

CHAPTER 1

INTRODUCTION

Social bots are social media accounts controlled completely or in part by computer algorithms. During the past two decades, public has progressively turned to the Internet and social media to find news, share opinions, and entertain conversations. What public creates and consumes on the Internet impacts all aspects of our daily lives, including our political, health, financial, and entertainment decisions. This increased influence of social media has been accompanied by an increase in attempts to alter the organic nature of our online discussions and exchanges of ideas. In particular, over the past 10 years the public have witnessed an explosion of social bots, a presence that does not show signs of decline. Therefore, defending from social bots raises serious research challenges. They can generate content automatically and interact with human users, often posing as, or imitating, humans. Automated accounts can be harmless and even helpful in scenarios where they save manual labour without polluting human conversations.

Despite high awareness, this survey also reveals that many people are not confident in their ability to identify social bots. Social bots that go undetected by platforms have become increasingly sophisticated and human-like, introducing an arms race between bot creators and the research community. Indeed, industry and academic research about social bots has flourished recently. Social bots are programmed to autonomously create posts or tweets. If they are able to simulate human behaviour well or rather poorly depends on their “intelligence”. The basic version responds to key words or hash tags and independently posts content written beforehand only adding trending hash tags.

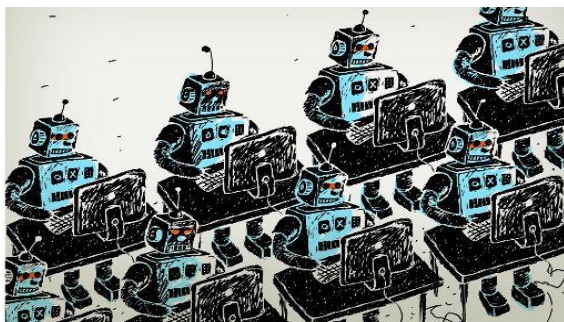


Figure 1.1: social bots

Fake accounts can be either human-generated, computer-generated (also referred to as “bots”), or cyborgs. A cyborg is a half-human, half-bot account. Such an account is manually opened by a human, but from then onwards the actions are automated by a bot. Variations exist between bots and human accounts. For example, bots are known as “Sybil” accounts when the accounts are fake, and not stolen from legitimate user. On the other hand, fake human accounts are known as “trolls” when their purpose is to defame the character of another person. Humans are just as responsible for the malicious intents found on SMPs and they therefore warrant the same attention. The difference, according to the authors, is that fake bot accounts target groups at large, whereas fake human accounts rather tend to target specific individuals. This could lead to severe consequences for the targeted individual.

Social bots appear to have played a significant role in the 2016 United States presidential election and their history appears to go back at least to the United States midterm elections, 2010. It is estimated that 9-15% of active Twitter accounts may be social bots and that 15% of the total Twitter population active in the US Presidential election discussion were bots. At least 400,000 bots were responsible for about 3.8 million tweets, roughly 19% of the total volume. Twitter bots are already well-known examples, but corresponding autonomous agents on Facebook and elsewhere have also been observed. Nowadays, social bots are equipped with or can generate convincing internet person as that are well capable of influencing real people, although they are not always reliable. Boto meter (formerly Bot Or Not) is a public Web service that checks the activity of a Twitter account and gives it a score based on how likely the account is to be a bot.

CHAPTER 2

LITERATURE REVIEW

Detecting Malicious Social Bots Based on Clickstream Sequences [1] proposed detecting and removing malicious social bots in online social networks. The most existing detection methods of malicious social bots analyse the quantitative features of their behaviour. These features are easily imitated by social bots; thereby resulting in low accuracy of the analysis. This method not only analyses transition probability of user behaviour clickstreams but also considers the time feature of behaviour. The result of this paper gives a novel method of detecting malicious social bots, including both features selection based on the transition probability of clickstream sequences and semi-supervised clustering, is presented in this paper.

In [2] Contrast Pattern-Based Classification for Bot Detection on Twitter is discussed. They have presented detection of non-human activities in social networks. This context helps in obtaining high detection accuracy with only desired quality, experts in the application domain would also like having an understandable model, with which one may explain a decision. Furthermore, it introduce a new feature model for social bot detection, which extends (part of) an existing model with features out of Twitter account usage and tweet content sentiment analysis. This paper uses a pattern-based classification mechanism to social bot detection, specifically for Twitter. This yields better classification results.

Detection of Fake Twitter Accounts with Machine Learning Algorithm [3] proposed certain machine learning algorithms for detection of fake twitter accounts. Many activities such as communication, promotion, advertisement, news, agenda creation have started to be done through social networks. Some malicious accounts on Twitter are used for purposes such as misinformation and agenda creation. This is one of the basic problems in social networks. In this proposed system, machine learning-based methods were used to detect fake accounts that could mislead people. For this purpose, the dataset generated was pre-processed and fake accounts were determined by machine learning algorithms, and proved to be more successful.

MMORPG was proposed to provide realistic user interactions in the paper Contagion of Cheating Behaviours in Online Social Networks. The diffusion process on social networks has also been leveraged to understand the spread of undesirable contagion. To detect and prevent cheating, it is beneficial to understand this dynamic as a contagion problem [4]. This paper gives the study of social contagion problem of cheating behaviour found in the massively multiplayer online role-playing game (MMORPG) that provides a lifelike environment with rich and realistic user interactions. To detect and prevent cheating, it is beneficial to understand this dynamic as a contagion problem. This paper shows the existence of the contagion of cheating.

Tracking and Characterizing the Competition of Fact Checking and Misinformation: Case Studies [5] paper proposed some of the methods that analyses the datasets collected by Hoaxy. Massive amounts of misinformation have been spreading over social media during the 2016 U.S. election season, causing wide public concern about our information ecosystem. To better understand how social bots spread misinformation, this paper conducts many case studies. By presenting strategies such as the production of a large number of original tweets, the alternating and hijacking of hash tags, and the injection of content into conversations, this demonstrate how social bots take advantages of the recommendation features of Twitter to amplify the spread of misinformation.

Machine Learning to Detect Fake Identities: Bots vs Humans paper proposed some of the engineered features such as friend-to-followers ratio. There are a growing number of people who hold accounts on social media platforms (SMPs) but hide their identity for malicious purposes [6]. These features were engineered from attributes, such as “friend-count” and “follower-count,” which are directly available in the account profiles on SMPs. In the case of bots, these machine learning models were dependent on employing engineered features, such as the “friend-to-followers ratio. The research discussed in this paper applies these same engineered features to a set of fake human accounts in the hope of advancing the successful detection of fake identities created by humans on SMPs.

Using Improved Conditional Generative Adversarial Networks to Detect Social Bots on Twitter paper [7] proposed an improved conditional generative adversarial network. The widely used bot detection method based on machine learning leads to an imbalance in the number of samples in

different categories. Classifier bias leads to a low detection rate of minority samples. To generate an auxiliary condition, we propose a modified clustering algorithm, namely, the Gaussian kernel density peak clustering algorithm (GKDPCA). The results compared with three common oversampling algorithms that achieve higher evaluation scores.

