The False positive problem of automatic bot detection in social science research

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Rauchfleisch, Adrian, and Jonas Kaiser. &quot;The False positive problem of automatic bot detection in social science research.&quot; Berkman Klein Center Research Publication No. 2020-3 (2020).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42645596">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42645596</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
The False positive problem of automatic bot detection in social science research

Adrian Rauchfleisch *1 and Jonas Kaiser †2,3

1Graduate Institute of Journalism, National Taiwan University
2Berkman Klein Center for Internet & Society, Harvard University
3Alexander von Humboldt Institute for Internet & Society

March 2020

Abstract

The identification of bots is an important and complicated task. The bot classifier Botometer was successfully introduced as a way to estimate the number of bots in a given list of accounts and, as a consequence, has been frequently used in academic publications. Given its relevance for academic research and our understanding of the presence of automated accounts in any given Twitter discourse, we are interested in Botometer’s diagnostic ability over time. To do so, we collected the Botometer scores for five datasets (three verified as bots, two verified as human; n=4,134) in two languages (English/German) over three months. We show that the Botometer scores are imprecise when it comes to estimating bots; especially in a different language. We further show in an analysis of Botometer scores over time that Botometer’s thresholds, even when used very conservatively, are prone to variance, which, in turn, will lead to false negatives (i.e., bots being classified as humans) and false positives (i.e., humans being classified as bots). This has immediate consequences for academic research as most studies using the tool will unknowingly count a high number of human users as bots and vice versa. We conclude our study with a discussion about how computational social scientists should evaluate machine learning systems that are developed to identify bots.

Keywords: bots, bot detection, twitter, social bots, botometer, machine learning, automatic classifier

*adrian.rauchfleisch@gmail.com
†jkaiser@cyber.harvard.edu
This report is part of the Berkman Klein Center for Internet & Society research publication series. The views expressed in this report are those of the authors alone and do not reflect those of Harvard University or the Berkman Klein Center for Internet & Society at Harvard University.

Acknowledgments: A prior version of this paper was accepted for 20th annual meeting of the Association of Internet Researchers in 2019 as well as for the annual conference of the International Communication Association in 2020. We thank Chung-hong Chan as well as Momin M. Malik for their valuable feedback that helped us to substantially improve our paper. We also want to thank Martin Fuchs for providing us with his manually validated list of German Members of Parliament.
1 Introduction

Identifying bots (here understood as fully automated accounts) on social media platforms like Twitter is a task that is as important as it is challenging. Indeed, being able to estimate how authentic any given discourse on social media is or how many “real” (i.e. human) followers a politician has, is an important baseline for the analysis of online debates. This is not only relevant in the context of political communication but especially so when it comes to disinformation campaigns and information security. Consequently, several methods have been introduced to get a handle on this problem, including a generic activity filter (Howard & Kollanyi, 2016), an analysis of posting patterns (Bastos & Mercea, 2019), as well as a trained classifier (C. A. Davis, Varol, Ferrara, Flammini, & Menczer, 2016). Especially the trained classifier Botometer (C. A. Davis et al., 2016; Varol, Ferrara, Davis, Menczer, & Flammini, 2017) has established itself in the social sciences to estimate the number of bots within a given dataset.

In April 2018 Pew Research Center published a large-scale study analyzing Tweets with links to popular websites finding that two-thirds of the analyzed tweets are posted by automated accounts (Wojcik, Messing, Smith, Rainie, & Htilin, 2018). Another study by Bessi and Ferrara (2016) estimated that 19% of the tweets published during the 2016 presidential election in the US were posted by “social bots”. A study conducted in the German context found that 9.9% of the follower of German national party accounts on Twitter are “social bots” (T. R. Keller & Klinger, 2019). What all of these studies have in common is that they used the tool Botometer (Varol et al., 2017) which was formerly known as BotOrNot (C. A. Davis et al., 2016). It is, however, worth noting that these are by far not the only studies making use of this particular form of bot detection method. A simple Google Scholar search for “‘botometer’ OR ‘botornot’” yields 739 results (as of January 29, 2020) with 281 results from 2019. It is thus fair to say that the tool is arguably the most popular bot detection method in the social sciences and applications of the tools have been published both in field specific journals such as Political Communication (T. R. Keller & Klinger, 2019) as well as high-ranked general-interest journals like Science (Vosoughi, Roy, & Aral, 2018) or Nature Communications (Shao et al., 2018)). This, then, would suggest that Botometer is the de-facto standard of bot detection in academia.

