#BDST: BOT DETECTION SUPPORT TOOL
FOR HUMAN ANNOTATORS OF TWITTER DATA

S. Barberi, F. Grisolia, S. Tardelli, M. Tesconi

IIT TR-03/2019

Technical Report

Maggio 2019
#BDST: Bot Detection Support Tool
for human annotators of Twitter data

S. Barberi, F. Grisolia, S. Tardelli, M. Tesconi

Contents

1 Introduction 2
  1.1 Why is social media a double edged sword? . . . . . . . . . . 2
  1.2 What are social bots? . . . . . . . . . . . . . . . . . . . . . . . . 3

2 Objectives and Methodology 5
  2.1 Aim of the project . . . . . . . . . . . . . . . . . . . . . . . . . . 5
  2.2 What is the usefulness of human annotation? . . . . . . . . . 6
  2.3 How does the web application work? . . . . . . . . . . . . . . . 7
  2.4 Structure of the web application’s interface . . . . . . . . . . . 9

3 Results 9
  3.1 Human accounts: verified and not verified . . . . . . . . . . . 9
    3.1.1 Verified account . . . . . . . . . . . . . . . . . . . . . . 9
    3.1.2 Not verified account . . . . . . . . . . . . . . . . . . . . . 10
  3.2 Bot accounts . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
    3.2.1 First bot account: tweeting content . . . . . . . . . . . . 16
    3.2.2 Second bot account: tweeting frequency . . . . . . . . . 19

4 Conclusion and future development 22

Abstract

The human annotation of social bots is an essential task for the training of new algorithms and bot detection techniques. However, identifying bot users on social media is tricky and error-prone, even for expert annotators. Furthermore, this task is often time-consuming and the results are modest. The Bot Detection Support
Tool (#BDST) is a tool developed in order to make it possible for a human annotator to better understand the behaviour of a Twitter account through intuitive visualizations of the user’s activity. Results show a usable and effective tool, with great potential for future research.

ACM Classification.
H.5 Information Interfaces and Presentation

1 Introduction

1.1 Why is social media a double edged sword?
Social media environments such as Twitter represent an extension of the concept of public sphere, traditionally constructed by Habermas [1] as the breeding ground of public opinion, a virtual space where the collective consciousness is expressed and challenged, and where the community sentiment is shaped through the mass media. In this regard, it can be argued that social media platforms have helped merging purely national public spheres into global networks where information is shared and consumed on a large scale [1] and where every citizen has access to the public debate. An example of how social media has been successful in bringing political change is its role in the Arab Spring, a series of uprisings that flowed into a Democratization movement joined by several Muslim countries, including Tunisia, Egypt and Libya. In this case, social media has been fundamental in building, organizing and coordinating political action both online and in the streets [2], being treated by users as a public information infrastructure where they could pursue civic engagement and exercise freedom of speech [3].

Twitter, in particular, has shifted from being a “personal microblogging site to a news and information dissemination platform” [4] and now serves as an online ecosystem where social, political and economic discourse is constructed and negotiated by participants through content publishing techniques such as tweets, retweets and replies. Twitter does not only provide a platform where content is generated by users: it also allows for information to be extracted from the user generated content in order to be used to predict real-life phenomena in areas such as finance [5, 6] or fashion [7].

In such a virtual space, where access is granted to everyone, participants can share a type of content which might not be beneficial nor useful for the
social media public. An example is the spread of fake news and the attempt to manipulate the public opinion: as reported in [8], false information on Twitter is retweeted more rapidly and by a larger amount of users compared to true information, especially when the topic has a political nature. As a matter of fact, manipulation on social media through the spreading of political misinformation has been studied and analyzed for more than a decade suggesting the possibility of fake news consumption being able to influence, control and therefore manage the public opinion [9, 10]. In addition, there is a spread of fake accounts on Twitter, used to inflate the followers count of a specific account: this phenomenon can be dangerous when accounts buy fake followers in order to gain more influence in different domains [11], from the social to the political and economic sphere. Consequently, it can be argued that social media is being exploited to manipulate online discussions, and that there is a need to control, regulate and manage the proliferation of fake news online. On this account, some social media platforms, including Facebook and Twitter, signed up in September 2018 to a voluntary Code of Practice on Disinformation for online platforms, commissioned by the European Union. Signing this Code of Practice means taking action in different aspects; for example it compels platforms to empower consumers to report misinformation and to address the issue of fake accounts and online bots.