Entropy Minimization Discretization (EMD) technique pre-processed some of the datasets that is proposed in the Twitter fake account detection paper [8]. The social networking sites such as twitter and Facebook attracts many users across the world. This situation can result to a huge damage in the real world. The results obtained from the Naïve Bayes algorithm was analysed based on the discretization technique.

Detecting Spam Accounts on Twitter [9] this paper introduced some of the ML classification algorithm namely k-Nearest Neighbour (k-NN), Decision Tree (DT), Naive Bayesian (NB), Random Forest (RF), Logistic Regression (LR), Support Vector Machine (SVM), and eXtreme Gradient Boosting (XG-Boost). These algorithms improved the existing spam detection mechanism. The experimental results show that the proposed set of features gives better performance than existing state of art approaches.

CHAPTER 3

ARMING THE PUBLIC WITH ARTIFICIAL INTELLIGENCE TO COUNTER SOCIAL BOTS

A social bot is also known as social networking bot, or social bot. A social bot is a type of bot that controls a social media account. Like all bots, a social bot is automated software. The exact way a social bot replicates depends on the social network, but unlike a regular bot, a social bot spreads by convincing other users that the social bot is a real person. This also acts as an agent that communicates more or less autonomously on social media, often with the task of influencing the course of discussion and/or the opinions of its readers. It is related to chat bots but mostly only uses rather simple interactions or no reactivity at all.

3.1 CHARACTERIZATION OF SOCIAL BOTS

- Some simple bots only do one thing: They post content automatically. The Twitter account at (@) big_ben_clock is a quintessential representative of this class it tweets every hour mimicking the real Big Ben. Similar accounts can automatically post or retweet /share to publicize information such as news and academic papers. These bots are naive and easy to identify: They share only one type of content and they do not try to misrepresent themselves or their motivations. In 2011, Lee et al identified thousands of social bots. They created bots that posted meaningless messages through automated scripts. The underlying expectation was that only other bots and no human would follow these honeypot accounts. The honeypot accounts did end up having many followers. Subsequent analysis confirmed that these followers were indeed bots and revealed the common strategy of randomly following accounts to grow the number of social connections.
- Social bot developers can populate bot profiles by searching and collecting material from other platforms. A more extreme example of these kinds of bots is identity thieves: they copy usernames, profile information, and pictures of other accounts and use them as their own, making only small changes.

- More sophisticated bots adopt various strategies to impersonate human users. Sophisticated bots can emulate temporal patterns of content posting and consumption by humans. They can even interact with other users by engaging in conversations, commenting on posts, and answering questions
- Bots employ a variety of strategies to form an audience. There are bots designed to gather followers and expand social circles, with the goal of exerting some form of influence. Some bots, for example, search the social networks for popular accounts, follow them, and ask to be followed back. This type of infiltration has been proven to be more effective when the target community is topically centred.
- Another class of bots, whose activity is less evident, is that of so called fake followers. These accounts are managed by entities that get paid in exchange for following customers who want to increase their perceived popularity. The fake followers often follow each other, forming a network that lends credibility to each member and allows them to elude being flagged for lack of followers.
- A group of social bots, sometimes referred to as a botnet, may act in coordination, posting the same content again and again to generate false popularity. This type of behaviour is generally harder to detect as the accounts involved, when inspected individually, may appear as genuine. Their behaviour becomes conspicuous only when one is able to detect a large number of such accounts acting in a strongly coordinated manner.

3.2 ACTIVITY AND IMPACT OF SOCIAL BOTS

- Bot activity has been reported in several domains, with the potential to affect behaviours, opinions, and choices. Health is one domain of particular concern, where we have observed social bots influencing debates about vaccination policies and smoking. Politics is another key domain. During the 2010 U.S. midterm elections, primitive social bots were found to support some candidates and attack their opponents, injecting thousands of tweets pointing to websites with fake news.
- There was no significant difference in the amount of retweets that humans generated by resharing content produced by other humans or by bots. In fact, humans and bots re

tweeted each other substantially at the same rate. This suggests that bots were very effective at getting messages reshared in the human communication channels. We further explored how bots and humans talked about the two presidential candidates. We noted that bots tweeting about Donald Trump generated the most positive tweets. The fact that bots produce systematically more positive content in support of a candidate can bias the perception of the individuals exposed to this content, suggesting that there exists an organic, grassroots support for a given candidate, while in reality it is all artificially generated.

- The spread of fake news online is another area in which the effect of bots is believed to be relevant. A study based on 14 million tweets posted during and after the 2016 U.S. presidential election revealed that bots played a key role in the spread of low-credibility content. The study uncovered strategies by which social bots target influential accounts and amplify misinformation in the early stages of spreading, before it becomes viral.

3.3 HOW CAN YOU SPOT A SOCIAL BOT






Studies estimate that between 9-15% of traffic on Twitter and as much as 66% of shared URLs are generated by bots. So if encountering one is likely, how can you spot a social bot? As they become more and more sophisticated and constantly advance their cover-up techniques, they are difficult to detect. However, there are certain characteristics that indicate a bot:



- Bursts of activity and periods of silence are untypical for humans and indicate bot accounts.



- Their profile pictures oftentimes feature graphical pictures instead of real user pictures.

-  Skewed friend/follower-ratio: As it is way easier to follow a user than get another user to follow back, bot accounts often follow far more people than follow the bot accounts.
-  Unrealistically high number of tweets per day: The number of messages a human user can release is far lower than the number some bot accounts produce.
-  Unrealistically high reaction speed: An account replies in a split-second when it is addressed.
-  Quality of comments: Bot accounts usually have limited vocabulary and may produce inadequate or imprecise responses.
-  Many bot accounts give likes so that the liked account follows back, resulting in an unusual high number of given likes compared to human accounts.

3.4 BOT DETECTION METHODS

An early attempt by Wang et al. Involved building a crowdsourcing social bot detection platform. This method proved to be effective, but scalability was a prominent issue: while bots can be multiplied at will at essentially no cost, human detectors cannot. Later efforts therefore mostly leveraged machine learning methods.

Approaches based on supervised machine learning algorithms are the most common. Supervised approaches depend and often start with the collection of an extensive dataset, with each account labelled as either human or bot. These labels usually come from human annotation, automated methods, or botnets that display suspicious behaviours. A critical issue with existing datasets is the lack of ground truth. There is no objective, agreed-upon, operational definition of social bot. One of the factors that explain this is the prevalence of accounts that lie in the gray area between

human and bot behaviour, where even experienced researchers cannot easily discriminate. Nevertheless, datasets do include many typical bots; using the training labels as proxies for ground truth makes it possible to build practically viable tools.

