Original Manuscript

Social Science Computer Review
1-20

DOI: 10.1177/0894439321034991
journals.sagepub.com/home/ssc


Automation on Twitter: Measuring the Effectiveness of Approaches to Bot Detection

Oliver Beatson[1], Rachel Gibson[1], Marta Cantijoch Cunill[1], and Mark Elliot[1]

Abstract

The effectiveness of approaches to bot detection varies, with real-time detection being almost impossible. As a result, this article argues that the general Twitter using public cannot be expected to judge which accounts are bots with certainty and therefore do not know to what extent they are being manipulated online. In this article, the challenge of detecting bots and fake accounts is demonstrated by constructing two distinct methods to bot detection. The first method takes a fixed criteria-based approach, by building on commonly cited identifiers for bots. The second method takes a more flexible, investigative approach in order to uncover bots involved in coordinated efforts to influence online debates. As well as profiling the specific mechanics of how each one operates, we argue that they can be compared against an evaluative framework that specifies a set of key criteria that bot detection methods should meet in order to perform. Here, we identify four key criteria on which these methods can be evaluated and then examine how they perform in terms of the key criteria of accuracy. The results of these methods are then compared and cross-checked against an existing and widely used bot detection service. The findings show that different bot detection methods can present significantly different results and that only confirmation from Twitter, through suspensions or announcements, can truly allow users to know whether an account is a bot or not. We argue that this development could have a significant effect on the level of trust that social media users have both in the information they receive through social media and also in the political process.

Keywords

Twitter, bots, social media, disinformation, political astroturfing, OSINT

Since the U.S. election in 2016 and the UK's decision to leave the European Union (EU) in the same year, the problem of computational propaganda has been central to discussions about social media and its effects on Western democracies. Computational propaganda (CP) is the use of data and autonomous agents across social media platforms to manipulate public opinion (Woolley & Howard,

[1] University of Manchester, United Kingdom

Corresponding Author:
Oliver Beatson, University of Manchester, Humanities Bridgeford Street, Manchester M13 9PL, United Kingdom.
Email: oliver.beatson@manchester.ac.uk

2016). During these high-profile events, CP was manifested in a number of ways, such as the spreading of disinformation, misinformation, and malinformation. The aim of CP is to try to subtly manipulate public opinion across the political landscape (Woolley & Howard, 2016), by stoking existing divisions and tensions and attempting to increase voter apathy (Lartey, 2018). Research has found that CP was used in the EU referendum of 2016 (Bastos & Farkas, 2019; Howard & Kolanyi, 2016), specifically with the use of bots and fake accounts that were used to spread disinformation and to try to amplify opposing sides of online political discussions.

Within this definition of CP, as stated in the previous paragraph, lays a number of techniques with which social media users are becoming more acutely aware, of which the aforementioned use of bots to amplify specific political messages is just one. Their usage feeds into a particularly fundamental aspect of CP, which is political astroturfing (PA). PA is best described as "... (a) campaign disguised as spontaneous, popular 'grassroots' behaviour that are in reality carried out by a single person or organisation" (Ratkiewicz et al., 2011, p. 1). Social bots, which are types of bots designed to mimic the conversational capabilities of a real person, are particularly useful for this aim as they can appear to be real social media users or they can be used to boost the apparent popularity of a social media post through likes and retweets. A distinctive component of PA is that the messages being amplified do not necessarily have to be instances of disinformation or misinformation but rather they often focus on a real message or individual with the aim of generating large scale support. Further understanding of the potential impact of PA on democracy is incredibly important, given that PA represents a form of political manipulation that could lead to a small unrepresentative group wielding undue influence on an election. There is also the possibility of foreign actors influencing the elections of another state.

Given the potential threat to democracy that PA represents, the ability of the public to be able to discern a real grassroots campaign on social media from an astroturf campaign is of great concern. One of the key methods for this is to be able to identify when social bots are influencing online conversations. However, the ability to be able to determine whether users are real and to what extent a conversation is organic is clouded as a result of an unclear approach to bot identification by Twitter and a lack of clear and consistent methods within academia and the media, both in terms of what constitutes a bot, an account that is wholly automated and differs from a fake account which relies on more human involvement and the best practices with which to identify them. As a result, social media users, and specifically users of Twitter, do not have the ability to determine instances of disinformation and attacks on democracies across the globe, a situation that can be exacerbated and exploited by actors seeking to benefit from the confusion and distrust that this situation can create.

The goal of this research is to attempt to demonstrate the difficulty for individual users and average members of the public in confirming the presence of bots and the way they can be exploited as part of an astroturf campaign but also the difficulty in confirming an account is a bot without relying on a ground truth data set. This will be achieved by devising two approaches for the detection of bots that are grounded in publicly available techniques and determined within the literature on social bots and automated accounts. These approaches, the first, a semi-automated method based on fixed criteria around the detection of bots, the second, a manual approach drawn from open source intelligence (OSINT), will then be compared with an existing automated approach in order to assess their effectiveness. The approaches will be tested on data drawn from Twitter and will be centered on the #FBPE (Follow Back Pro European) hashtag and the campaign for a second referendum on the United Kingdom's membership of the EU. The implication, however, is that the difficulty in confirming bots for the average user is, to some extent, irrelevant as there isn't a method available that can provide users with the ability to spot a bot in real time. As a result, users will continue to be blind to the manipulation that they are encountering. This potentially means that the issue of bots is greater than initially feared, as users will not be able to identify when they are manipulated in the

immediate term, but some level of confirmation of bot detection over the longer term will only encourage a greater level of skepticism and potentially disengagement with the political process.

The Growth of PA

PA provides an important starting point with which to attempt to uncover inorganic behavior on Twitter as activity on an event or message provides a larger target than a focus on individual accounts. In addition to this, it is the larger scale, coordinated campaigns that have the greatest impact as one bot in itself is unlikely to have a large impact on social media users, however, as part of a larger group of bots, there is a greater opportunity to influence. This section will seek to present the debate and the often-fluid understanding around PA as well as the myriad and contested methods used for the detection of instances of PA and bots. What becomes evident is the difficulty in assigning the status of bot to an account both in academic and journalistic investigations, which brings in to focus the difficulty that is faced by an average social media user.

The previously noted definition given by Ratkiewicz et al. (2011) regarding PA is a useful starting off point but research has gone further in narrowing down the definition. Henrie and Gilde (2019) develop this further by giving an informed overview of the development of astroturfing and its increasing prominence online. Here, there is an emphasis on five distinct factors that play into PA in a specifically digital arena; it should be on the Internet, politically initiated, manufactured, strategic, and deceptive.

The purpose of PA can be varied, from seeking to boost opposing positions in a political debate to promoting a rival candidate or an incumbent leader. However, one of the ways in which PA differs from other forms of political manipulation, such as disinformation, is with regard to the use of promoting a message or individual that is rooted in reality and fact. The difference between PA and disinformation is that disinformation is used with the purpose of misleading and disincentivising participation in democratic processes (Jackson, 2017). PA, meanwhile, seeks to forward the aims of a particular organization or state with the aim of presenting that aim as a widely held belief. While this does involve a degree of dishonesty in misleading the public, the information itself is not misleading but rather the nature of the organization presenting it is. In some respects, this makes PA more troubling and consequential for democracy, because it is not just the message that needs to be debunked, which is the case with disinformation and misinformation, but rather the actors carrying out the PA that need to be pinpointed.

