Arming the Public with Artificial Intelligence to Counter Social Bots

Ravi Gatti
Assistant Professor, Department of Electronics and Communication Engineering,
Sri Venkateshwara College of Engineering, Bangalore
Bangalore, India
ravi.ncg@gmail.com

Meghana R
UG Student, Dept. of Electronics and Communication Engineering
Sri Venkateshwara College of Engineering, Bangalore
Bangalore, India
meganar98@gmail.com

Abstract- The increased relevance of social media in our daily life has been accompanied by efforts to manipulate online conversations and opinions. Creating artificial intelligence (AI) apparatuses to arm the general population in the battle against social bots. The paper reviews the literature on different types of bots, their impact, and detection methods. Utilization of the contextual analysis of Botometer, a famous bot discovery apparatus created at Indiana University, to represent how individuals communicate with AI counter measures. A user experience survey suggests that bot detection has become an integral part of the social media experience for many users. However, barriers in interpreting the output of AI tools can lead to fundamental misunderstandings. The weapons contest between AI techniques to create complex bots and compelling counter estimates makes it is important to refresh the preparation information and highlights of recognition apparatuses. The Botometer case is used to illustrate both algorithmic and interpretability improvements of bot scores, designed to meet user expectations. This helps in discussing how future AI developments may affect the fight between malicious bots and the public.

Keyword- Botometer, social bots, AI techniques.

I.INTRODUCTION

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. Social bots are social media accounts controlled completely or in part by computer algorithms. 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. 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. 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. Botometer (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 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.

II. LITRARTURE SURVEY

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.

III. METHODOLOGY

In this section, the present project gives an intelligent detecting malicious social bots in to detect the social bots. The block diagram in the below given figure1 shows the detail steps in detecting the malicious social bots. 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.

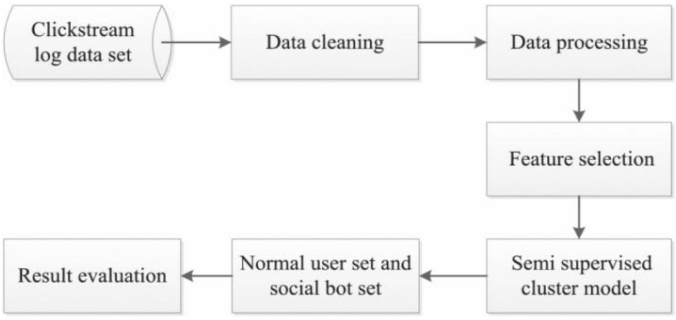

Figure 1: Experimental procedure for detecting malicious social bots

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(play,play) , P(play,like) , P(play,feedback) , P(play ,comment) , P(play,share) and P(play,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 valuates 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.

IV. 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.

V. BOT DETECTION TOOLS

In this segment Botometer technique 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.

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.

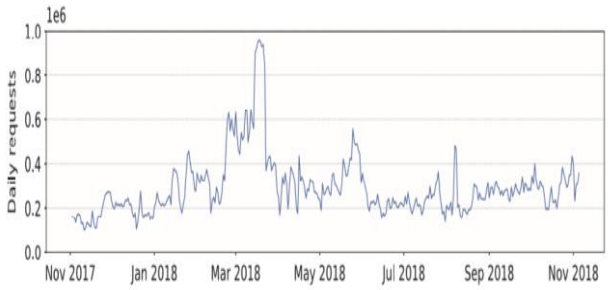

Figure 2: Number of daily requests handled by Botometer.

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 2, 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, this 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. This 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 3).