The choice of relevant features used to describe entities to be classified is a critical step of machine learning classifiers. Different choices have been considered, but in general six broad categories of features have been identified as relevant for discriminating between human and bot accounts: user metadata, friend metadata, retweet/mention network structure, content and language, sentiment, and temporal features. In the case of supervised learning, after extraction and pre-processing, the features are fed into supervised machine-learning models for training, and then the trained models are used to evaluate previously unseen accounts.

While supervised methods have proven to be effective in many cases, they do not perform well at detecting coordinated social bots that post human-generated content. As mentioned earlier, those coordinated bots are not usually suspicious when considered individually. Their detection requires information about their coordination, which becomes available only once the activity of multiple bots is considered. Unsupervised learning methods have been proposed to address this issue.

By comparing the time series of accounts sampled from the Twitter streaming application program interface (API), Chavoshi et al built an unsupervised tool called DeBot that is able to find accounts tweeting in synchrony, suggesting they are automated. Chen and Subramanian adopted a similar method to detect bots by finding accounts tweeting similar content.

A recent research direction is to test the limits of current bot detection frameworks in an adversarial setting. The idea is to propose methodologies to engineer systems that can go undetected. The researches proposed the use of evolutionary algorithms to improve social bot skills. Employed a hybrid approach involving automatic and manual actions to achieve bots that would be classified as human by a supervised bot detection system. Despite the good intention of pointing to weaknesses in existing systems, this research might also inspire bot creators and give them a competitive advantage.

3.5 WHAT EFFECTS DO SOCIAL BOTS HAVE?

Social bots can have a considerable impact on society, democracy or the economy. The number of scientific studies on social bots is increasing. However, research is still in its early stages. Here are some effects that already have been investigated:

- **Impact on democracy:** Social bots have been used in elections to spread fake news and to polarize the political discussion by giving the false impression that certain information, regardless of its accuracy, is highly popular.
- **Impact on stock markets:** In the past social bots have influenced stock prices. Investment decisions are increasingly being made by automatic trading systems that promptly react to news on social media channels.
- **Impact on economy:** Bots have the potential to harm the reputation of a company or its products and lead to considerable financial damage.
- **Cybercrime:** Other studies have demonstrated how bots have gained access to private information, such as phone numbers and addresses that in turn could be used for cybercrime.
- **Distorted popularity:** First analyses of Sound Cloud, the largest social media platform for sharing music, indicate that bots are used to promote certain songs, thus influencing their popularity and leading to a wider distribution.

CHAPTER 4

BOT DETECTION TOOLS

In this chapter Botometer is used to detect fake accounts. Botometer is a machine learning algorithm trained to classify an account as bot. this checks the activity of Twitter accounts and gives them a score based on how likely they are to be bots. Some of the Botometer survey is discussed in this segment.

4.1 USER ENGAGEMENT WITH BOT DETECTION TOOLS

Research efforts aimed at bot detection may be valuable to mitigate the undesired consequences of bot activity, but their efficacy is limited by two factors: limited public awareness of the bot problem, and unwillingness to adopt sophisticated tools to combat it. The success of research efforts also critically depends on the capacity to adapt to ever changing and increasingly sophisticated artificial accounts. This, in turn, depends on the ability to engage the public and collect the feedback it provides. It is therefore important to understand how the public adopts, interacts with, and interprets the results of bot detection tools.

- Botometer is based on a supervised machine learning approach. Given a Twitter account, Botometer extracts over 1,000 features relative to the account from data easily provided by the Twitter API, and produces a classification score called bot score: the higher the score, the greater the likelihood that the account is controlled completely or in part by software, according to the algorithm. Because some of the features are based on English-language content, the bot score is intended to be used with English-language accounts. To evaluate non-English accounts, Botometer also provides a language-independent score that is produced by a classification model trained excluding linguistic features. Botometer additionally reports six sub scores, each produced by a model based on a distinct subset of features. The score names refer to the feature classes: user meta-data, friends, content, sentiment, network, and timing. The sub scores are provided to help users identify which features contribute to the overall score.



Figure 4.1: Number of daily requests handled by Botometer

- The earliest version of Botometer became available to the public in May 2014. Free access is offered through both a web interface and an API. In the past 4 years, Botometer has constantly increased its basin of adoption, and has provided data for a number of influential studies, including by Vosoughi, Roy, and Aral (2018) and the Pew Research Centre. Even more importantly, Botometer is used by many regular Twitter users. Currently, it handles over a quarter million requests every day as per the analysis in the above Figure 4.1, while the website receives over 500 daily visits. Botometer also supports several third-party tools, including browser extensions, bot-hunting bots, and localized versions in non-English speaking countries.
- Between August and October 2018, we conducted a user experience survey with participants recruited among the visitors of the Botometer website. The survey contained two questions with required answers and a few optional ones. It also allowed respondents to enter some free-text comments. We collected usable answers from 731 participants; they are listed in Figures below together with the questions. A few interesting facts emerge.
- First, more than one third of participants use Botometer at least weekly, implying that detecting bots is becoming a recurring need for at least some users (Figure 4.2).

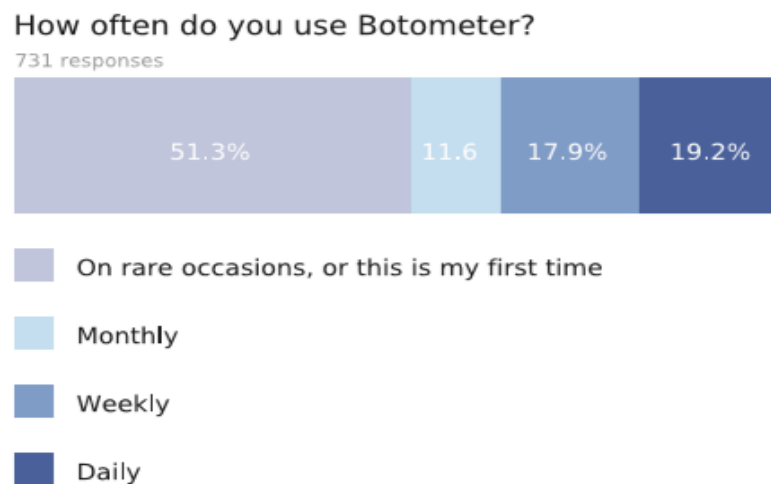


Figure 4.2: One third of participants using Botometer

- Second, over 80% of the users believe Botometer is accurate in classifying bots and humans (Figure 4.3).