As with all areas of CP, the specific role of PA has been of particular interest to researchers seeking to understand its effect on the political process. In their paper, Keller et al. (2019) used publicly available data released by Twitter to better understand how astroturfing campaigns operate and how they can be better identified in future work. The data focused on the 2012 South Korean Presidential election where it was found the South Korean National Intelligence Service were coordinating an astroturf campaign to boost the apparent popularity of the conservative candidate. Importantly, the authors of this article stress that bots are not the only means of coordinating an astroturf campaign as many such campaigns involve real people who operate fake accounts. As a result, studies which just focus on automation may miss important indicators of an astroturf campaign. This point is supported by the work undertaken by Bolsover (2019) on astroturfing in China. What was found in this context was that bots were not widely used across Chinese social media platforms but rather civil servants who work for the government were found to be using fake accounts to promote the actions of the Chinese government.

In certain political contexts, Bots have been shown to be an integral part of an astroturfing campaign on social media as they help with the fundamental aspect of the tactic, namely, amplification. Amplification is the use of bots and automated accounts to boost the retweet, share and like numbers of posts supportive of a particular position or candidate. Work by Arnaudo (2019) focused

on the role that astroturfing has had in Brazilian politics. During the 2014 presidential election campaign, an investigation uncovered evidence that one of the candidates opposing incumbent Dilma Rousseff had paid a private company to coordinate an astroturf campaign to make him appear more popular. In this instance, hashtags supportive of the challenging candidate, Aecio Neves, received a much higher level of popular support and engagement than those of Dilma Rousseff. It was also found that automated accounts had been retweeting content attacking Rousseff.

Although social media sites, such as Facebook and Twitter, are attempting to tackle disinformation, it is difficult to know what effect these changes will have (Ghosh & Scott, 2018). Twitter has a number of criteria in place to try to remove accounts that display bot behavior, such as hashtag spamming, automated retweets, automated favorites, and automated replies, however, it does not appear that these rules are consistently upheld (Marechal, 2016). Twitter have released data containing tweet and account information on accounts they confirmed were fake and were active during the 2016 U.S. Presidential Election (Twitter, 2018) in order to acknowledge the scale of the issue they faced in cleaning up the problem, something they have continued to do with their biannual transparency reports (Twitter, 2018). There has also been an ongoing attempt to try to remove fake accounts, however, the scale of the problem and the ability of fake accounts to adapt mean that the problem is ongoing. Twitter have also attempted to defend their practice of bot detection as well as trying to give some guidance on what they consider to be bot-like activity (Roth & Pickles, 2020). The problem, however, still remains that Twitter maintain a black box approach to bot detection and don't release detailed explanations for why certain accounts are suspended and others are not. While this could have a practical benefit in not giving too much away to those running such accounts, it leaves researchers and users confused as to what the threshold is for an account to be suspended other than being reported by numerous users (Siriwardane, 2020).

Since the public became more aware of bots in the wake of the U.S. presidential election and the EU referendum in the UK in 2016, the meaning and usage of the term "bot" has become commonly overused and misapplied. Work by Gorwa and Guilbeault (2018) has sought to highlight the different ways in which bots are defined and how this can have a knock-on effect when it comes to identifying them and then eradicating them. In their work, the authors found that much of the debate and the definitions around bots remain ill-defined and vague. This issue appears to be born out in much of the literature, with various characteristics associated with the term bot being contested.

Detection Methods for Identifying Bots

When attempting to identify bots, one of the primary issues is the difficulty in pinning down examples of coordinated activity, as there are no hard and fast rules for what constitutes a "real" user or a fake one, only guidelines and predictions (DFRLabs, 2017). Many automated bot accounts will often have certain features that make them possible to detect, such as lack of location information and a substantial difference between how many people they follow compared with followers and a structured or unusual tweet schedule (Baker, 2015). Many of these accounts will be created or start tweeting frequently at around the same time as an important political event (Baker, 2015) and will often operate in a network, regularly retweeting and replying to each other rather accounts outside of that network (Mittal & Kumaraguru, 2014). While there has been debate on how many tweets per day an account would have to submit for it to be considered a bot (Gallagher, 2017), work carried out by DRFLabs established a threshold of 72 tweets per day (2017). This means that an account would have to average 72 tweets per day since its creation date for it to qualify as an automated account.

The literature on bot detection falls, largely, into two distinct streams. The first stream is built on automation and the goal of identifying threats through automated methods of detection. This is an area that would encompasses approaches such as that devised by Botometer. The second stream is

more focused on a manual approach and the application of OSINT methods. The subsequent subsections will seek to outline these two streams.

Automated Approaches for Bot Detection

A number of machine learning techniques have been applied to the detection process in academic work that use some of the aforementioned rules as their basis. Bello et al. (2018) constructed a framework for the understanding of bot behavior by using knowledge of existing bots and then applying those criteria to their own bots and attempting to measure the impact that they had. One of the methods they used was an average mean sentiment analysis to measure the response of bots to real world events. Shao et al. (2017) focused more on the spread of fake news by bots and found that many targeted the mentions of famous and influential social media users.

The ability to garner a greater understanding of the behavior of bots has meant that it is possible to provide the public with the tools to be able to attempt bot detection themselves. Yang et al. (2019) give a breakdown of much of the literature that is available on this subject, giving particular attention to the need for machine learning techniques in the fight in uncovering bots. An important tool in this regard is Botometer, created by Indiana University, which allows a user to search for a username on their website, which then produces a score based on the likelihood of that username being a bot. Botometer has become the de facto tool for bot detection particularly in academic research. Using machine learning algorithms, the Botometer application programming interface (API) assesses a number of characteristics of a specified account in order to provide a comparison with known bot accounts. This process is detailed in a paper by Davis et al. (2016), which sets out how the algorithm focuses on certain key areas, such as the following:

- an accounts network, which focuses on the tweets of a specified account and how they are diffused;
- user information, such as account creation date and the language the account tweets in;
- user friends and how they interact with the specified account;
- temporal features which highlight the behavior of the account such as the potential for tweets scheduling and spam;
- content features which detect repetition; and
- sentiment analysis which explores the type of language published by a specified account.

These features used together give an indication as to whether an account is a bot or not, however, due to the scoring methods used by Botometer is it hard to say with absolute certainty whether an account is a bot, with Botometer urging caution in interpreting their bot scores. Criticism has arisen on the effectiveness of Botometer in academic research. Work undertaken by Rauchfleisch and Kaiser (2020) has shown that Botometer can be an unreliable source when it comes to identifying automation in Twitter accounts particularly when working across languages, which can lead to false positives and false negatives. The researchers also show that the wider academic field often employ differing interpretations of the scores given by Botometer and establish different thresholds for determining automation.

Investigative Approaches to Bot Detection: Use of OSINT Methods

An alternative to automation for detecting bots that has gained popularity, particularly among the nonacademic and journalistic community, relies on more manual and investigative methods to detect bots and PA. Using OSINT provides a manual and investigative approach to the detection of bots and PA. OSINT is defined as a method for the acquisition and the analysis of information, particularly information that is publicly available such as social media data (Ziolkowska, 2018). OSINT is

a catch-all term that encompasses many different investigative techniques. Most of the techniques centered on the cross-checking of similar types of accounts, particularly for identifying networks of accounts through similar follower lists. In terms of specific tactics, DFRLabs (2017) have produced a blog post that outlines the key areas to look out for when trying to assess whether an account is a bot. Many of these areas crossover with the characteristics highlighted by Botometer, specifically, areas around intensive activity, amplification, and common or repetitive content. These are all characteristics of accounts that would be put to use in a PA campaign in order to drive the apparent popularity of a message. Importantly, while potentially being more time-consuming, this approach gives the user/investigator of these accounts a greater feel and understanding of the activity of the suspect account or accounts, which will allow them to make a judgment based on their experiences of bot accounts and suspect behavior. Further investigative techniques involve image searching in order to detect stolen or commonly used photos, alphanumeric names, and commonly shared commercial content.

