

Identifying Active, Reactive, and Inactive Targets of Socialbots in Twitter

Mohd Fazil

Department of Computer Science
Jamia Millia Islamia, New Delhi, India
mohdfazil.jmi@gmail.com

Muhammad Abulaish, SMIEEE*

Department of Computer Science
South Asian University, New Delhi, India
abulaish@sau.ac.in

ABSTRACT

Online social networks are facing serious threats due to presence of human-behaviour imitating malicious bots (aka socialbots) that are successful mainly due to existence of their duped followers. In this paper, we propose an approach to categorize Twitter users into three groups – *active*, *reactive*, and *inactive* targets, based on their interaction behaviour with socialbots. Active users are those who themselves follow socialbots without being followed by them, reactive users respond to the *following* socialbots by following them back, whereas inactive users do not show any interest against the *following* requests from anonymous socialbots. The proposed approach is modelled as both binary and ternary classification problem, wherein users' profile is generated using static and dynamic components representing their identical and behavioural aspects. Three different classification techniques viz Naive Bayes, Reduced Error Pruned Decision Tree, and Random Forest are used over a dataset of 749 users collected through live experiment, and a thorough analyses of the identified users categories is presented, wherein it is found that *active* and *reactive* users keep on frequently updating their tweets containing advertising related contents. Finally, feature ranking algorithms are used to rank identified features to analyse their discriminative power, and it is found that *following rate* and *follower rate* are the most dominating features.

KEYWORDS

Social network analysis, Twitter data analysis, User profiling, Socialbot characterization, Socialbot identification

1 INTRODUCTION

Twitter, one of the most popular microblogging platforms, is quite open in nature. It allows the users to follow their friends, family members, celebrities, politicians, etc. and get updated with their views about events, revelations about their personnel life, etc. in real-time. Twitter is more about information propagation and ideas expressions rather than entertainment like other online social networking platforms. It is very democratic in nature and allows any user to follow any other user and get subscription of his/her tweets.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WI '17, Leipzig, Germany

© 2017 ACM. 978-1-4503-4951-2/17/08...\$15.00

DOI: 10.1145/3106426.3106483

Twitter facilitates users to share ideas and thoughts about any event without restrictions. This democratic and open nature along with the huge user-base attract malicious users creating new kinds of problems that are very different from conventional problems in terms of sophistication level, scalability, robustness, and so on. Emergence of socialbots is one such problems being faced by the online social media.

Socialbots are the automated programs who mimic human behaviour to resemble human-beings. Socialbots require followers and friends to built-up trust and reputation in online social networks and in the process they are facilitated by other similar users or by those who randomly follow any user. In existing literature, different aspects of socialbots have been analysed. Researchers have carried out live experiments in different online social networks to observe the efficacy and potential of socialbots to influence network users and structure. Different researchers have come up with varied approaches for characterization, identification, and detection of socialbots [11, 12], but there are hardly approaches, except [25], that monitor the users who provide helping-hand to socialbots. Moreover, to the best of our knowledge, none of the existing approaches profile users who interacted with and/or targeted by the socialbots. However, users can be categorized based on their interaction and content behaviour with socialbots and they can be profiled accordingly.

In this paper, we propose an approach to categorize Twitter users into three groups – *active*, *reactive*, and *inactive* targets, based on their interaction behaviour with socialbots. All those users who follow socialbots without being followed by them are considered as active users. On the other hand, users who respond to the *following* socialbots are considered as reactive users, and those who do not show any interest in anonymous socialbots are considered as inactive users. To this end, we carried out a live experiment by injecting an army of 98 socialbots and collected both content and structural data for analyses and experimental evaluation purposes. Our proposed approach is evaluated on a sample dataset consisting of 749 users collected through the experiment. Profiling is the characterization of users with their personal and behavioural information that differentiate every individual from others [1]. Among the users characteristics, there are attributes that do not or hardly change with time, and such attributes form the static component of users profile, which are generally identity-related features including name, age, sex, Twitter handle, etc. Other significant component of users profile is based on their daily activities, interactions with other users in the network, forming behavioural and attitude-related component. Characteristics extracted from such data form the dynamic component of a user profile that grabs nature of the user reflecting conditional and temporal changes. In the proposed

approach, static and dynamic components of users profile are used to learn classifiers for classifying users as active, reactive, or inactive users. Topical distribution and favourite topics of the users are also explored and analysed. *Active* and *reactive* users are found to be frequent tweet posters with the majority of tweets inclined toward advertisement-related topics. Dominating features have also been identified using feature selection and ranking algorithms and among the dominating features *following rate* and *follower rate* are found to be the most dominating features across the ranking algorithms.