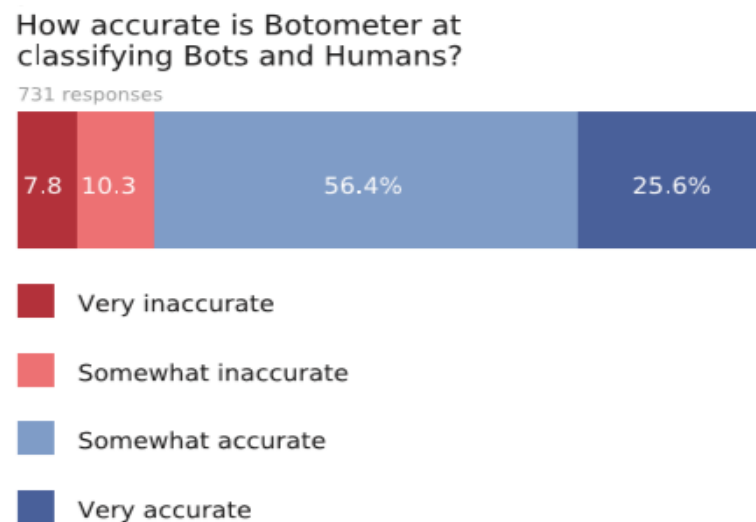


Figure 4.3: 80% of the participants believe Botometer

- Third, over 80% of the users find the bot scores presented by Botometer easy to understand (Figure 4.4). Although these numbers are encouraging, we are aware of self-selection bias, as respondents of the survey tend to be active users of Botometer.

Are the Botometer scores easy to understand?

692 responses

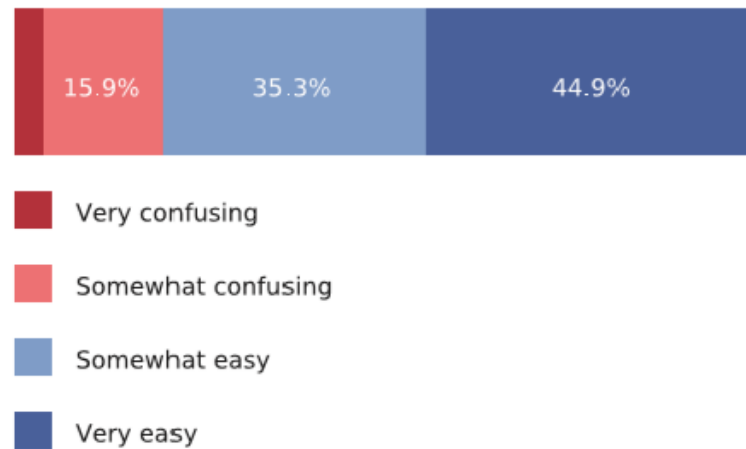


Figure 4.4: 80% of members utilizing Botometer feels are straightforward

- Finally, users seem to be equally worried by false positives (humans misclassified as bots) and false negatives (bots misclassified as humans) as shown in (Figure 4.5).

What errors should we focus on?

622 responses

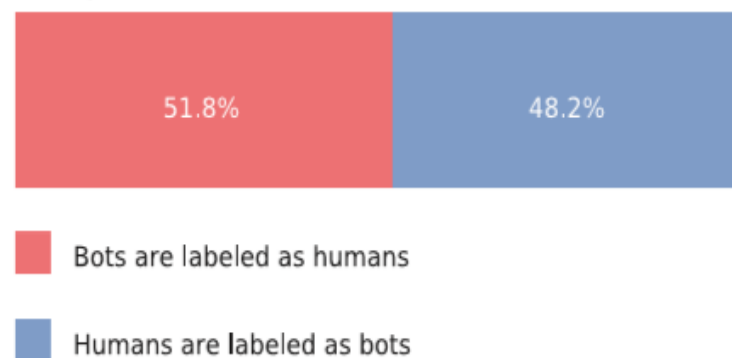


Figure 4.5: Users seem to be equally worried by false positives and false negatives.

Figure 4.6 shows that occasional users care more about false positives. A typical example of usage that might elicit this concern is when an individual checks whether their own account looks like a bot. On the other hand, frequent users care more about false negatives. For example, an advertiser interested in the number of fake followers of a celebrity paid to endorse a product may be more worried about missing bots. Respondents who use Botometer daily or weekly are considered frequent users.

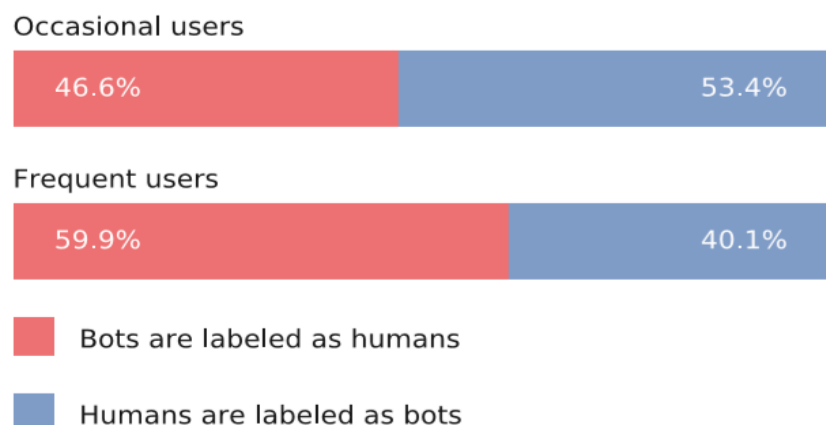


Figure 4.6: Responses to the Botometer user experience survey question about error type concerns, grouped by frequency of usage.

4.2 MALICIOUS SOCIAL BOTS DETECTION

Data set cleaning and screening, data feature processing, data classification, and a series of operations were conducted after acquiring clickstream data set of the user. The detailed steps are shown in Figure 4.7.

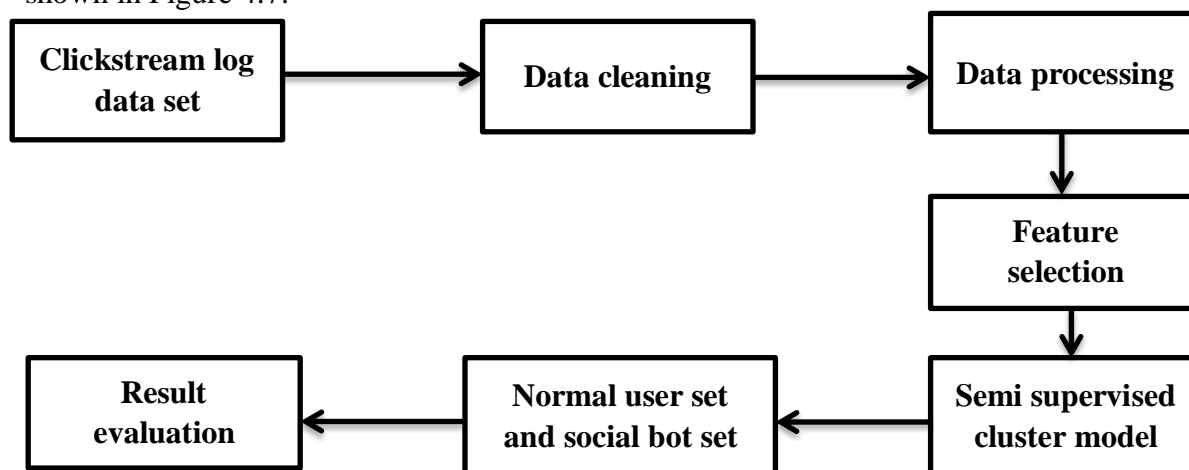


Figure 4.7: Experimental procedure for detecting malicious social bots

- **Data cleaning:** Data that are clicked less must be cleaned to remove wrong data, obtain accurate transition probability between clickstreams, and avoid the error of transition probability caused by fewer data.