OSINT is also used in academic research, particularly in the realm of policing and intelligence services (Yeboah-Ofori & Brimicombe, 2018; Ziolkowska, 2018), however, there is relatively little in the way of published work relating to the use of OSINT on social media and in the detection of acts of PA and disinformation. But the relative ease with which information can be obtained through OSINT tools has led to an increase in the use of the practice being undertaken in journalism. Bellingcat, the online investigation website, has conducted a number of such investigations targeted at fake accounts in Twitter, including an attempt at astroturfing by Amazon (Toler, 2019).

The important part of OSINT that the method for doing so is not prescriptive. OSINT represents a number of techniques, with some being more useful than others based on the particular traits that the creator or manager of a fake account is trying to disguise. This is the greatest strength in the use of OSINT, in that it can be applied in various ways in the investigation of bots and astroturf campaigns. With such a high-profile bot searcher, such as Botometer, it is possible for an account to do just enough to create a level of uncertainty around its behavior which prevents it from being labeled definitively as a bot. However, despite the positives associated with using manual or OSINT methods to detect bot accounts, it remains true that ultimately, the decision to assign the tag of bot to an account rest with the researcher or investigator. This means that there is unlikely to be alignment across different studies and investigations meaning there could be an issue of reproducibility.

Classifying the Effectiveness of Bot Detection Methods

The methods included in this literature review go some way to delivering on different criteria or standards that bot detection methods require. These methods largely aim to deliver accurate results in terms of bot classification, although, there are other criteria that are prioritized over others, such as efficiency or simplicity. The implementation of a fixed classification provides a framework for this research with which to assess the effectiveness and quality of the bot detection methods used based on the results derived from previous research. The four criteria put forward are:

- accuracy,
- user friendliness,
- efficiency or time costs, and
- simplicity of operation.

A combination of these four criteria would provide the best method for bot detection, however, this would present an almost impossible task. Therefore, a method that maximizes the greatest number of these criteria would represent the best option. The most important question is which

method performs better on the key criteria of accuracy and the extent to which they deliver with regard to the other criteria. This research will assess the methods included by these criteria.

Approach 1: Applying fixed criteria in an automated or semi-automated method would maximize efficiency but may fall somewhat short in terms of simplicity and is not likely to rank high on user friendliness. There could also be issues on accuracy if the criteria established contain some bias.

Approach 2: OSINT meets the aims of being user friendly and potentially more accurate but fails on efficiency and simplicity grounds as the techniques used can often be time intensive and there is no single process that can be applied.

Data and Methods

This research will compare the accuracy and reliability of two alternative methods for identifying bots. It will do so by analyzing a recent campaign that we expect was likely to be subject to PA. In particular, we have selected the UK campaign for a second referendum or Peoples Vote on Twitter. To build our sample, we use the #FBPE hashtag on Twitter in order identify accounts that are seeking to present themselves as real users or organizations, therefore disguising their real purpose. For this research, a bot is an account that demonstrates automation while seeking to conceal that fact by attempting to operate as a normal "real" user.

Analysis will then assess the level of amplification, through bots that are present in the debate on a second referendum on the UK's membership of the EU. The focus will be on Twitter accounts that are promoting the desire to have a second referendum with the goal of remaining a part of the EU. In order to achieve this, certain criteria will be set out on how fake accounts are identified and this will then be applied to the data set relating to #FBPE.

These criteria will be applied using two separate approaches. The first approach will rely on automation to detect suspicious accounts based on predetermined characteristics. These characteristics are focused on two key areas: anonymity and high levels of activity. These criteria will be applied to all of the accounts in the data set. Automation will be tested by exploring the number of tweets being produced per day by each of the accounts and also the number of tweets they have contributed to the #FBPE data set. The limit for daily activity will be set at 72 tweets per 24-hr period in line with the recommendations set out by DFRLabs (2017). Anonymity will be assessed through the level of information that has been provided by the user to the accounts. This will rely on information regarding the use of a profile picture, whether the account has been altered from the default state and whether the account has a description or biography. If an account meets all of these criteria, they will then be checked to see if they are still active.

The second approach is concerned with the detection of unusual patterns within the data that suggest unnatural activity, which could in turn be as a result of bots. Suspicious accounts will then be investigated using manual approaches. All accounts that are identified will then be cross-checked using Botometer. This will give an indication of how various approaches can show different results when it comes to the detection of bots.

These approaches are not completely extensive nor are they fool proof. The goal in applying these techniques is to demonstrate the difficulty with which it is possible to identify automated accounts using accessible methods that can be applied by the average user and the avenues for further investigation that this can open up.

The data for this research have been collected from Twitter using the rest API, over the course of a 7-month period beginning in November and culminating on June 2, 1 week after the results of the European Election results were announced with the number of individual tweets standing at just over

1.2 million. The data collection used the key word #FBPE, which means that all tweets that contain the hashtag along with all accounts that have the hashtag in their bio or names will be included.

Case Study—#FBPE

Before the 2019 General Election, pro-EU activists had positioned themselves at the center of a debate on Brexit and the possibility of remaining in the EU. This mainly manifested itself in the People's Vote (PV) campaign, a group seeking to commit the British government to a second referendum on the UK's membership of the EU. In addition to frequent television appearances and newspaper interviews, PV has sought to boost its impact on the political landscape through creating a vociferous presence on social media and online campaigning. This can be seen in the number of high-profile Twitter accounts and sponsored content promoting PV. This online presence is also particularly visible due to the methods of promoting the pro-European position such as the use of hashtags and emoji's such as "#FBPE," "#Peoplesvote," and the EU flag emoji either in Tweets, usernames, or user descriptions (Belam, 2018). This allows like-minded users to stay in contact with each other and boost content that they agree with in the hope that this will influence former leavers in to changing their mind.

#FBPE is central to this movement. This hashtag became an important online shorthand during the campaign to raise awareness and build a larger network around the campaign for a second referendum and the remain cause. FBPE stands for "follow back pro-European." The hashtag is used in tweets relating to the fight for a second referendum and also commonly used in the bios of users to indicate their position. Primarily, this hashtag was used to encourage others of a similar political position to follow each other (Belam, 2018). The goal of the hashtag, however, means that it is particularly open to being exploited by automated accounts seeking to build up their following. Bot networks could target #FBPE in order to build up their network by tweeting pro-second referendum content but then switch their focus. This could provide a false sense of legitimacy in the accounts and lead to a greater dissemination of disinformation as a result.

Recent work has focusing on the role of bots on Twitter that were active in the buildup to the EU referendum and the following debate on a second referendum. Patel (2019) focused on accounts relating to both Leave and Remain supporters. This study found that there are a large number of accounts that aren't associated with real people who are very prominent and receive lots of interactions, they also often followed some of the patterns previously discussed with regard to unusual behavior and the sharing of articles from untrustworthy media sources. However, the focus was very much on accounts associated with a pro-leave sentiment that were often based outside of the UK. Howard and Kolanyi (2016) also researched bot activity on Twitter at the time of the 2016 EU referendum to assess the level of influence that the bots had. They found that the main activity of bots was to retweet messages relating to the UK either remaining or leaving in the EU. While there has been a great deal of research investigating the buildup to the referendum on Britain's membership of the EU, there is seemingly little research looking at how the conversation has continued, particularly in respect of the demand for a second referendum.