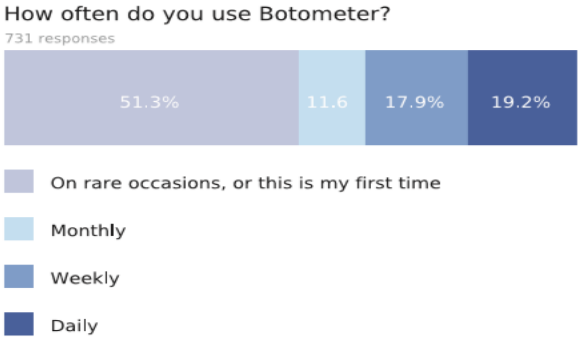

Figure 3: One third of participants using Botometer

Second, over 80% of the users believe Botometer is accurate in classifying bots and humans (Figure 4).

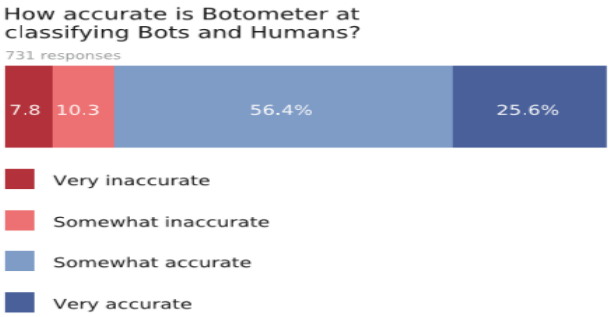

Figure 4: 80% of the participants believe Botometer

Third, over 80% of the users find the bot scores presented by Botometer easy to understand (Figure 5). Although these numbers are encouraging, we are aware of self-selection bias, as respondents of the survey tend to be active users of Botometer.

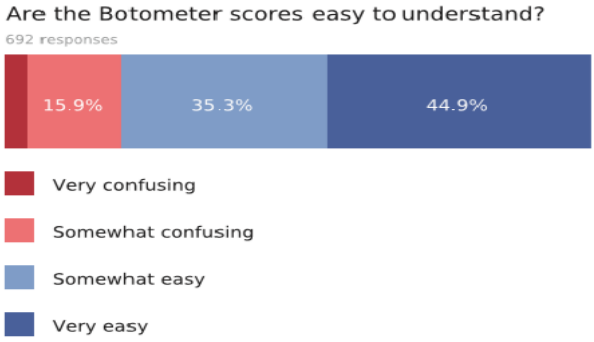

Figure 5: 80% of members utilizing Botometer feels are straightforward

VI. 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. 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 6. 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 7 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..

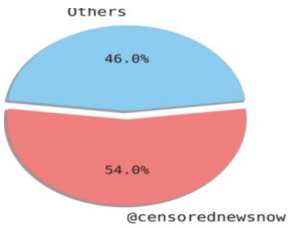

Figure 6: Share of tweets by the account

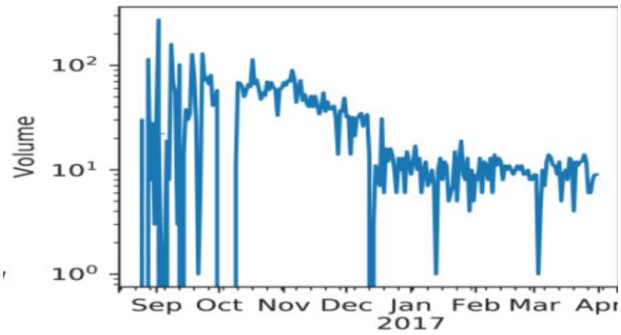

Figure 7: Timeline of original tweets posted.

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.

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 8, 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.