- **Data processing:** Some data are selected randomly from the normal user set and social bots set to the label. Normal user account is labeled as 1, and the social bots account is labeled as -1. Seed users are classified as the category of clusters.
- **Feature selection:** In the spatial dimension according to the main functions of the CyVOD platform, we select the transition probability features related to the playback function: $P(\text{play}, \text{play})$, $P(\text{play}, \text{like})$, $P(\text{play}, \text{feedback})$, $P(\text{play}, \text{comment})$, $P(\text{play}, \text{share})$ and $P(\text{play}, \text{more})$; in the time dimension: we can get the inter-arrival times (IATs). Because if all transition probability matrixes of user behavior are constructed, extremely huge data size and sparse matrix can increase the difficulty of data detection.
- **Semi-supervised clustering method:** First, the initial centers of two clusters are determined by labeled seed users. Then, unlabeled data are used to iterate and optimize the clustering results constantly.
- **Obtain the normal user set and social bots set:** The normal user set and social bots set can be finally obtained by detecting.
- **Result evaluation:** This evaluates results based on three different metrics: Precision, Recall, and F_1 Score. In the meantime, this uses Accuracy as a metric and compares it with the SVM algorithm to verify the efficiency of the method. Accuracy is the ratio of the number of samples correctly classified by the classifier to the total number of samples.

Clickstream are the user tabs that navigate few sites. Clickstream analysis is a process of collecting and analysing the data about the pages. This shows the user tabs how much they have reviewed, how long they have stayed online and time saved when they have left.

CHAPTER 5

DETECTION MODELS

A supervised machine learning tool is only as good as the data used for its training. Social bots evolve rapidly, and even the most advanced algorithms will fail with out-dated training datasets. Therefore it is necessary to update classification models, using newly available data as well as feedback collected from users. At the same time, one must continuously evolve the set of features that may discriminate between human behaviours and increasingly complex bot behaviours. Let us again use the Botometer case to illustrate the need for retraining bot detection models via new data and feature engineering.

5.1 TRAINING DATASETS IN BOTOMETER

In the past few years, several bot datasets collected by colleagues in academia have been included in our training data. Such data are now shared with the public at large via a platform called Bot.

Table 1 lists the datasets currently included and used for training, while Table 2 summarizes the different versions of Botometer and the growing training dataset it employed. Retraining does not only mean adopting new datasets as they become available. In light of journalistic evidence and other academic research, the list of features employed by Botometer was recently enriched to incorporate new ones that are designed to capture bots employed in information operations.

Examples include:

- **Time zones:** An account would be suspicious if its profile indicates the US Eastern Standard Time zone while most of its followers appear to be in the Moscow Standard Time zone.
- **Language metadata:** Anomalous patterns in language use and audience language preferences can be revealed by a low fraction of neighbours (friends, followers) having the same language as a target account.
- **Device metadata:** This capture the types of devices and platforms used for posting tweets, as well as the entropy across different platforms.

- **Content deletion patterns:** Highly active accounts frequently create and delete content to hijack user attention without revealing too much information on their excessive posting behaviour.
- **TABLE 1** Training data used by Botometer. CNetS refers to the Centre for Complex Networks and Systems Research at Indiana University.

Dataset name	#Bots	#Human	Notes
caverlee	22,179	19,276	Honeypot-lured bots and sample human accounts (Lee et al., 2011)
varol-icwsm	826	1,747	Manually labeled bots and humans sampled by Botometer score deciles (Varol, Ferrara, Davis, et al., 2017)
cresci-17	10,894	3,474	Spam bots and normal humans (Cresci et al., 2017)
pornbots	21,963	0	Pornbots shared by Andy Patel (github.com/r0zetta/pronbot2)
celebrity	0	5,970	Celebrity accounts collected by CNetS team
vendor-purchased	1,088	0	Fake followers purchased by CNetS team
botometer-feedback	143	386	Botometer feedback accounts manually labeled by author K.-C.Y.
political-bots	62	0	Automated political accounts run by @rzazula (now suspended), shared by @josh_emerson on Twitter
Total	57,155	30,853	All datasets available at botometer.iuni.iu.edu/bot-repository

- **TABLE 2** Datasets and numbers of features used for training subsequent versions of Botometer models.

Version (activity period)	#Features	Training datasets	Notes
v1 (1 May 2014–3 May 2016)	1,150	caverlee	Initial version of Botometer (known as BotOrNot at the time)
v2 (3 May 2016–11 May 2018)	1,150	caverlee, varol-icwsm	Version used in results presented by Varol, Ferrara, Davis, et al. (2017)
v3 (11 May 2018–Present)	1,209	caverlee, varol-icwsm, cresci-17, pornbots, vendor-purchased, botometer-feedback, celebrity	New features introduced to capture more sophisticated behaviors and to comply with Twitter changes.

5.2 STRATEGIES USED BY SOCIAL BOTS: CASE STUDIES

As mentioned earlier that social bots can play an important role in the spread of misinformation. Here these provide some concrete examples to show how social bots amplify misinformation diffusion. Based on the social reinforcement theory and social contagion models that a user prefers to adopt a social behaviour if it has received reinforcement from various social groups, hypothesizing two rules (**R1** and **R2**) that are followed by social bots to amplify the spread of misinformation. Considering the recommendation mechanisms of the Twitter platform, the case study introduces the following strategies that are being used by social bots.

➤ **Case (I) Producing a Large Number of Original Tweets:**

The simplest strategy that follows rule **R1** is to produce a large number of original tweets.

The numeric statistics of the tweets sharing this article are shown in Figure 5.1. Among the total 11,944 tweets, 6,454 (54%) come from one single account. And 95% of these tweets from are original tweets. To avoid detection of such anomalous behaviour by Twitter, this account posted a low daily volume of tweets over a long time rather than a large burst of tweets in a short time. This can observe the behaviour in Figure 5.2. However, researchers have found that potential readers increase fast only in the early stage of a trend, and most of the active periods of spread are a week or shorter. Therefore, this behaviour of slow posting over seven months is not normal. At the time of this writing, the account is still active, suggesting that Twitter's anti-abuse algorithms fail to detect this kind of stealth tactic.