Bot detection, however, is notoriously tricky. Not only will bot creators (especially in the context of spam or fraudulent social bots that pretend to be human) try to convince users as well as researchers applying their
detection methods that their bots are legitimate humans but they will also adjust their bots to the newest identification methods and will thus often be one step ahead of researchers. This, of course, goes hand in hand with a certain uncertainty when it comes to detecting bots. There are numerous cases of Twitter accounts that look like bots but are actually human and vice versa; some accounts are both: semi-automated and semi-manual. To account for this complexity, Botometer does not say whether a user is a bot or not but rather will give a probability estimate between 0-1 (1 = most likely bot, 0 = most likely being human; the scale can also be 0-5 as shown on Botometer’s homepage). Researchers consequently have to define a manual threshold that seems reasonable to them if they want to classify accounts in a binary fashion. If an account has a higher score than the defined threshold, it will be defined as bot by the researchers. Wojcik et al. (2018), for example, chose 0.43, while T. R. Keller and Klinger (2019) opted for 0.76. In contrast, Zhang et al. (2019) use a rather low threshold of .25 for the complete automation probability (CAP) which is also provided by Botometer.

For researchers, established tools like Botometer that come with a large body of peer-reviewed publications that lend authority and legitimacy to the tool, are an important resource in their goal to shine a light into the role of bots on social media platforms such as Twitter and their potential impact on the public discourse. Indeed, understanding whether an online discussion is organic (i.e. consists of human accounts) or compromised by bots is an important question, especially in the context of political communication and misinformation. However, Botometer is not above criticism. Grimme, Assenmacher, and Adam (2018) show in their study, for example, that Botometer could not classify the hybrid and full automation bot accounts that the authors had created precisely. The original creators are also aware of the potential limitations of their tool and admit that “many machine learning algorithms, including those used by Botometer, do not make it easy to interpret their classifications” (Yang et al., 2019, p. 58). They further suggest in the official FAQ on the homepage of Botometer that “[i]t’s tempting to set some arbitrary threshold score and consider everything above that number a bot and everything below a human, but this is probably not the best way to think about it [the bot score]” (Botometer, 2019). Yet, as we have already indicated above, many studies relying on Botometer as a tool do use a fixed threshold (T. R. Keller & Klinger, 2019; Vosoughi et al., 2018; Wojcik et al., 2018). Still, there are a few exceptions in the literature where researchers have validated existing tools like Botometer in a first step and, then, opted for a self-developed solution in-
stead. While not explicitly discussing the diagnostic ability of Botometer, Fernquist, Kaati, and Schroeder (2018) indicate in their study of Swedish Twitter data that Botometer struggles with non-English language tweets and report that their own supervised classifier outperforms Botometer.

In this paper we will use the term “bot” throughout. We acknowledge that other researchers have talked about “automated accounts” (Howard & Kollanyi, 2016) or “social bots” (T. R. Keller & Klinger, 2019) but since the creators of Botometer specifically talk about identifying “bots” we will use the same terminology (Botometer, 2019). It is noteworthy that the creators of Botometer will generally speak of “social bots” in their academic work (Bessi & Ferrara, 2016; C. A. Davis et al., 2016; Ferrara, Varol, Davis, Menczer, & Flammini, 2016; Yang et al., 2019), but do not use the term in the tool’s FAQ system. One of the original Botometer creators even understands bot as short for “social media bot” in a recently published study (Luceri, Deb, Giordano, & Ferrara, 2019), adding to the confusion. While there is agreement on the definition of “bot” (automation as a deciding factor), the social in “social bot” is contested. Yet, a debate about this part is beyond the scope of this paper and Gorwa and Guilbeault (2018) have already clarified many aspects of the conceptual confusion in their bot typology.

The goal of this paper is twofold. Firstly, we want to critically assess, as we have shown above, one of the most popular bot detection tools used in social science. Secondly, our contribution should help to reflect on the role of computational methods in computational social science research as some of our findings help to better understand the limitations of machine learning systems that are developed to identify bots on social media platforms.

### 2 Research questions

Given that it is very complicated to identify bots and Botometer’s prominence in the social sciences we ask: How precise is Botometer in detecting bots? More specifically we are interested in 4 different research questions. Firstly, we are interested in the general diagnostic ability of Botometer. We want to assess Botometer with the Receiver Operating Characteristics (ROC) curve as this is the common approach used, for example, by Varol et al. (2017) themselves. This approach allows us to assess how well Botometer can distinguish between bots and human users overall.