1.2 What are social bots?

One of the most powerful means of manipulation on social media, in particular on Twitter, is through the usage of social bots. Bots, which is short for software robots, are automated agents which interact with other users within a system, with the intention to emulate and possibly reshape and manipulate their behavior [12]. When the system is a social media platform, they are specifically called social bots. They are created for a number of different reasons. On Twitter, social bots can generate informative tweets which can enrich the user’s experience on the platform and help with some tasks. An example is a bot called @WhatTheFare: it estimates a user’s Uber fare based on a user’s tweet containing the pick up and drop off locations. @WhatTheFare is only one of the many examples of benevolent Twitter social bots. Just as users can interact with social bots which are created to benefit their experience, they can also encounter social bots which are programmed to disrupt the user’s experience on social

media through various approaches.

In the political election sphere, it is advantageous to refer to these social bots as political bots, which are designed to, for example, follow particular politicians in order for them to be perceived as more popular, or to spread propaganda in support or against a certain topic or political party. Doing so enables political bots to polarize the political conversation and endangers democracy [13]. For instance, as detected in [12], roughly 3.8 million tweets, which represented one-fifth of the conversation surrounding the online political discussion about the 2016 US Presidential election, were not produced by human-managed account but by bot-managed accounts. The danger of bot activity on Twitter is not confined to political manipulation: for example, social bots are being exploited by spammers to spread spam and malicious content, following a great amount of users in hopes of being followed back by a good amount. The main objective of their phishing activity is to target users through a keyword search, and subsequently attract them through captivating tweet content, which persuades users to click on the links that social bots have embed in the tweet. When users click on the links, which are often shortened through an URL shortener service such as Bitly to make the link’s content illegible, they are redirected to malicious sites or spam content [14].

As researched in [5], social bots spreading spam material are also present in stock microblogs. As mentioned before, content generated on platforms such as Twitter can be used to predict real life phenomena [15]: such pre-
<table>
<thead>
<tr>
<th>TYPE OF ATTACK</th>
<th>USED APPROACH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abusive usage of topics</td>
<td>A topic can be hijacked (changing its initial meaning to a new one) or censored (destroying its meaning)</td>
</tr>
<tr>
<td>Unsolicited communication</td>
<td>Communicating in an unsolicited way, e.g. tricking users into clicking on links embedded in an unobtrusive context (clickjacking)</td>
</tr>
<tr>
<td>Spoofing</td>
<td>Impersonating a specific user</td>
</tr>
<tr>
<td>Setting traps</td>
<td>Using bot accounts as honeypots to attract users to malicious websites</td>
</tr>
</tbody>
</table>

Table 1: Social bots attacks

dictions become inaccurate if not completely misfigured when the content they are based on is fake and possibly purposely created to mislead algorithms and users alike.
In addition, social bots can produce content with the intention to create panic during emergencies or to damage personal or corporate reputation [4]. Malicious bot activity on Twitter has been categorized by [16] according to the type of attack perpetuated.

2 Objectives and Methodology

2.1 Aim of the project

After discussing the danger posed by the presence of social bots on Twitter, it is clear that an intervention of identification and elimination of all those automated accounts which are jeopardizing user’s activities on the platform is required. This particular project aims at providing a tool for the recognition and identification of social bots on Twitter, through a web application that offers an overview of different features that have been linked to bot activity through various researches and studies. The web application can be used by human annotators to analyze and label those accounts which seem to present attributes correlated to bot activity and to identify them as automated malicious accounts. Following this task, the analyzed accounts can be collected and incorporated into a dataset of automated accounts, which can be used as training data for a bot detection algorithm.
2.2 What is the usefulness of human annotation?

To make bot detection simpler, researchers from the Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS) at Indiana University have developed a tool called Botometer\(^2\) which lets the user insert the screen name of a Twitter account in order for the tool to check the account’s activity and give it a score based on its likeliness to be or not to be an automated account. Although Botomter is a great tool to detect online bots, its accuracy is not perfect yet and it is still a work in progress. The project’s objective is to create a tool to complement bot detection using human annotation rather than machine learning algorithms. As asserted by Appen, an Australian company which specializes in producing human-annotated training data for machine learning, “humans are simply better than computers at managing subjectivity, understanding intent, and coping with ambiguity”\(^3\), all of which is an essential ability when identifying automated accounts which might possess human-like features. As a matter of fact, modern social bots are becoming more and more sophisticated, making the “boundary between humanlike and bot-like behavior [...] fuzzier” [12], with automated accounts incorporating human-like features, from tweeting timing to tweeting content.

\(^2\)Botometer by OSoMe, https://botometer.iuni.iu.edu/#/.
2.3 How does the web application work?