Findings

Automated Criteria-Based Bot Search

The initial search for automation within the #FBPE data set focused on an automatic search for bot accounts. The goal of this investigation is to identify the accounts that meet the greatest number of predetermined criteria. Given that the primary focus of the research is to identify accounts engaging

in amplification through PA, accounts that displayed high levels of unusual activity were targeted. The criteria included are the following:

- over 72 tweets per day since the creation of the account,
- over 100 tweets contributed to the #FBPE data set,
- over 90% of tweets published by the account are retweets,
- tweets published by an uncommon source,
- no account description/biography,
- default profile picture, and
- default account settings.

Accounts were then investigated based on the number of criteria that applied in Figure 1. When accounts were required to meet at least five of the seven criteria in order to be kept, this method produced a total of 324 accounts, meaning that only 0.035% of all the accounts captured in the original #FBPE data set were identified as being suspected bot accounts. Of the 324 accounts, 56 have subsequently been suspended meaning that the percentage of suspended accounts stands at 17%. When cross checked with Botometer, 31 of the accounts have a bot score of over 4.5 meaning that Botometer considers them very likely to be bots. However, of the 31 accounts with a bot score exceeding 4.5, only one of the accounts had contributed more than one tweet to the #FBPE data set, with that account only contributing three tweets. Meaning that these accounts, individually at least, had little influence over the overall conversation.

When the criteria required to identify bots is increased to six of the seven, the number of accounts decreases dramatically to 16. Of the those 16, only two have subsequently been suspended and none of the accounts register a Botometer score of above 4.5, meaning that they are not considered to be highly suspicious. This method did, however, isolate a number of the high content accounts, with eight of the 16 accounts having contributed over 100 tweets each to the data set.

When the number of required criteria is increased to seven of seven, only one account is then isolated. The account in question while being a high-volume account is still in operation and when cross-checked with Botometer does not meet the 4.5 threshold to be labeled as a bot in this research.

OSINT-Based Bot Search

Exploration of the data in Figure 2 shows that there are clear peaks and troughs with regard to activity, with a sharp increase throughout January and a large spike in mid to late March.

The higher level of activity in January coincides with a number of votes in the houses of parliament on the deal that Theresa May had negotiated with the EU. The spike in late March coincides with the point at which Britain was originally meant to officially leave the EU and also a PV march which took place on March 23. Given the importance of these events, it is natural to assume that there would be a spike in activity, but it is important to try to gauge how organic this activity was.[1]

Retweet activity gives an indication as to the extent of the amplification that was taking place. Research by Liu et al. (2014) found that of all tweets published, between 25% and 30% were retweets. Overall, the percentage of tweets that were retweets in the #FBPE data set was considerably higher at 75.3%. During the month of January, there was a very slight reduction in the proportion of tweets that were retweets at 74.7%. However, during the second peak in activity of March 23, the proportion of retweets rose to 83%. This figure takes into account the 24-hr period for March 23. This means that only 17% of all tweets that were published during this time period were original tweets by the tweet publisher. To put this into a political context, a study by Bastos et al. (2012) found that in relation to an estimated 2 million tweets associated with a number of political

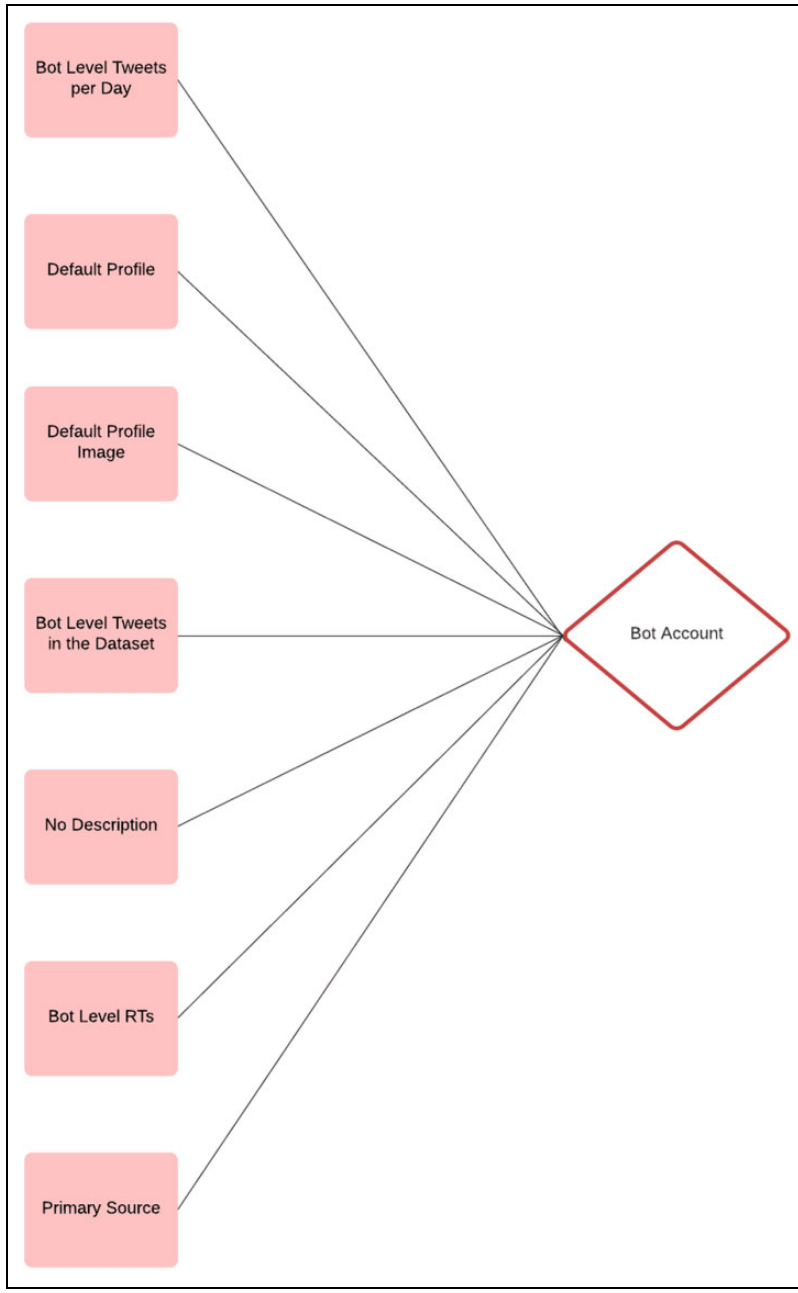

Figure 1. Visualisation of the seven distinct criteria used in catagorizing an automated account.

hashtags,[2] 791,968 were retweets. This means that 39.59% of the data set were retweets. This suggests that not only is there a high level of amplification in the #FBPE data set, compared with all tweets and hashtags relating to other political events, but this was a particularly intense period of amplification. Whether this is as a result of bot accounts is yet to be determined and requires further investigation of the accounts included in this reduced cohort.

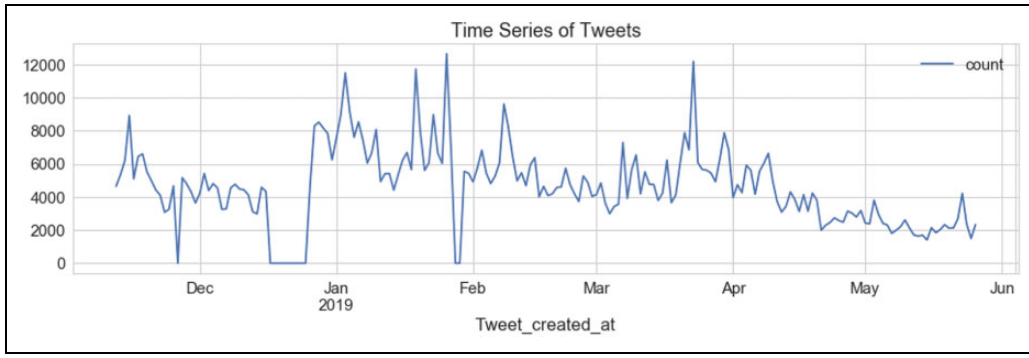

Figure 2. Time series representing the volume of tweets containing the target hashtag over the course of the data collection period.