**RQ 1 How good is the diagnostic ability of Botometer when used for five distinct sets of Twitter accounts?**
Yet, in practice many researchers analyze data sets that are sampled from the general Twitter population and manually define thresholds of what qualifies as a bot and what not. We are thus interested in the precision (percentage of accounts that are classified as bots) as well as the recall (percentage of identified bots from the whole population).

**RQ 2** *How good is the precision and the recall of Botometer scores when used for five distinct sets of Twitter accounts representing the bot/human ratio in the general Twitter population?*

As Fernquist et al. (2018) analysis already indicates, there might be differences between languages with regard to the performance of Botometer. The data analysis for the first two research questions allows us also to specifically assess whether there are differences between languages.

**RQ 3** *With regard to the diagnostic ability are there differences with the performance of Botometer between languages?*

Grimme et al. (2018) showed in their study with 3 bots in an experimental setting that the classification score is not stable over time. Therefore, we are interested in not only measuring the Botometer score once but instead tracking accounts over a longer period of time.

**RQ 4** *How stable are the Botometer classifications over time?*

In answering these four questions, we want to add to the literature on bots by shining a light on some of the issues that the Botometer has and discuss what consequences this has for social scientists conducting research on bots.

### 3 Data and Methods

To answer these research questions, we rely on 5 data sets. We included all official accounts of German members of parliament based on a list of personally confirmed official accounts (German politicians: n=532) and all members of the 115th U.S. Congress with a Twitter account (US politicians: n=516) (Social Feed Manager, 2017) as “human” accounts. We also created a list of obvious bots (new bots: n= 935). We used all bots listed on https://botwiki.org/, a data base with the goal to preserve interesting and creative Twitter bots as well as a number of other labelled bots we identified on Twitter. Most of these accounts are labelled as bots and for many even their code is available on GitHub. As most of these bots use the
English language, we also collected an additional list of German language bots (German bots: n=27). As a fifth dataset we use a manually annotated data set with human (n=1747) and bot accounts (n=826) that was created by the makers of Botometer (Varol et al., 2017). We combined these data sets for our data analysis (see Table 1). For the first three research questions we need bots as well as human accounts combined as we are evaluating how well Botometer can distinguish bots from human accounts. We therefore combine the different datasets (see Table 1). For RQ4, however, we can use single datasets consisting only of bots or humans.

The selection of the data sets is based on one basic assumption: bot detection tools like Botometer will perform better in distinguishing between bots and humans when the cases are clear-cut. As can be seen in Figure 1, clear cut bot accounts will often outright state in their description that they are, in fact, bots; occasionally even sharing the code on GitHub. Similarly, clear-cut human cases will often have indicators that signal their authenticity such as a blue checkmark (Twitter’s verification badge) or their personal website. To make it even clearer, we chose politicians, i.e. people that are in the public limelight and have often high follower numbers, where any suspicious Twitter activity would automatically lead to media attention and increased scrutiny. Finally, we also included Varol’s list that included both bots as well as humans. While we were not always able to individually verify whether an account was a bot or a human, we trust in the initial categorization which parts of Botometer’s training is based on. As a result, we have four data sets of accounts that human coders would have no difficulty in differentiating between bots and humans and one which was used to train the Botometer classifier on. Based on these criteria, we would expect the Botometer tool to perform rather well as the “real world” data from Twitter is much messier.

To conduct this test, we created a Python script that ran every day at 00:00 UTC time on an Ubuntu Linux server for 3 months (3rd March – 2nd of June, 2019). The data collection went uninterrupted. For every account