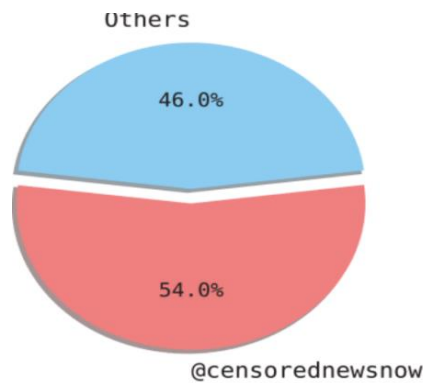


Fig 5.1

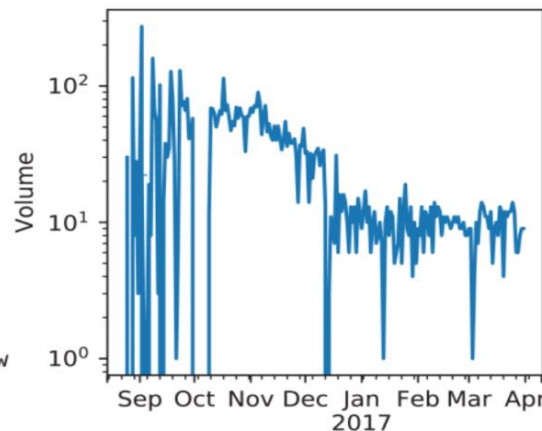


Fig 5.2

Figure 5.1: Share of tweets by the account. Figure 5.2: Timeline of original tweets posted.

➤ **Case (II) Alternating and Hijacking Hash tags (#)**

A hash tag is a keyword or a phrase used to describe a topic or a theme. Users make use of hash tags to categorize and summarize their posts on Twitter. Moreover, there is a recommendation mechanism behind hash tags: Twitter may recommend trending hash tags based on a user's profile. Bots could take advantage of the hash tag feature by posting tweets with trending hash tags, so that their tweets would have a high chance to catch the attention of users who follow the trends. This strategy is an indirect way to achieve rule **R2**.

- To illustrate the hijacking of hash tags, let us consider another example the *#MeToo* hash tag. *#MeToo* spread virally in October 2017 as a hash tag used on social media to help demonstrate the widespread prevalence of sexual assault and harassment. However, the method also noticed that social bots utilized this viral hash tag for unintended purposes. Figure 5.3, illustrates a diffusion network for the *#MeToo* hash tag. Nodes represent Twitter accounts and links represent the propagation of the hash tag through retweets and mentions/replies.

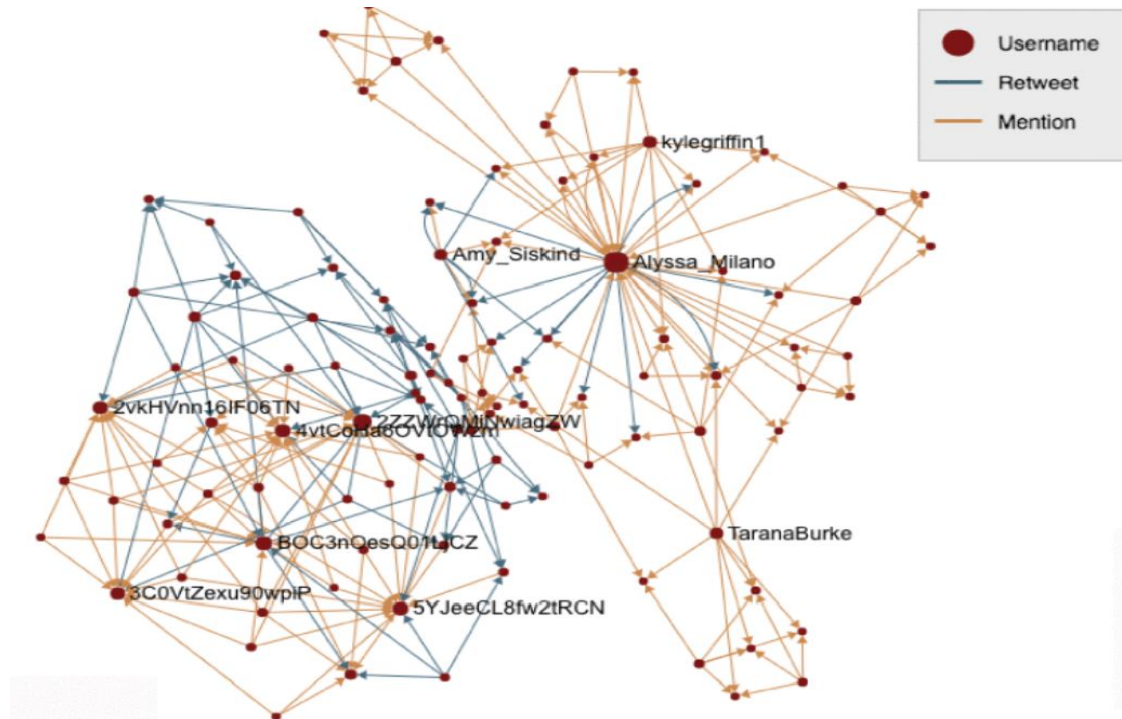


Figure 5.3: Case (II) hijacking the hash tags. Diffusion network for the *MeToo* hash tag. Nodes represent Twitter accounts. When account A retweets B, the direction of this edge is from B to A. And when account A mentions B, the direction of this edge is from A to B

The network above can roughly be divided into two clusters. In the left cluster, we observe that most of the accounts have suspicious screen names, such as 5YJeeCL8fw2tRCN, 3C0VtZexu90wpiP, 4vtCoHa6OVtOWzm, etc. This believes that accounts were social bots; in fact, Twitter has since suspended them. However, due to the suspension, we cannot fetch the actual content of these tweets. Anyway, this is a good example to show how social bots hijack trending hash tags.

➤ *Case (III) Injecting into the Conversation*

Conversations happen all the time on Twitter, and they all start with just one reply to a tweet. When a user navigates a tweet, all replies to this tweet are listed below it, so that users can read through these comments and even join the conversation by replying. Moreover, the recommendation of the Twitter platform makes it easy to find and join popular conversations which users may be interested in. The social bots can take advantage of this feature to inject content into these conversations. Among the total 4,068 tweets, 1,935 (48%) are from the single account @garydixson, which has since been suspended (see Figure. 5.4(a)). And of these tweets from @garydixson, 1,407 (73%) are replies, indicating that the replies dominate the spreading. User looked at who are the users that are replied to Figure. 5.4(b) shows a word cloud of the replied-to users, and can see that the top four frequently replied-to users are CNN, CNN Politics, the hill and NBC News, all mainstream news organizations.