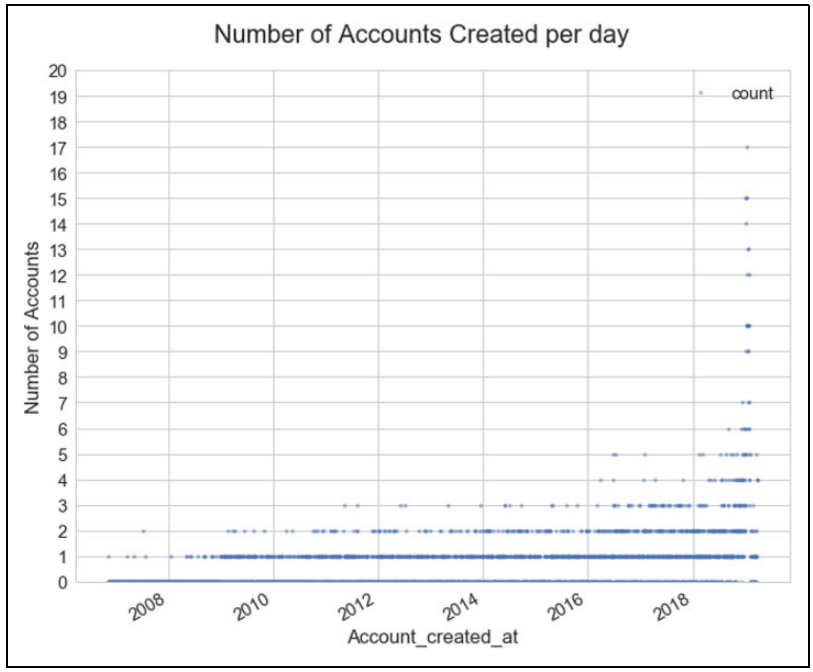

Figure 3. Depiction of the creation dates of the target 2,309 accounts.

Given that it appears there was an unusually high rate of activity at the previous points mentioned, January 2019 and March 23, 2019, further criteria were applied to the accounts to be able to separate out the more suspicious accounts from the rest. In order to do this, accounts that were found to have tweeted within the target time frame and also had tweet counts of over 72 tweets per day were separated from the main data set. This resulted in 28,054 tweets generated by 2,309 individual accounts.

To further assess amplification and test whether the accounts that were tweeting at key times of high activity were genuine, analysis sought to explore the creation dates in order to detect patterns that may have been present in the creation of accounts. It is possible that if there are a significant number of accounts being created at a particular moment that they could be linked, especially if they are tweeting about similar topics.

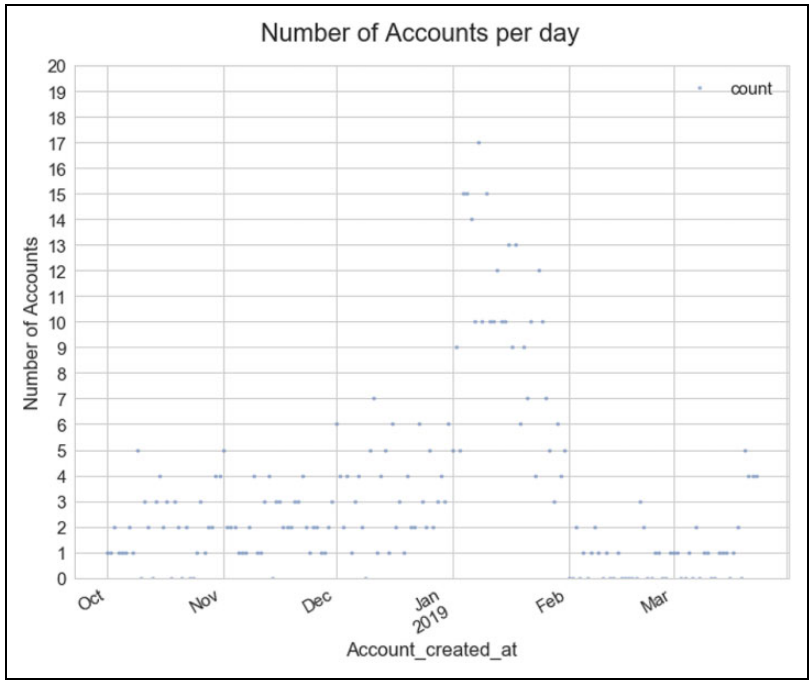

Figure 4. Depiction of the higher concentration of account creation within the period of increased tweet activity.

Figure 3 shows the times at which the 2,309 accounts in the reduced data set were created. Each dot represents the total number of accounts created on a given day, with 0 meaning that there were no accounts created on that specific day. The range of dates for the creation of the accounts is quite broad, with some accounts being over 10 years old. There is also a steady increase in the number of accounts being created the closer we get to the point of the data collection. There appears to be a significant spike in the number of accounts being created in January of 2019, which suggests that a number of the accounts that were prolific at the point at which there was a significant spike in activity were created at around that time.

Figure 4 shows the extent to which account creation increased over a 2-week period in January. The rate of the creation of new accounts then subsided over the course of February and then into March. Given the sharp rise in new accounts at the point of a particularly sharp increase in activity, these accounts seem highly suspicious.

Further analysis was then applied to accounts that were created in the month of January, starting with a sentiment analysis of the content created by the accounts. A sentiment analysis allows for the quantification of the degree to which the general discussion on the #FBPE data captured is either positive or negative and also the degree to which the debate is subjective, that is, whether the debate is based upon opinion and feeling more than objectivity. Polarity, which measures the positivity or negativity of the tweet text, is on a -1 to 1 scale with 1 being *positive*. Subjectivity is measured on a scale of 0 to 1 with 1 being *highly subjective*. A sentiment analysis can be useful for monitoring the fluctuations in the level of sentiment attached to a specific argument. This is one way of detecting when there may be a large increase in activity of automated accounts.

The sentiment analysis provides an important first step in detecting an increase in the output of fake accounts, bots, and also any attempt at PA that may be made. Given the aims of PA, it is likely that there will be surges both in the quantity of tweets about a particular topic but also a significant

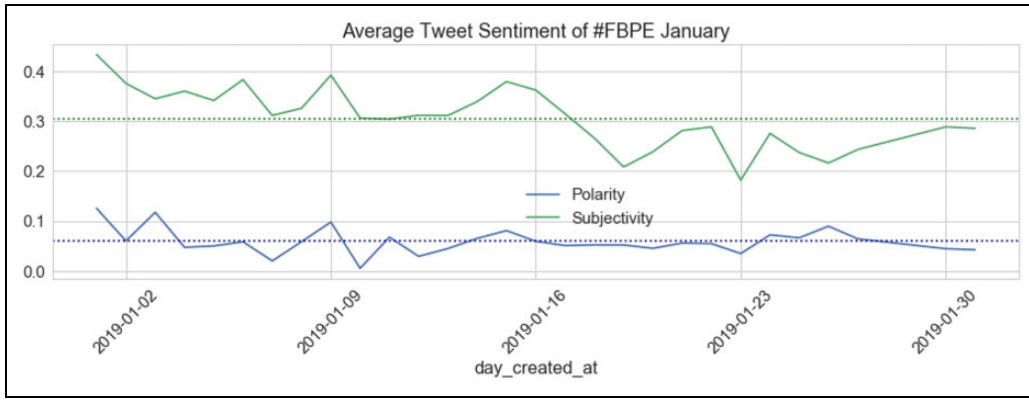

Figure 5. Average tweet sentiment for tweets containing the #FBPE hashtag through January 2019.