<table>
<thead>
<tr>
<th>dataset</th>
<th>In our analysis</th>
<th>Valid accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>German politicians</td>
<td>all/ German politicians and bots</td>
<td>human=516, bots=27</td>
</tr>
<tr>
<td>US politicians</td>
<td>all/ US politicians and bots</td>
<td>human=502, bots=928</td>
</tr>
<tr>
<td>New bots</td>
<td>all/ US politicians and bots/ German politicians and bots</td>
<td>bots=699, human=1462</td>
</tr>
<tr>
<td>German bots</td>
<td>all/ German politicians and German bots</td>
<td>bots=27</td>
</tr>
<tr>
<td>Varol</td>
<td>all/ Varol et al.</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Data sets and how we combined them for our analysis
Figure 1: Example accounts for each data set
in our sample (n=4,583) we requested the user scores via the Botometer API giving us a total of 374,724 valid scores. Besides an empty timeline (Botometer needs at least one tweet – 2,021 calls had to be excluded), missing accounts (mainly from Varol’s data set which includes accounts that do not exist anymore – 11,197 calls) and private accounts (also mainly from Varol’s data set – 33,608 calls had to be excluded), only 86 calls (less than 0.05% of all calls) returned an error because of a technical API issue from Botometer or the Twitter API. Eventually we could analyze 4,134 existing accounts that could be accessed at least once during the process of data collection and have at least one tweet on their timeline (see Table 1:). For the following analysis we use the language-independent universal Botometer score between 0 and 1 (which corresponds to the score 0-5 also used by Botometer – e.g., 0.5 will be 2.5), as well as Botometer’s newly integrated complete automation probability (CAP) (ranging between 0 and 1). While we are aware that the Botometer creators offer more fine-grained feature scores besides the overall universal scores, we focus on the universal overall scores as they are the scores used by social scientists in their studies and are independent from the language that the accounts tweet in.

For RQ1 we compared the ROC-AUC (Receiver Operating Characteristics-Area Under The Curve) with the old ROC-AUC reported in older studies for Botometer to evaluate the diagnostic ability of Botometer. ROC-AUC indicates how well a model discriminates between two classes (bot and no bot). The true positive rate on the y-axis is plotted against the false positive rate on the x-axis. While this statistic is useful to directly compare the overall performance of different models and labelled data sets with regard to the accuracy of the classifications, this evaluation method also has its limitations (J. Davis & Goadrich, 2006; Halligan, Altman, & Mallett, 2015). Researchers, for example, are often not interested in all potential score thresholds as used in the AUC-ROC approach and instead only in specific classification values (Lobo, Jiménez-Valverde, & Real, 2008). The AUC-ROC approach does not consider the cost of false-positive and false-negatives and does not fully consider imbalances in the general population (Saito & Rehmsmeier, 2015). However, in their original paper for the first version of the tool (BotOrNot) the authors transparently warn that the accuracy of their model (they received a ROC-AUC of 0.95) overestimates future performance due to the age of the training data (C. A. Davis et al., 2016).

As the ROC-AUC approach has limitations, we also use a different approach for RQ2. In bot research the biggest problem is the lower occurrence of bots in relation to humans. In the general population of all Twitter users
there are more real users than bots. However, the ROC-AUC approach does not consider the baseline in the Twitter population. This is a general problem with classifiers if the structure of the population in which the classifier should be later used is ignored. In the case of bots, we want to clearly identify bots and make statements about the number of bots in a data set. This becomes even clearer when the main goal as well as the wording in bot studies is considered. Researchers are interested in the number of bots and not mainly in the number of human users. Therefore, even a small false-positive rate (FPR) has a large impact on the results. The data used to test (as in our case) and train Botometer is often more balanced than the rather imbalanced general Twitter population. Saito and Rehmsmeier (2015) advocate in such cases to evaluate the performance of a classifier based on the Precision-Recall (PR) curves as they “allow for an accurate and intuitive interpretation of practical classifier performance”. To simulate a population that represents the same ratio as in the general Twitter population we resampled the different data sets for RQ2.¹

4 Results

Our analysis (see Figure 2 and Table 2) shows that the ROC-AUC is worse with our complete data (AUC=0.85) than with the data sets used in the original papers of the Botometer creators which received an AUC of 0.94 (Varol et al., 2017). However, their baseline model received also only an AUC of 0.85 in the original paper. The nearer the ROC curve (Figure 2) is to the upper-left corner and the more space is under the curve, the better Botometer can distinguish between bots and humans. In the lower-left corner the curve starts with a threshold of 0 for the Botometer which means the false-positive rate is 0 but at the same time 0 of the bots in the population will be identified. The curve ends with a threshold of 1 in the upper-right corner as then all bots in the population will be identified. However, such a threshold also means that all human accounts will be wrongly classified as bots (false-positive rate = 1). Each point in the curve thus represents a specific threshold for which the true positive rate as well as the false positive rate is shown.