The rest of the paper is organized as follows. Section 2 presents a brief review of the existing literature on users profiling in online social networks in the context of socialbots. Section 3 presents detail of the proposed users' categorization and profiling approach. Section 4 presents statistics of the dataset used to evaluate the proposed approach. It also presents evaluation of the experimental results using different performance measure metrics. Finally, section 5 summarizes the paper in addition to highlighting possible future directions.

2 RELATED WORKS

2.1 Online Social Network and Socialbots

In literature, various experiments have been performed to observe socialbots' behaviour, their impact in terms of infiltration, and extent to which they can manipulate and pollute the online social networks [4, 13, 27]. Socialbots can easily rise to one of the top influencers of an online social network without putting much effort as proved by the authors in [4] through a experiment in *aNobii* network. Experiment proved that even technologically aware users get trapped to socialbots' social engineering tactics [13]. In [8], authors thoroughly analysed the economic feasibility of running socialbots campaign and presented the inherent vulnerabilities that are generally exploited by the socialbots. It has also been observed that socialbots can easily get followers with little efforts [25].

2.2 User Profiling and Socialbots Detection

User profiling is the characterization and identification of attributes to represent objects and human-beings. With the evolution of online social networks and availability of big amount of data, researchers conceive various user profiling strategies and used it for characterization and detection of malicious entities such as spammers, bots, spambots, socialbots, etc. [2, 3, 6, 24]. Authors in [22] profiled Twitter users and classified them into three broad categories – broadcast, consumption, and spam bots. In [23], authors predicted users political orientation and ethnicity using linguistic and profile features along with the topical distribution of the users to observe their interests. To filter users time-line as per their interest, an interest-based profiling and filtering approach has been proposed in [14]. It identifies interests of the users by analysing contents of the pages redirected by the URLs used in the users' tweets. In existing literature, researchers have also characterized content polluters, spammers, and bots and proposed detection approaches for various online social networks [9, 11, 20]. Profiles that are compromised by malwares or other means have been profiled using the profile's behaviour analysis through– browsing sequence,

first activity performed after login, and so on. Another sequence-based approach inspired from DNA sequencing is used in [12] for profiling individual activities that are grouped-by activity-type and activity-sequence to segregate social spambots from benign users. Another very interesting experiment has been carried out in [25] to identify users who are susceptible to reply the socialbots, either by following them back or through messaging them. But, there is no study to profile users who follow socialbots without being followed by the socialbots, users who follow back the socialbots along with the users who do not follow back or respond even though socialbots followed them.

3 PROPOSED APPROACH

In this section, we present the functional details of our proposed approach for profiling susceptible users who can be the victim of social engineering traps of socialbots. Since trapped users, either active or reactive, are responsible for socialbots' trust building and success in online social networks, their characterization and profiling is crucial to observe how they differ from other users. In real-world scenario, two types of identity information are associated with every user for their recognition. First category is the users physical or implicit identities containing information like name, age, home-town, gender, etc. that hardly change with time, whereas the second category includes users behavioural and interactional characteristics, constituting personality-related information describing users behaviour, nears and dears, and so on. The second category of identity represents the dynamics in a user behaviour representing how it changes and evolves temporally and conditionally with time. Our proposed approach considers both static and dynamic components of identity for user profiling. A work flow of the proposed approach for socialbots targets profiling and identification is shown in Fig. 1. Before presenting a detailed discussion of static and dynamic profiles, a brief description of different categories of socialbots' targets profiled by the proposed approach is given in the following paragraphs, where $N(U, E)$ represents the socialbots injected network, U is the set of users comprising both socialbots S and benign users B , and E is the set of connections between the users. Obviously, $S \subset U$ and $B \subset U$.

Active targets: In a socialbots injected network $N(U, E)$, a user $u_i \in B$ is considered as an *active target* of a socialbot $s_j \in S$ if u_i starts following s_j without any initiation from s_j . Such users are called active as they are always ready to follow anyone without any familiarity or verification.

Reactive targets: In a socialbots injected network $N(U, E)$, a user $u_i \in B$ is considered as a *reactive target* of a socialbot $s_j \in S$ if u_i starts following s_j in response to its being followed by s_j . Such users are called *reactive* due to the fact that they need some push-up action to get activated and trapped.