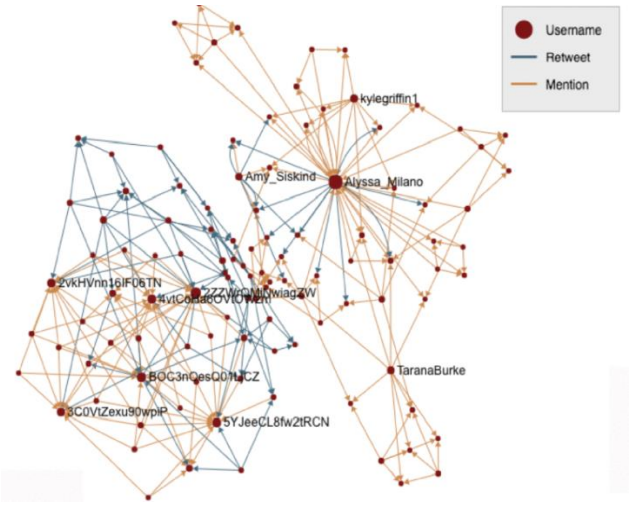

Figure 8: Hijacking the hash tags. Diffusion network for the MeToo hash tag. Nodes represent Twitter accounts. When account A retweets B, the direction of this edge is from B to A. And when account A mentions B, the direction of this edge is from A to B.

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. 9). 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. 10 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.

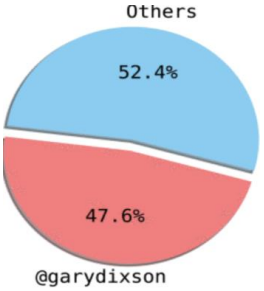

Figure 9: Injecting conversations Share of tweets by the account.

Figure 10: Word cloud of screen names.

VII. RESULTS AND DISCUSSION

To calibrate Botometer, the system employs Platt's scaling, a logistic regression model trained on classifier outputs. Figure 11 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 11 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.

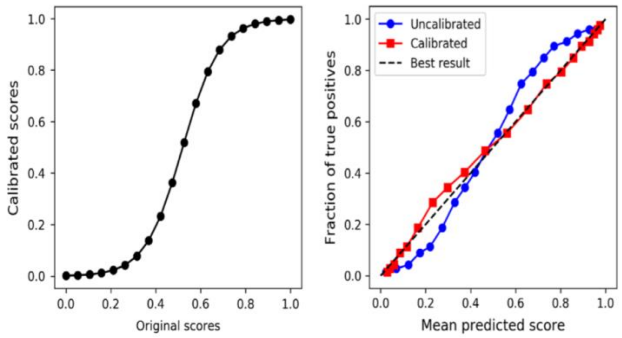

Figure 11: Calibration of the bot scores

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 12 and 13 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.

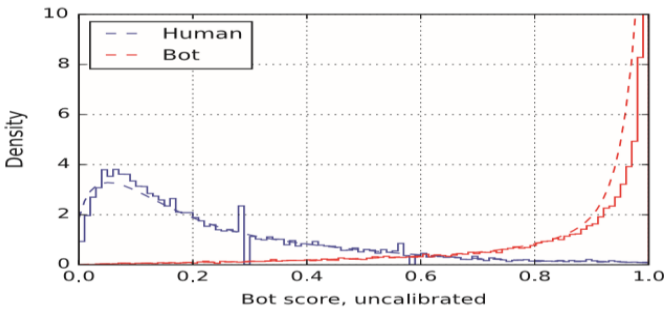

Figure 12: Likelihood distributions and posterior probabilities

The plot shown in the Figure 12 is the binned, empirical bot score distribution for accounts labelled 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 13. 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 (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.

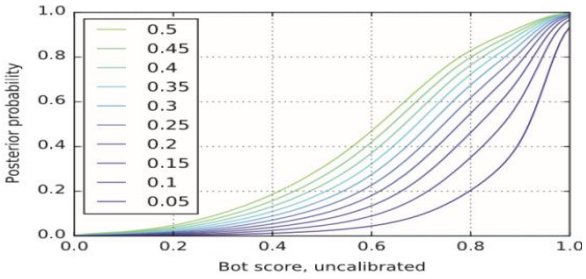

Figure 13: The complete automation probability calculation (CAP).

The plot in the Figure 13 shows the posterior probability curves are calculated for several choices of the prior, P (Bot). In Boto meter, the posterior is called CAP. As shown in Figure 13, 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

In this work 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 behaviour 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.

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. 447-460,1July-16