When applying this to our data sets, we see that the classifier has the highest AUC score for the US politicians and bots data set (0.93) followed by the Varol et al. (2017) labeled data set (0.86). The score for the German

¹The data and the code are available on Harvard Dataverse: https://doi.org/10.7910/DVN/XVCKRS
politicians and German bots yields a lower score (0.76). As we only had a few German bots we also tested the German politicians and bots (with the new bots instead of the German bots) which had a slightly higher AUC (0.78).

We, then, calculated the ROC-AUC for the CAP that indicates the level of automation (Figure 3 and Table 3). We could observe similar results as before for the universal score: the US politicians received the highest AUC (0.94) followed by Varol et al. (0.86), the complete data set (0.86), the German politicians with the new bots (0.78) and the German politicians and the German bots (0.77).

For RQ2 we suggest a thought experiment based on the data sets analyzed in this study. While the data sets are not perfectly balanced, they definitely do not represent the imbalanced ratio of bots and human accounts of the real Twitter population. We thus assume for our simulation that the Twitter population has 15% bots (Varol et al., 2017) and 85% non-bots. We first gave every single account a sampling probability that reflects the occurrence of cases in the Twitter population (n bots/0.15 and n humans/0.85) and took for every analyzed data set 100,000 random samples with replacement. As a result, we created new data sets that have 15,000 bots and 85,000
non-bots. We then calculated for all the data sets the ROC-AUC as well as the PR-AUC (see Table 2 and Table 3). As expected the ROC-AUC is for all analyzed data sets the same with the new sampled data sets which is also in line what Saito and Rehmsmeier (2015) show in their study. However, the PR-AUC shows a different score for the resampled data sets as the PR-AUC changes depending on the imbalance in a data set. The only exception is the in its original form highly skewed data set (German politicians and German bots) shows an improvement but overall still an extremely low score (from 0.10 to 0.27).

The PR curves allow us to better evaluate how many bots will be amongst the accounts that Botometer classified as bots, and how many bots were not identified if we use a high threshold (as many studies do with this tool; e.g., T. R. Keller & Klinger, 2019). PR curves start on the right with a Botometer score threshold of 0. If an extremely low threshold is used, all bots in the population will be identified (recall = 1) but the number of false-positive cases will be extremely high. The precision is thus .15 as 85% of the accounts in the sample are humans. On the left, the curve ends with an extremely high threshold for the Botometer score which means that the precision is extremely high (the as bots classified accounts are all bots) but the recall is extremely low (almost all bots in the general population are not identified). We can now also check the precision and recall for specific thresholds. For example, if 0.76 is used as a threshold for the resampled all data set, 59% real bots (precision = 0.59) and 41% false-positive human will be amongst the accounts classified as bots (see Figure 4). However, around 80% (recall = 0.2) of the bots in the population will not be identified (false-negatives). If we consider the German politician and bots data set, the results are even worse. Of the accounts classified as “bot” only around 24% (precision=0.24) would be bots and 76% of the accounts would be false-positive human accounts. Moreover, around 90% (recall=0.1) of the

<table>
<thead>
<tr>
<th>Data set</th>
<th>ROC-AUC</th>
<th>ROC-AUC sample</th>
<th>PR-AUC</th>
<th>PR-AUC sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.85 (0.84-0.86)</td>
<td>0.85 (0.85-0.86)</td>
<td>0.77 (0.75-0.79)</td>
<td>0.50 (0.49-0.51)</td>
</tr>
<tr>
<td>German politicians and German bots</td>
<td>0.76 (0.70-0.83)</td>
<td>0.76 (0.76-0.77)</td>
<td>0.10 (0.07-0.14)</td>
<td>0.27 (0.27-0.27)</td>
</tr>
<tr>
<td>German politicians and bots</td>
<td>0.77 (0.74-0.79)</td>
<td>0.76 (0.76-0.77)</td>
<td>0.81 (0.78-0.84)</td>
<td>0.30 (0.30-0.31)</td>
</tr>
<tr>
<td>US politicians and bots</td>
<td>0.93 (0.92-0.94)</td>
<td>0.93 (0.93-0.93)</td>
<td>0.96 (0.95-0.97)</td>
<td>0.79 (0.79-0.80)</td>
</tr>
<tr>
<td>Varol et al.</td>
<td>0.86 (0.84-0.88)</td>
<td>0.86 (0.86-0.86)</td>
<td>0.76 (0.73-0.79)</td>
<td>0.58 (0.58-0.59)</td>
</tr>
</tbody>
</table>

Table 2: ROC-AUC as well as the PR-AUC scores for the original data sets as well as the weighted resampled data sets (sample = 100,000) for the universal score. The 95% confidence intervals based on 10,000 stratified bootstrap replicates are shown in brackets.
Figure 3: Receiver Operating Characteristics curve for Botometer and the complete automation probability (CAP) (average over 3 months for each account).

bots in the Twitter population would not be identified (see Figure 4).