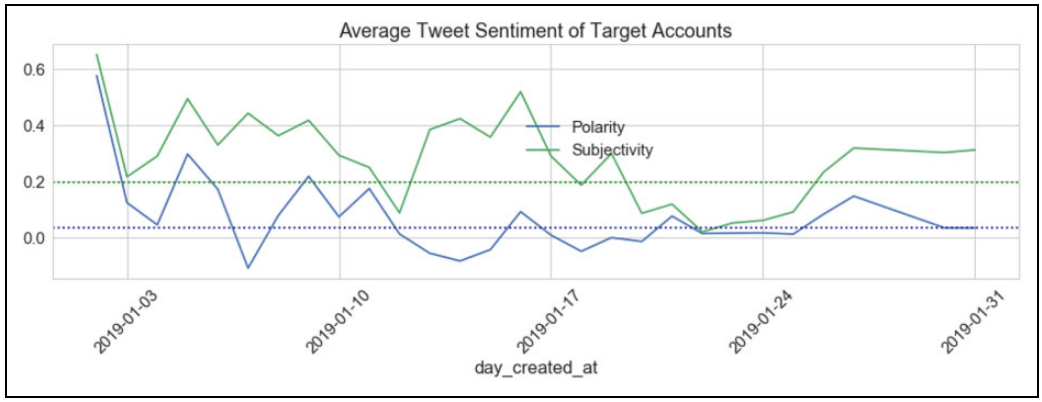

Figure 6. Average tweet sentiment for tweets containing the #FBPE hashtag through January 2019 produced by accounts drawn from target accounts.

change in the level of sentiment particularly if the debate is flooded with positive content in order to try to influence social media users. By plotting the changes in the level of sentiment, it is possible to see where there has been a significant increase and therefore reduce the focus of the search.

Figure 5 shows the level of sentiment of all tweets through the month of January in order to offer a point of comparison. On average, the tweets are just about positive, although they are close to the threshold of 0 which would represent neutrality, with the average score of 0.07. There is also a reasonable level of objectivity, with an average of 0.31 across the month.

When the sentiment analysis was applied to the reduced data set of accounts that were created in January 2019, there is a notable difference in the average of both the polarity and subjectivity of the tweets included. The polarity score is at 0.02, and the subjectivity score is at 0.2. This does seem somewhat surprising, as the expectation would be that should these accounts be involved in astro-turfing or amplification of a particular viewpoint, it would be likely that there would be a higher level of subjectivity contained within the included tweets in Figure 6.

This drop in the level of subjectivity and polarity among the accounts coincided with the highest level of activity as shown in Figure 7.

The total number of accounts identified using this method was 290. Of that 290, 110 accounts have subsequently been suspended by Twitter meaning that the percentage of suspended accounts

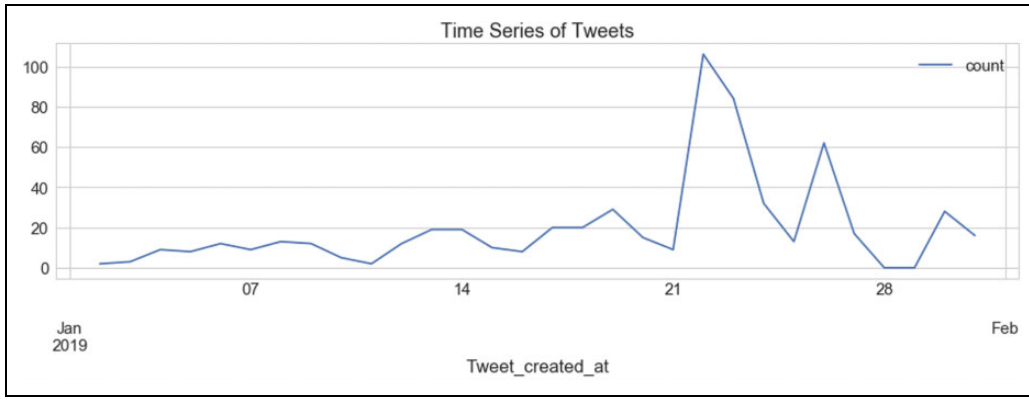

Figure 7. Time series depicting the level of tweet activity through January 2019.

stands at 37%, significantly higher than the 17% obtained using the previous method. Forty-two of the accounts identified, that are still active on Twitter, have a Botometer bot score of higher than 4.5, which again is higher than the 31 accounts using the previous method. Thirty-eight of the accounts were stated as not existing, meaning they had gone through a name change. A change in screen name has been identified as a common method across social bots, as a user redirects their focus toward a new topic or is sold to a different operator for a new purpose. This method also helps disguise the nature of the account. While the screen name of the account can change, the account id will often stay the same. As a result, it is possible to search across the full data set to try to find the different versions of these accounts. Further checks to these accounts found that two of the accounts featured in the data set under a different screen name but that screen name had subsequently been changed and did not feature again in the data set. One of the accounts was found to have changed their screen name which also featured in the data set but then was subsequently suspended, raising the number of suspended accounts to 111.

Figure 8 shows the activity of the accounts from the January cohort. There is a clear rise in activity from the accounts from mid to late January through to mid February. Activity then reduces dramatically up to until the end of June. This could have largely been as a result of many of the accounts being suspended by Twitter.

The content produced by these accounts is heavily influenced by one particular account that is still operating, which tweeted at a high volume. This account is an Italian language account that tweeted regularly about the #facciamorete hashtag. #facciamorete is an anti-government hashtag (*la Repubblica*, 2019), that appears to have a substantial crossover with #FBPE. The top 10 hashtags present are given in Table 1.

When looking exclusively at the accounts that had either subsequently been suspended by Twitter or those that had a Botometer bot score of over 4.5, the contents largely seem to centre on the tweet displayed in figure 9. What is most interesting about this tweet is the retweet to likes ratio. Usually, a tweet will often receive more likes than retweets, however, the tweet in question has only received two likes but has received over 400 retweets, 82 of which were from accounts in this data set. This is a common feature of amplification as it signifies that the engagements are unlikely to be real but rather are in place to try to spread the tweet to as many users as possible and the tweet is not being interacted with organically. Of the other four tweets in the top 5, three appear to be pro-Brexit and use the #FBPE hashtag in order to mock the FBPE community in Figure 9.