The web application developed for this project lets the user select the Twitter account they wish to analyze, insert the account’s screen name into the application’s form, in order for the application to display the most important features to be used for the labeling task. As figure 2 shows the most important features to be considered have been selected through a careful literature review of existing studies. The following features are characteristics of bot accounts:

- Lack of intelligent and/or original content: an example might be an account which only retweets other accounts’ content, or tweets and replies that are very similar to each other, if not duplicates [14];

- An abundant presence of URLs in tweets: links might redirect users to spam or malicious content [17];

- Content posted regularly and at small and fixed intervals [18];

- Tweets containing words and hashtags such as offer, free, click, prize, debt, deal, credit, and sex, which suggest the presence of spam or malicious content [19];

- Accounts that are not verified [20] or geolocated, and tweets are not protected;

- Lack of profile picture or biography [21];
• Usage of automated tools for tweeting [20];

• Aggressive following behaviour, following and unfollowing other users in a short amount of time [14];

• Very high or very low followers/friends ratio [17].

Once the features had been selected, the following step was to design a system to download the necessary data using the Twitter API\(^4\). It is possible to interact with Twitter’s data using its API, logging oneself in through the API credentials: once one makes a request to Twitter, it returns a result in JSON format. To download the account’s personal information, such as the name or followers count, the Tweepy\(^5\) Python library has been used. The library provides the class tweepy-api, which supplies a wrapper for the Twitter API. The data, which is contained in the Twitter User Object\(^6\), is then transferred dynamically into the web application. To download the account’s timeline, containing the account’s tweets, the data is crawled directly through the Twitter API using a PHP script. The accounts’ timelines are stored in a MySQL database, which is composed of three tables:

• users (contains information about accounts)

• entities (contains entities such as hashtags and URLs incorporated in the users’ tweets)

• tweets (contains information about users’ tweets)

In order to use the MySQL database and the scripting language PHP, the XAMPP\(^7\) development environment has been adopted, which also provided the Apache HTTP server. Apart from the script used to download the User Object data, which has been developed in Python, the other scripts used to retrieve data from the Twitter API have been developed in PHP. In addition to PHP and Python, the HTML and CSS markup languages have been used to develop the web application’s interface, together with the Bootstrap framework. The DOM elements of the HTML web page are dynamically managed through the usage of JQuery, a Javascript library for DOM elements manipulation. The Twitter data is transformed and normalized using the Javascript programming language. Furthermore, AJAX technology has been used for the dynamic update of the web page; as a matter of fact, AJAX allows to update web pages by exchanging data with a web server without the need for the user to reload the web page.

\(^4\) developer.twitter.com/en/docs/tweets/search/api-reference.html

\(^5\) www.tweepy.org

\(^6\) developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.html

\(^7\) www.apachefriends.org/
2.4 Structure of the web application’s interface

The web application is divided into four main sections:

- account
- tweeting activity
- tweeting content and sources
- tweeting frequency

The account’s personal information, which even alone can be indicative of bot or human behaviour, is displayed in the first section of the web application. The following section displays the timeline, twenty of the most recent replies (if present), and some tweeting details. The third section contains a word cloud showing the most used hashtags, and a pie chart displaying the used tweeting sources. The final section contains a heat map showing the tweeting frequency.

3 Results

3.1 Human accounts: verified and not verified

3.1.1 Verified account

It can be argued that the official Twitter account of the Italian politician Matteo Salvini is not managed by a social bot because the condition “verified” is true, as showed in figure 3.

As we further investigate, we can see in figure 4 that the published content, such as the account’s replies, could not be possibly produced by an automated agent because the lexicon is diverse and the content in general has a human-like originality.
As displayed in figure 5, the tweeting sources are not automated, being mainly Twitter for iPhone and Twitter Web Client, as well as the most used hashtags are not a sign of bot-created content.

Finally, the tweeting frequency, which is described in figure 6 in a heatmap that takes the days of the week on the y-axis, and the day hours on the x-axis, shows a human-like behaviour with a lower density of published tweets in the hours generally associated with sleeping (1am - 7am).

3.1.2 Not verified account

Following is an example of a not verified account taken from a dataset containing only human accounts.

The user has a followers/friends ratio close to 1, which is common for human managed accounts, where the ratio is usually not too high and not too low. In addition, the presence of a profile picture and a biography are additional signs of human behaviour (figure 7).

Moreover, the content of the user’s replies is original and could not be produced by a bot. The tweets/age of account ratio is relatively high for a human account, with an average of 14 tweets produced per day. On the other hand, the tweets with URLs are 25% of the total produced tweets, which is not particularly high (figure 8).
Figure 4: @matteosalvinimi - Section 2
Figure 5: @matteosalvinimi - Section 3

Figure 6: @matteosalvinimi - Section 4
As for the tweeting content, the most used hashtags do not present any of the words and expressions linked to bot-produced content. Also, the sources are mostly Twitter Web Client, Twitter for Android and Twitter Web App, which are not automated tools (figure 9).