We then also calculated the PR-AUC for the CAP with the newly weighted resamples (see Table tab:my-table-3). With regard to the PR-AUC and the ROC-AUC, the results are almost the same as for the universal Botometer scores. When calculating the CAP and using a threshold of 0.25 – as used by Zhang et al. (2019), the accounts, classified as bots by Botometer, would consist of around 55% bots (precision=0.55) and 45% humans. However, around 71% (recall=.29) of the bots in the population would not be iden-

<table>
<thead>
<tr>
<th>Data set</th>
<th>ROC-AUC</th>
<th>ROC-AUC sample</th>
<th>PR-AUC</th>
<th>PR-AUC sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.86 (0.85-0.87)</td>
<td>0.86 (0.86-0.86)</td>
<td>0.78 (0.76-0.80)</td>
<td>0.51 (0.50-0.52)</td>
</tr>
<tr>
<td>German politicians and bots</td>
<td>0.77 (0.70-0.83)</td>
<td>0.77 (0.76-0.77)</td>
<td>0.10 (0.07-0.14)</td>
<td>0.27 (0.27-0.27)</td>
</tr>
<tr>
<td>German politicians and German bots</td>
<td>0.78 (0.75-0.81)</td>
<td>0.78 (0.77-0.78)</td>
<td>0.81 (0.79-0.84)</td>
<td>0.31 (0.30-0.31)</td>
</tr>
<tr>
<td>US politicians and bots</td>
<td>0.94 (0.93-0.95)</td>
<td>0.94 (0.94-0.94)</td>
<td>0.97 (0.96-0.98)</td>
<td>0.81 (0.81-0.82)</td>
</tr>
<tr>
<td>Varol et al.</td>
<td>0.86 (0.84-0.88)</td>
<td>0.86 (0.86-0.87)</td>
<td>0.76 (0.73-0.79)</td>
<td>0.58 (0.57-0.59)</td>
</tr>
</tbody>
</table>

Table 3: ROC-AUC as well as the PR-AUC scores for the original data sets as well as the weighted resampled data sets (sample = 100,000) for the universal CAP. The 95% confidence intervals based on 10,000 stratified bootstrap replicates are shown in brackets.
Figure 4: Precision-recall curves for the resampled data sets considering the population baseline on Twitter (15% bots) for the universal Botometer score, black points indicate the precision and the recall for the Botometer score 0.76. With the German politicians and bots data set, for almost every threshold level the identified sample of bots has more humans than real bots (precision).

In our third research question we are interested in the difference between languages. A comparison between the different data sets shows that the tool works better with data from the US/English (see Figure 2, Figure 3, Figure 4, and Figure 5). However, Botometer is neither working well with the German politicians nor the new English-language bots we tested. We further tested with unpaired DeLong tests (DeLong, DeLong, & Clarke-Pearson, 1988; Sun & Xu, 2014) whether the ROC-AUC of the German data sets are
Figure 5: Precision-recall curves for the resampled data sets considering the population baseline on Twitter (15% bots) for the CAP Botometer score. Black points indicate the precision and the recall for the CAP 0.25. With the German politicians and bots for almost every threshold level the identified sample of bots has more humans than real bots (precision).

statistically different from other data sets. Our test shows that the ROC-AUC for the German data sets are significantly different from the ROC-AUC of the US politicians and bots with regard to the universal score (German politicians and bots: $D = -10.58$, $df = 2133.1$, $p < 0.001$; German politicians and German bots: $D = -4.67$, $df = 586.12$, $p < 0.001$; Bonferroni correction $\alpha = 0.005$) as well as the CAP (German politicians and bots: $D = -4.81$, $df = 579.12$, $p < 0.001$; German politicians and German bots: $D = -10.65$, $df = 2066.3$, $p < 0.001$; Bonferroni correction $\alpha = 0.005$; see also Table 2 and 3). Furthermore, if Botometer is used with German accounts, the real rate of false-positives is higher than researchers might assume when their only indicator for performance is the ROC-AUC. While the ROC-AUC approach is a good indicator for the original study and data set, researchers should be cautious to assume the same level of discriminatory power can be reached with a new data set from a different account population. Additionally, researchers should consider the absolute number of false-positives in the classified sample and report the precision as well as the recall for the threshold used in a
In RQ4 we are interested in the temporal volatility of Botometer scores. We identified substantive changes in the bot scores on an individual level. We calculated the standard deviation of the universal score as well as the CAP for every single account and then plotted the density for every individual data set. Our analysis shows that especially the bot datasets show high standard deviations for individual accounts and thus are rather volatile over time (Figure 6). The only noteworthy difference between the universal score and the CAP is that the CAP overall seems to be slightly less volatile overall.