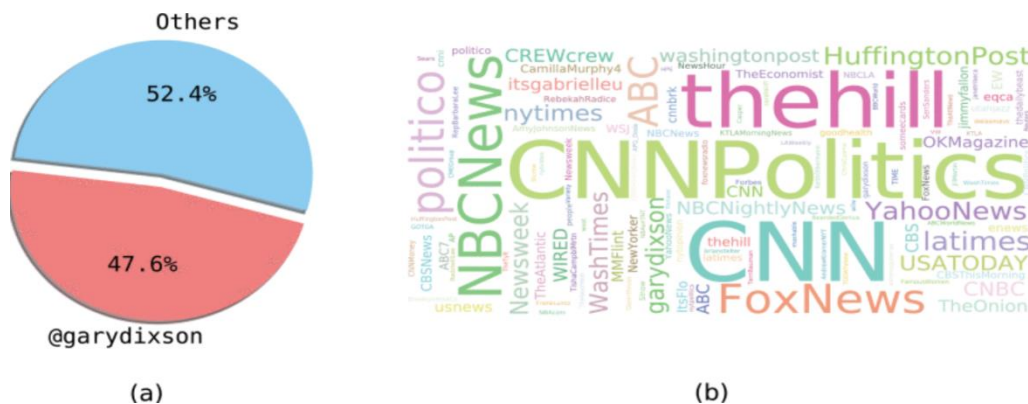


Figure 5.4: Case (III): Injecting conversations. (a) Share of tweets by the account. (b) Word cloud of screen names.

5.3 ADVANTAGES AND DISADVANTAGES OF SOCIAL BOTS:

Advantages of Using Social bots:

- Faster Customer Service
- Increased Customer Satisfaction
- Lower Labor Costs
- Variety of Uses

Disadvantages of Using Social bots:

- Limited Responses for Customers
- Customers Could Become Frustrated
- Complex Social bots Could Cost More
- Not All Business Can Use Social bots

CHAPTER 6

RESULTS AND DISCUSSIONS

The score produced by such a model is the fraction of trees that classify the account under examination as a bot. This interpretation of the bot score is obscure to the typical user, who is unlikely to have background knowledge in machine learning. Because a bot score is defined in the unit interval, it is tempting to interpret it as the probability that the account is a bot if the score is 0.3, say, then 30% of accounts with similar feature values are bots. Another interpretation is that such an account is 30% automated. These changes were incorporated into the latest version of Boto meter (V3), launched in May 2018, together with other improvements. Users seem to appreciate these changes, as shown in Figure 6.1.

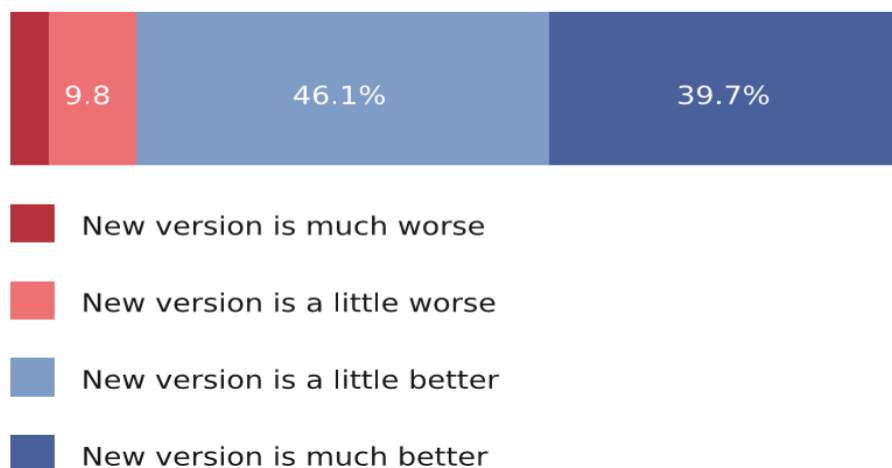


Figure 6.1 Botometer user experience survey result [3].

To calibrate Botometer, the system employs Platt's scaling, a logistic regression model trained on classifier outputs. Figure 6.2 represents the initial model outcomes and the calibrated scores. Note that this mapping shifts scores within the unit interval but preserves order, therefore leaving the AUC unchanged. The Figure also shows reliability diagrams for raw and calibrated scores. For each bin, the mean predicted score is computed and compared against the fraction of true positive cases. In a well-calibrated model, the points should align with the diagonal.

Observe that the blue line on the right side of Figure 6.2 is steepest in the middle of the range; this is because most uncalibrated bot scores fall near the middle of the unit interval. Since users, when presented with a single uncalibrated bot score, do not know that most scores fall into this

relatively narrow range, they are misled into perceiving uncertainty about classification of most accounts.

The flatter red line in the plot shows that each bin has approximately the same number of scores in the calibrated model.

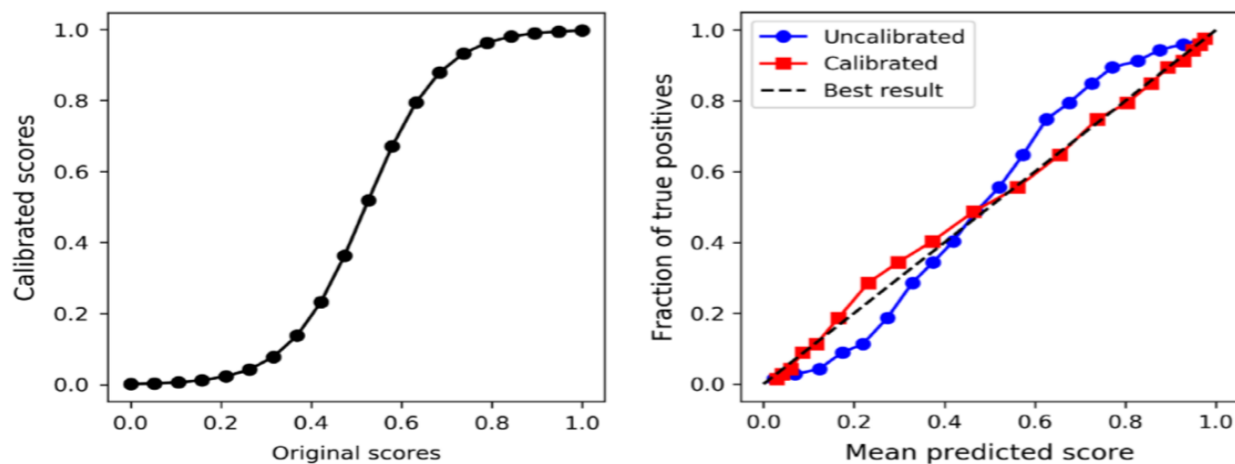


Figure 6.2: Calibration of the bot scores [3].

The mapping function projects raw classifier outputs to calibrated scores (left). Reliability curves plot true positive rates against mean predicted scores (right). The calibrated curve indicates higher reliability because it is closer to the unbiased diagonal line.

The model's training data provides empirical distributions of scores for both humans and bots. Figure 6.2(a) and (b) density estimation can be used to find a probability density function likely to produce a particular empirical distribution. Binning is the simplest approach to density estimation, sometimes employing a sliding window. However, this approach proved unsuitable because quantization artifacts in the Botometer classifier output lead to discontinuities in the density functions.