The tweeting frequency heat map shows human behaviour, with a low activity during the sleeping hours, and a higher activity during the day, in particular in the evening (figure 10).
Figure 8: Human account - Section "Tweeting Activity"
Figure 9: Human account - Section “Tweeting Content and Sources”

Figure 10: Human account - Section “Tweeting Frequency”
3.2 Bot accounts

The analyzed bot accounts have been selected from a dataset containing only social bots.

3.2.1 First bot account: tweeting content

In this example, the account section suggests bot behaviour for three reasons: a fake name, lack of biography, a low followers/friends ratio (figure 11).

An important indicator of bot behaviour also lies in the content of replies: the user mostly replies adopting the same hashtag. In addition to that, the account’s activity alternates tweets composed by a single hashtag, just like the replies, and retweets.

Another anomaly is offered by the tweeting frequency heatmap, which suggests that the bot might have been programmed to tweet only in a certain time slot (figure 14).
Figure 12: Bot account 1 - Section “Tweeting Activity”
Figure 13: Bot account 1 - Section “Tweeting Content and Sources”

Figure 14: Bot account 1 - Section “Tweeting Frequency”
3.2.2 Second bot account: tweeting frequency

The second example shows a bot account which has a human-like follower/s/friends ratio, but a biography with a randomly produced content, probably automatically generated (figure 15). As the tweeting information suggests, all of the account’s tweets are actually retweets, and there are no replies (figure 16). Another red flag is represented by the tweets/age of account ratio, which is very high: in fact, the account posts an average of 85 tweets per day. The complete absence of replies and the 100% of retweets is also indicative of bot behaviour (figure 16).

The tweeting content is particularly indicative of bot behaviour, as the user retweets only memes (figure 17).

Finally the tweeting frequency is very revealing too: the heatmap shows a bot-like behaviour, with content being tweeted all the time, without any break (figure 18).
Figure 16: Bot account 2 - Section “Tweeting Activity”
Figure 17: Bot account 2 - Section “Tweeting Content and Sources”

Figure 18: Bot account 2 - Section “Tweeting Frequency”
4 Conclusion and future development

The massive increase of social media usage can be perceived as an opportunity to develop a public sphere where a “democratization of discussions about policy, politics and social issues” [22] is in place. Nevertheless, in such an open environment, the risk of these discussions being manipulated and transfigured by automated agents such as social bots is very high. Social bots do not only try to exploit public thought, they can also share malicious content such as spam and malware. Therefore it is fundamental to enable users to identify and ask for removal of those accounts which are not human-managed.

The web application developed for this project has the purpose to specifically help annotators to identify social bots on Twitter. Once the annotator searches for the selected account, the web application presents four distinct sections. Each section contains different account’s characteristics aimed at helping with the identification.

The current project could be further developed adding different features. A helpful addition could be a fifth section in which a different area of bot detection is explored: the conversational network structure in which bots communicate with each other. As argued in [23], bot conversations are different from human conversations. Bots “tend to get involved in isolated conversations, and the followers of the bot are very loosely connected. The network created from a human virtual interaction on Twitter, is highly connected due to shared friendship, shared interests, and shared experiences in the real world” [23]. Social bot analysis built on the structure of their conversations could be helpful in detecting social botnets, i.e. networks of social bots.

Furthermore, a functional incorporation could be a machine learning tool such as a classifier to help the annotator with their final decision regarding the nature of the account.

Another interesting integration could be represented by the possibility for the annotator to manually create a dataset of bot or human accounts based on the accounts analyzed on the web application, connecting it to an external database.

Last but not least, a fundamental improvement for the tool would be to focus not only on traditional social bots, which exhibit explicitly different features compared to human accounts, but also on a new wave of social spambots, which “mimic the human behavior of [...] genuine users” [24]. The rise of these bots, which are hardly recognizable from human accounts, calls for a new type of analysis, based on collective behaviour of groups of accounts rather than on single features of individual accounts: according
to [24], this is possible through a new approach to online behaviour exploration, based on *digital DNA*. The idea of digital DNA derives from the analysis of biological DNA: similarly to the latter, a user’s specific behaviour is associated with a character; thus, a sequence of behaviours generates a string of characters, representing the digital DNA, which “encodes the user’s behavioral timeline” [25]. The usefulness of digital DNA in social media analysis has been already proven in [24], exposing that automated accounts share longer sequences of behaviours with each other, compared to human-managed accounts. In this regard, the developed tool would greatly benefit from the integration of an algorithm which calculates users’ digital DNA, giving the annotator the possibility to compare different accounts in terms of behaviour similarity, rather than focusing only on single features.

References