Lastly, we also checked how many accounts in each data set are classified differently during the three months. The volatility alone is only a problem if single accounts are depending on the day below or above a chosen threshold used by researchers to classify bots. If we chose a threshold of 0.76, 27.2% of the new bots, 22.2% of the German bots and 13.9% of the Varol bots have at least once a daily measurement above and below the threshold. For the humans the volatility is less severe with regard to the classification based on a high threshold. 7.4% of the German politicians, 3.1% of the Varol bots and only 0.6% of the US politicians have at least once a daily measurement above and below the threshold.

We also calculated the change of classification for a threshold of 0.25 for the CAP. 37.5% of the new bots, 33.3% of the German bots, 10.7% of the German politicians as well as 1% of the US politicians have at least once a
Figure 7: Percentage of accounts (y-axis) that have at least once a score in the three months below as well as above the threshold for all thresholds between 0 and 1 in steps of 0.05. Left for the universal score, right for the CAP.

score in the three months below as well as above the threshold. Of the Varol bot data set 17.7% and of the Varol human data set 4.7% are at least once a bot or a human.

We additionally calculated the percentage of accounts that have at least once a score in the three months below as well as above the threshold for all thresholds between 0 and 1 in steps of 0.05 (see Figure 7. Our analysis shows that overall bots are more prone to have at least once a score below and above the threshold for both the CAP and the universal score. However, the CAP, works better for the new bots as well as the German bots.

5 Discussion

In a time where terms like “computational propaganda” (Bradshaw, Howard, Kollanyi, & Neudert, 2019), “disinformation campaigns” or “orchestrated in-authentic behavior” are often used, identifying the automated accounts that spread disinformation or attempt to shift the conversation seems paramount. Indeed, being able to measure the number of inauthentic behaviors within a given online discourse and understanding their patterns could lead to the identification of coordinated disinformation campaigns. Consequently, there was and still is an academic demand for tools like Botometer that did not only come with an API but, more importantly, also with the necessary aca-
demic credentials (i.e. institution and citations). Coming from computer science, Botometer was and is a very good example of how machine learning models can be trained and what the potential outcome might be.

Yet, as we show in this study, social scientists should be cautious with tools created and tested in computer science. A high ROC-AUC score with the original dataset and research setting does not translate into a good “precision” for computational social science studies investigating a different context. Even a good diagnostic ability leads to problematic outcomes when the population baseline is considered. In other words: just because the model works with the test data set, does not mean that the results are reliable for a specific use case. As social scientists, we are very rigorous with the quantitative as well as qualitative methods that we use. We care about reliability, validity, and reproducibility. In many of the social science studies that make use of Botometer, these criteria have been disregarded. As we have shown with our thought experiment, Botometer has problems in identifying false positives and false negatives in our collections when applying approaches from social sciences; especially with thresholds used in prior studies. This is noteworthy, as our data sets consisted either of accounts that Botometer was directly trained on or where the cases were clear-cut and easily identifiable for a human coder. Furthermore, we show that the scores change over time, thus, effectively, making studies that use Botometer data from one day or a week hard to replicate. Botometer worked, unsurprisingly, better on accounts that the algorithm was trained on even in our analysis over time. However, the tool was imprecise over time for the new bots as well as the German dataset. In addition, bots that are not aimed at entertaining other users but rather aim at obfuscating their automated identity and real goals (e.g., spreading misinformation) will pretend to be humans, thus making it even harder for bot detection tools to identify them. In this sense, distinguishing clearly between humans as well as bots is the minimum that bot detection tools should accomplish. Indeed, as practitioners we should be careful when relying on tools that have been created by scientists from other fields and, as our thought experiment highlighted, also have to be mindful of the population we apply the classifiers on. If the numbers of bots are in the minority for the general population (as we assume on Twitter), even a small false-positive rate might lead to a high number of human accounts wrongly classified as bots.