The word cloud in Figure 10 shows how influential this tweet is on the list of the most used words by the suspicious accounts. Almost every word included in the word cloud is a word that features in

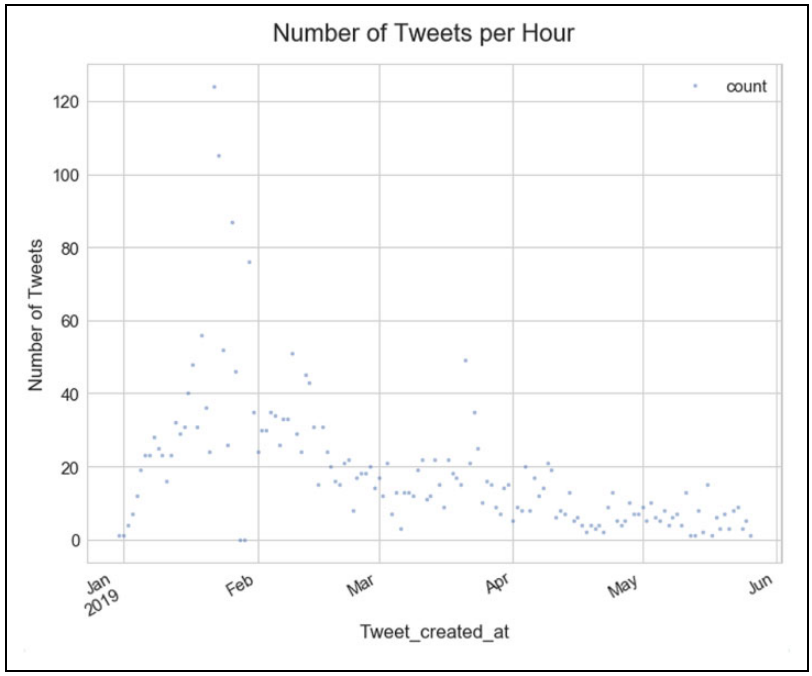

Figure 8. Depiction of the tweet activity of the target accounts through the first half of 2019.

Table 1. Count of Most Commonly Used Hashtags.

Hashtag	Count of Occurrences
#FBPE	179
#facciamorete	173
#siamoEuropei	86
#SiamoEuropei	67
#VERGOGNALega5Stelle	19
#RestiamoUmani	11
#PeoplesVote	6
#siamotanti	5
#LosersVote	4
#siamoveri	3

the tweet in question minus stopwords. While the amplification of this tweet may not have been as a result of any malpractice by the tweet publisher, the tweet does appear to have been used in an attempt to discredit Jeremy Corbyn and the position of the Labour Party on Brexit at the time of the publication of the tweet.

Assessing the Effectiveness of the Bot Detection Methods

A comparison of the two approaches outlined here indicates that the manual method is the more accurate of the two. However, there are still issues with the ability of this method to detect, with any degree of certainty, whether an account is a bot. There is also a reliance on the researcher to make subjective judgments over each individual account. This is a factor that has less influence over the rules-based approach.

Table 2. Results of the Bot Detection Methods.

Criteria	Method 1	Method 2
Accounts identified	324	290
Accounts suspended	56	111
Percentage suspended	17.2%	38.2%
Accounts still in operation labeled bot by Botometer	31	42

Figure 9. An example of a tweet drawn from the dataset that was likely subject to amplification.

Table 3 shows the breakdown of the approaches based on the classification system devised for this research.

Of the two approaches, Method 1 ranks higher across three of the four criteria due to the efficiency with which it can be applied to obtain suspected accounts and also its relative user friendliness and simplicity of use. However, they main issue with this approach is that the accuracy with which bots are identified is lower compared with the manual approach.

Discussion

This article has sought to demonstrate two potential routes to seek out instances of automation in Twitter accounts and their application for the purpose of PA. A semi-automated method removes the need for manual checking, and a manual approach relies on searching for unusual patterns of behavior. The suspicious accounts generated by these two methods were also investigated using Botometer. Between both methods, there was very little crossover in terms of the accounts that were identified. Of the 608 accounts identified across both methods, only 55 accounts were present in both of the groups identified by each method and of the top 50 most prolific accounts, in terms of quantity of tweets produced by an account, identified by each method, there was no cross over at all. This means that these two approaches both largely found different groups of accounts. Of the accounts that were checked against Botometer, the majority were found to be unlikely to be bots or the results were inconclusive. Viewed through the prism of the classifications devised for this research, the manual process would appear to be the most accurate, having uncovered the greater number of accounts that have subsequently been suspended. However, this method is the most time intensive and therefore least efficient as well as producing accounts for which there is no way to verify their validity.

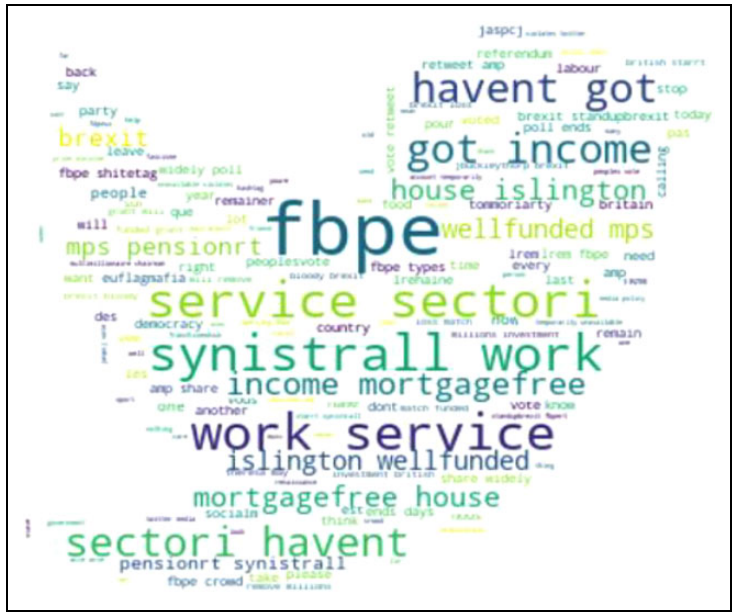

Figure 10. Work cloud of most commonly used words drawn from the tweets of target accounts.

Table 3. Classification for the Bot Detection Methods.

Criteria	Method 1	Method 2
Accuracy	Low	Moderate
User friendliness	Moderate	Low
Efficiency	High	Low
Simplicity of use	Moderate	Low

What these findings suggest is that bot detection, and particularly the detection of an orchestrated PA campaign using bots, is extremely difficult and often throws up results which are impossible to verify. This is particularly the case when the analysis takes place either during or shortly after the period of time of the event being investigated without the benefit of a ground truth data set. This then begs the question of how the average user is expected to be able to identify when another user isn't real or trying to mislead them. This is also the case within academia. Without a ground truth data set, that has been confirmed by Twitter, it is impossible to say with absolute certainty that an account is a bot.

Given that Twitter and bot detection services like Botometer can be unreliable, it would appear that for the time being at least, the onus is on the individual user to explore suspected Twitter accounts and report them if they are seen to be acting against the rules of Twitter. Central to this, however, is the ability of individuals to be able to detect these accounts and make informed decisions on them. At present, it would appear that the sole measure of success that can be applied in this respect is down to the time difference between the point of the data collection and of the analysis which allows for some level of confirmation of the malpractice of certain accounts, as a result of having been suspended. However, if accounts have not been suspended, there is little that can be done to prove with outright certainty that they are fake or bots. This means that any attempt at detecting bots or an astroturf campaign in real time are all but impossible. As a result, actors can

carry out such campaigns with relative impunity, and users will only be aware they have been manipulated after the fact. This is an issue that social media companies in general have to face up to, especially in relation to future elections. These organizations have done very little in terms of efforts to work with academics and journalists to make data available and to try to assess the impact that their products are having on individuals and on democratic processes. The black box practices which they operate mean that they have access to data we could never have before dreamed of, with the possibility to create an unparalleled understanding of human nature. It is, therefore, not beyond the realms of possibility to believe that they have the ability to know which accounts are real, label all instances of disinformation in a timely manner, and ultimately to remove accounts that are seeking to do harm in real time.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Economic and Social Research Council.

Notes

1. The missing data observed during 5 days in December and 1 day in January were a result of technical issues during data collection and cannot be interpreted as real drops in the time-series data.
2. The political hashtags included in this study were #FreeIran, #FreeVenezuela, #Jan25, #SpanishRevolution, and #OccupyWallSt.

