implications for research.

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Each subsection in our Methods section starts with a brief discussion of the presented method's relevance and suitability of its use for supporting our arguments.

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the Data Collection section.

(e) Did you describe the limitations of your work? Yes, see the Discussion & Limitations section.

(f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethics Statement.

(g) Did you discuss any potential misuse of your work? Yes, see the Ethics Statement.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, see the Ethics Statement.

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we tried our best to follow the recommendations and requirements mentioned in the ethics review guidelines.

2. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? Not applicable.

(b) Have you provided justifications for all theoretical results? Not applicable.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not applicable.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not applicable.

(e) Did you address potential biases or limitations in your theoretical framework? Not applicable.

(f) Have you related your theoretical results to the existing literature in social science? Not applicable.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Not applicable.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? Not applicable.

(b) Did you include complete proofs of all theoretical results? Not applicable.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. All code, data - as much as we are allowed to share - and further instructions for the reproduction of all our experiments are available via https://github.com/Dennis1989/YaB. We encourage other researchers to contact us to discuss the options for access to our full data.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. Model training details are described in subsection Accusation Detection with BERT.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No, because our model development does not aim to strictly achieve optimal performance on a task or benchmark, but to provide us with a practical classifier with acceptable accuracy. However, we ensured generalizability through hyperparameter optimization. The results can be inspected via https://github.com/Dennis1989/YaB.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, see the Methods Section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. We report evaluation scores in subsection Accusation Detection with BERT, and discuss our considerations with respect to the evaluation in the Discussion & Limitations section.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. We discuss the consequences of misclassifications in the Discussion & Limitations section.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Yes. Where possible, we reference relevant publications accompanying those assets. If no publication is available, we mention the creators and provide links to the assets in footnotes.

(b) Did you mention the license of the assets? Yes. We distribute our materials under the GNU General Public License v2.0 license, as specified in https://github.com/Dennis1989/YaB.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. We make our new assets accessible via https://github.com/Dennis1989/YaB.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see the Ethics Statement.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, see the Ethics Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? No, because we cannot fully share Twitter data according to Twitter's ToS.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, because we only share Tweet IDs.

6. Additionally, if you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots? Not applicable.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable.

(d) Did you discuss how data is stored, shared, and deidentified? Not applicable.