This, then, raises the question if social scientist should make use of Botometer at all. And while an argument in favor of Botometer would be that, while not perfect, it at least is an approximation towards identifying bots on Twitter, our counter-argument would be that the issue of false pos-
Figure 8: Left: Density plots for the different combined data sets in our analysis showing the distribution of Botometer’s universal score. We used the resampled data sets with 15% bots and 85% humans with a total n=100,000 for each data set. Right: Density plots for the human accounts data sets. Lines indicate the median, a bandwidth of 0.04 was used for all data sets.

Negative and false negative rates for specific thresholds with datasets based on the general Twitter population as well as the tool’s vulnerability to temporal changes or different languages do call the tool’s classifications into question. A solution researchers might come up with would be to shift the focus from thresholds to the overall distribution of the raw scores to evaluate the prevalence of bots in a data set. However, based on our analysis we would not recommend such an approach as different data sets with the same human to bot ratio as well as our human data sets can look differently in density plots (see Figure 8).

However, we also have to add as a limitation that we tested the tool with a few specific data sets. We have chosen data sets that make it easy for us to evaluate the ground truth and we still expect very similar results with future analyses. Furthermore, we have used a data set of bots that includes some of the most obvious bots. If researchers manually validate the tool again with their own unique data analysis, they of course should make use of the tool. Our analysis also shows what Saito and Rehmsmeier (2015) recommend in their study: the PR-curve is more informative if we are coping with imbalanced data in the population.

With that having said, we want to emphasize that we do not think that Botometer is a bad tool in general; we rather suggest that researchers should be aware of Botometer’s limitations when using the tool and care about
validation with their own labeled data sets like Fernquist et al. (2018) did in their study. We do think that Botometer is a great tool in the context of computer science. Journalists as well as citizens can also make use of the tool to investigate suspicious accounts. But Botometer, in this context, would only be one step of a multi-step process that is built around human investigation of individual accounts. We believe that for social scientists as well as computer scientists working at scale, the uncertainty is too high. If social scientists and internet researchers want to use Botometer for their datasets or even develop their own new classifiers we suggest certain steps going forward:

1. Manual classification: They should manually analyze a substantial sample of the classified bot and human accounts that were classified by the machine learning system. Researchers could then calculate the intercoder-reliability between the machine learning system’s classification and the human coder classification. While the percentage agreement is already an interesting indicator, as communication scientists we would highly recommend the use of a measurement like Krippendorff’s alpha (Hayes & Krippendorff, 2007) that takes into account chance agreement. They should also always manually validate, if possible, all the accounts that were classified as bots by the machine learning system if researchers want to make statements about the number of bots in a population.

2. Time: Researchers should validate their results over time. As we’ve shown with Botometer, the scores are prone to changes over time. Consequently, researchers using Botometer or developing a new machine learning classifier should test their data set over time and look at the changes over time.

3. Language: Researchers should be aware of Botometer’s limitations (that are even stated by the creators of the tool) when it comes to classifying accounts that tweet in other languages. Hence, scholars will have to consider using other tools to checking the amount of bots in their data set; e.g., manual coding. Future machine learning systems should also consider different languages.

4. Data sharing: Researchers analyzing bots should share the user ids of the classified “bots” as the Twitter developer terms allows researchers to share the id of users. While the creators of Botometer share their
lists of bots on their official homepage, social scientists using Botometer or similar machine learning systems in their studies often do not even share the ids of the accounts that got classified as bots.

5. Improvements: Going forward, researchers should support tools that are not based on black boxes but rather commit to the idea of open science. We should be able to work together on finding solutions and improving those solutions over time. This, then, would also allow us to deal more transparently with potential weaknesses.

Finally, there are other ways to identify bots on Twitter. Bastos and Mercea (2019), for example, have investigated suspicious patterns based on communication metrics as well as network analytics. And while this method might be more complicated and computationally expensive, it offers social scientists a way forward in identifying bots and understanding patterns of disinformation campaigns. Overall it is also important to broaden the perspective and focus more on political digital astroturfing (F. B. Keller, Schoch, Stier, & Yang, 2019) for which bots are just one potential dimension (Kovic, Rauchfleisch, Sele, & Caspar, 2018).

References


Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), e0118432. doi: 10.1371/journal.pone.0118432


61. doi: 10.1002/hbe2.115