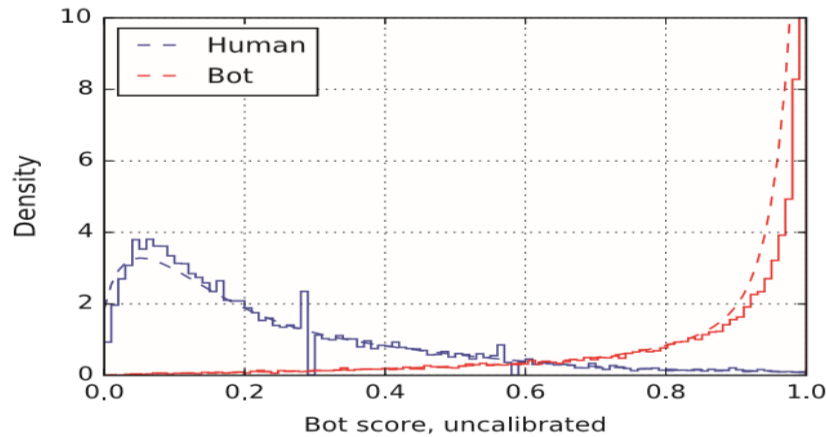


Figure 6.2(a): Likelihood distributions and posterior probabilities [3].

The plot shown in the Figure 6.2(a) is the binned, empirical bot score distribution for accounts labeled human and bot, along with dashed lines displaying the density estimate for each.

Experiments showed that a 40th-degree polynomial fit produced satisfactory results Figure 6.2(b). With likelihood curves generated from our classifier testing data, it is straightforward to calculate the evidence term in Bayes' rule. The other term is the prior $P(\text{Bot})$, the background probability of a given account being a bot. This is necessary because the training data does not include information on how common bots are in the wild.

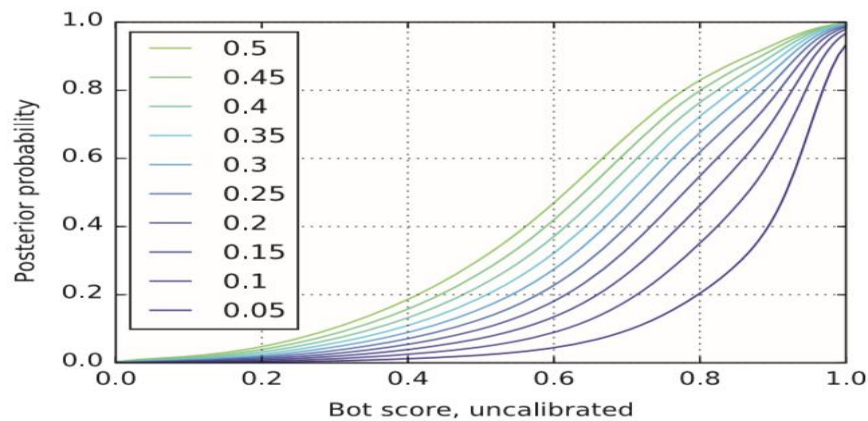


Figure 6.2(b): The complete automation probability calculation (CAP [3].

The plot in the Figure 6.2(b) shows the posterior probability curves are calculated for several choices of the prior, $P(\text{Bot})$. In Boto meter, the posterior is called CAP. As shown in Figure 6.2(b), this probability is generally more conservative than the calibrated bot score, reflecting the relative rarity of bots. The reported CAP estimates the probability that the account is indeed a bot, and gives end users the information they need to act on the data returned.

CONCLUSION AND FUTURE SCOPE

In this report artificial intelligence for detecting bot behaviors on the social network, Twitter is described. Bot detection problems have become a significant research area due to many companies is investing resources in detecting abnormal behavior in their accesses. As a consequence, this method has been proposed for bot detection, but there is no proposal using understandable models for detecting bots on social networks. Botometer helps in detecting social bots in twitter accounts. This is one of the novel methods to accurately detect malicious social bots in online social networks.

Algorithmic interpretability would make it easier to spot such biases. Nevertheless, social media are particularly vulnerable because they facilitate automatic interactions via software. As a result, social media platforms have to combat a deluge of attacks. The advanced machine learning algorithms could further be extended to other social networks like Facebook, quora, Instagram etc. As a result bot detection tools are used to make important decisions, such as whether an account should be suspended, the issue of algorithmic bias is an important direction for future research. The main future scope of the social bots are summarized as follows,

- Bots will become regular and acknowledged parts of the social media ecosystems of the future.
- Successful implementation and further advances of these technologies can pose serious threats, since discriminating between organic human behavior and automation is likely to become a more and more challenging task, even for experts.
- The future of social bots may be shaped by legal, ethical, and political considerations as much as technological ones.
- Political interference could also be counterproductive.

REFERENCES

- [1] P. Shi, Z. Zhang and K. R. Choo, "Detecting Malicious Social Bots Based on Clickstream Sequences," in *IEEE Access*, vol. 7, pp. 28855-28862, 2019.
- [2] O. Loyola-González, R. Monroy, J. Rodríguez, A. López-Cuevas and J. I. Mata-Sánchez, "Contrast Pattern-Based Classification for Bot Detection on Twitter," in *IEEE Access*, vol. 7, pp. 45800-45817, 2019.
- [3] İ. AYDIN, M. SEVİ and M. U. SALUR, "Detection of Fake Twitter Accounts with Machine Learning Algorithms," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, 2018, pp. 1-4.
- [4] J. Woo, S. W. Kang, H. K. Kim and J. Park, "Contagion of Cheating Behaviors in Online Social Networks," in *IEEE Access*, vol. 6, pp. 29098-29108, 2018.
- [5] C. Shao, P. Hui, P. Cui, X. Jiang and Y. Peng, "Tracking and Characterizing the Competition of Fact Checking and Misinformation: Case Studies," in *IEEE Access*, vol. 6, pp. 75327-75341, 2018.
- [6] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," in *IEEE Access*, vol. 6, pp. 6540-6549, 2018.
- [7] B. Wu, L. Liu, Y. Yang, K. Zheng and X. Wang, "Using Improved Conditional Generative Adversarial Networks to Detect Social Bots on Twitter," in *IEEE Access*, vol. 8, pp. 36664-36680, 2020.
- [8] B. Erşahin, Ö. Aktaş, D. Kılınç and C. Akyol, "Twitter fake account detection," *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, 2017, pp. 388-392.
- [9] Z. Alom, B. Carminati and E. Ferrari, "Detecting Spam Accounts on Twitter," *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, 2018, pp. 1191-1198.
- [10] M. Egele, G. Stringhini, C. Kruegel and G. Vigna, "Towards Detecting Compromised Accounts on Social Networks," in *IEEE Transactions on Dependable and Secure Computing*, vol. 14, no. 4, pp. 447-460, 1 July-Aug. 2017.