References

Arnaudo, D. (2019). Brazil: Political bot intervention during pivotal events. In S. Wooley & P. Howard (Eds.), *Computational propaganda: Political parties, politicians and political manipulation on social media* (pp. 128–152). Oxford University Press. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780190931407.001.0001/oso-9780190931407

Baker, V. (2015). Battle of the bots. *Index on Censorship*, *44*, 127–129. https://doi.org/10.1177/0306422015591470

Bastos, M., & Farkas, J. (2019). "Donald Trump is my president!" The Internet research agency propaganda machine. *Social Media + Society, 5*. https://doi.org/10.1177/2056305119865466

Bastos, M., Travitski, R., & Raimundo, R. (2012). Tweeting political dissent: Retweets as pamphlets in #free-iran, #freevenezuela, #jan25, #spanishrevolution and #occupywallst. *OII*. http://blogs.oii.ox.ac.uk/ipp-conference/2012/programme-2012/track-a-politics/panel-6a-the-arab-spring-and-political/marco-bastos-rodrigo-travitzki-rafael.html

Belam, M. (2018). #FBPE: What is the pro-EU hashtag spreading across social media? *The Guardian*. https://www.theguardian.com/media/2018/jan/17/fbpe-what-is-pro-eu-hashtag-spreading-across-social-media

Bello, B. S., Heckel, R., & Minku, L. (2018). *Reverse engineering the behaviour of Twitter bots* [Conference session]. 2018 Fifth International Conference on social networks analysis, management and security (SNAMS) (pp. 27–34), IEEE. https://doi.org/10.1109/SNAMS.2018.8554675

Bolsover, G. (2019). China: An alternative model of a widespread practice. In S. Wooley & P. Howard (Eds.), *Computational propaganda: Political parties, politicians and political manipulation on social media* (pp. 212–238). Oxford University Press.

DFRLabs. (2017). #BotSpot: Twelve ways to spot a bot. https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. In *Paper presented at the Proceedings of the 25th International Conference on Companion on World Wide Web* (pp. 273–274). Montreal, Canada.

Gallagher, E. (2017). Bot or not? *Medium*, https://medium.com/@erin_gallagher/bot-or-not-dfabdd4e7f56

Ghosh, D, & Scott, B. (2018). #digitaldeceit: The technologies behind precision propaganda on the Internet. *New America: Public Interest Technology*. https://d1y8sb8igg2f8e.cloudfront.net/documents/digital-deceit-final-v3.pdf

Gorwa, R., & Guilbeault, D. (2018). Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*. https://doi.org/10.1002/poi3.184

Henrie, K., & Gilde, C. (2019). An examination of the impact of astroturfing on nationalism: A persuasion knowledge perspective. *Social Sciences*, *8*, 38. https://doi.org/10.3390/socsci8020038

Howard, P. N., & Kollanyi, B. (2016). Bots, #strongerin, and #Brexit: Computational propaganda during the UK-EU referendum. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2798311

Howard, P. N., & Woolley, S. C. (2016). Political communication, computational propaganda, and autonomous agents—Introduction. https://ora.ox.ac.uk/objects/uuid:ce0c948c-be5b-4864-9bc1-2dd5c162b4b8

Jackson, D. (2017). Distinguishing disinformation from propaganda, misinformation, and "fake news". *National Endowment for Democracy*. https://www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news

Keller, F. B., Schoch, D., Stier, S., & Yang, J. (2019). Political astroturfing on Twitter: How to coordinate a disinformation campaign. *Political Communication*, 1–25. https://doi.org/10.1080/10584609.2019.1661888

la Repubblica. (2019). Il successo dell'hashtag #facciamorete: Un'opposizione al governo che parte dal web [The success of the #facciamorete hashtag: An opposition to the government that started on the web]. *la Repubblica*. https://www.repubblica.it/politica/2019/01/11/news/_facciamorete_opposizione_twitter-216329909/

Lartey, J. (2018). Race and Russian interference: Senate reports detail age-old tactic. *The Guardian*. https://www.theguardian.com/world/2018/dec/24/race-russian-election-interference-senate-reports

Liu, Y., Chloe, K-S., & Mislove, A. (2014). The Tweets They are A-Changin: Evolution of twitter users and behavior. In *Proceedings of ICWSM* (Vol. *30*, pp. 5–314).

Marechal, N. (2016). When bots tweet: Towards a normative framework for bots on social networking sites. *International Journal of Communication*, *10*, 5022–5031.

Mittal, S., & Kumaraguru, P. (2014). Broker bots: Analyzing automated activity during high impact events on Twitter. arXiv:1406.4286 [Physics]. https://arxiv.org/abs/1406.4286

Patel, A. (2019). Analysis of Brexit centric Twitter activity. *F-Secure*. Retrieved March 21, 2019, from https://labsblog.f-secure.com/2019/03/12/analysis-of-brexit-centric-twitter-activity/

Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., & Menczer, F. (2011). *Truthy: Mapping the spread of astroturf in microblog streams* [Conference session]. Proceedings of the 20th International Conference Companion on World Wide Web—WWW '11 (p. 249), ACM Press, Hyderabad, India. https://doi.org/10.1145/1963192.1963301

Rauchfleisch, A., & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PLoS One*, *15*(10), e0241045.

Roth, Y., & Pickles, N. (2020). Bot or Not? The facts about platform manipulation on Twitter. *Twitter Blog*. https://blog.twitter.com/en_us/topics/company/2020/bot-or-not

Shao, C., Ciampaglia, G. L., Varol, O., Flammini, A., & Menczer, F. (n.d.). The spread of fake news by social bots. https://www.arxiv-vanity.com/papers/1707.07592/

Siriwardane, V. (2020). Why we desperately need more research on social media's effect on democracy. *Medium*. https://medium.com/the-center-for-social-media-and-politics/why-we-desperately-need-more-research-on-social-medias-effects-on-democracy-2a73af1022d8

Toler, E. (2019). Amazon's online Bezos Brigade unleashed on Twitter. *Bellingcat*. https://www.bellingcat.com/news/americas/2019/08/15/amazons-online-bezos-brigade-unleashed-on-twitter/

Twitter. (2018). Elections integrity. *Twitter*. https://about.twitter.com/en_us/values/elections-integrity.html#data

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*, 48–61. https://doi.org/10.1002/hbe2.115

Yeboah-Ofori, A., & Brimicombe, A. (2018). Cyber intelligence & OSINT: Developing mitigation techniques against cybercrime threats on social media a systematic review. *International Journal of Cyber-Security and Digital Forensics (IJCSDF)*, *7*(1), 87–98.

Ziółkowska, A. (2018). Open source intelligence (OSINT) as an element of military recon. *Security and Defence Quarterly*, *19*, 65–77. https://doi.org/10.5604/01.3001.0012.1474

Author Biographies

Oliver Beatson is currently studying toward a PhD at the University of Manchester as part of the ESRC funded CDT: Data Analytics and Society. He is particularly interested in social media analytics and the public understanding of misinformation across social media.

Rachel Gibson is a professor in Politics at the University of Manchester. Her research focuses on the impact of new information and communication technologies on political parties, particularly with regard to their activities in the elections and campaigning sphere.

Marta Cantijoch Cunill is a lecturer in Politics at the University of Manchester. Her research interests include political behaviour, political participation, elections and voting, political communication and the effects of new media.

Mark Elliot is a professor within the School of Social Sciences at the University of Manchester. In addition to Privacy his research interests include AI and Society and substantive social science topics under the broad heading of Psychological Sociology (including studies of fathering, personal relationships and social and political attitudes).