Inactive targets: In online social networks, genuine or benign users generally respond to only those whom they know. In contrast, malicious users are always ready to follow and respond to any one to increase their followers and consequently network reachability. In a socialbots injected network $N(U, E)$, if a socialbot $s_j \in S$ follows a user $u_i \in B$ and u_i neither follows back nor sends direct messages

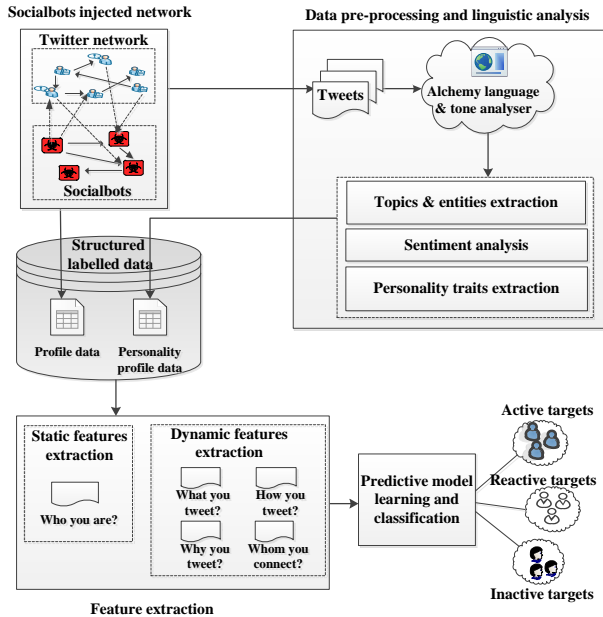


Figure 1: Workflow of the proposed approach for socialbots' targets profiling and identification

to the socialbot s_j then the user u_i is called an *inactive target* of the socialbot s_j .

As discussed earlier, profile of a user u consists of both static S_u and dynamic \mathcal{D}_u components. A brief discussion about these components is presented in the following sub-sections.

3.1 Static Profile

For characterizing users, information related to their implicit identity such as name, sex, date of birth, etc. either do not change or rarely change with time. In Twitter and other online social networks, users are asked to provide certain personal details while registering for profile creation, which is generally used in the network for their recognition by other users. Though the information provided by the users may be wrong, there is no universal mechanism to insure the authenticity of the provided information. However, online social networks use different mechanism to shield and secure their network from malicious users and bots. But above all, online social networks still have vulnerabilities that are exploited by the malicious users for fake profiles creation and other malicious activities [7]. Based on user-supplied information during profile creation, *online identity: Who you are?* component described in the following sub-section, is considered as a static constituent of the user profiles.

3.1.1 Online Identity: Who you are? In online social networks, whenever a friend request is received by a user, receiver verifies the sender using the public information available from the sender's profile. Unlike other online social networks, a Twitter user does not need to provide much information for account registration and profile is created with limited information. In Twitter, there are

users who blindly follow anyone without any restriction and reason, and such users are most pleasing and helpful for the socialbots. Similarly, when socialbots follow users, some of them blindly follow back, whereas aware and conscious users avoid such deceiving requests. Users who follow back the socialbots are either trapped by socialbot's profile appearance or generally they themselves are bogus, fake, or follower seekers. However, sometimes users follow back socialbots due to topical similarity or social etiquette. Here challenge is to profile *active* and *reactive* users based on the information available from their profiles and analyse how they differ from the users who avoid follower requests initiated by the socialbots. Static component of user profiling is a set of 9 attributes viz - $\{twitter\ age\ (S_u^1),\ geo\text{-}enabled\ status\ (S_u^2),\ profile\ description\ length\ (S_u^3),\ special\ character\ count\ in\ profile\ description\ (S_u^4),\ special\ character\ ratio\ in\ profile\ description\ (S_u^5),\ profile\ image\ status\ (S_u^6),\ handle\ length\ (S_u^7),\ special\ character\ ratio\ in\ handle\ (S_u^8),\ name\ and\ handle\ similarity\ (S_u^9)\}$, with major emphasis on profile description and Twitter-handle chosen by the users. Hence, the static profile of a user u , S_u , can be defined as a nine-dimensional real-valued vector as given in equation 1

$$S_u = \{S_u^1, S_u^2, \dots, S_u^9\} \quad (1)$$

3.2 Dynamic Profile

Users' implicit attributes are used for their identity, but in a network of discussion other users recognize and judge them by their behaviour. In the proposed approach, dynamic profile of a user covers his/her behavioural aspects in the network, such as whom the user interacts, contents used in interaction, why he/she connects and interacts, topics discussed in the interactions, and so on. Due to complex behavioural dynamics, it is very difficult to profile these aspects of a user. In online social media platforms, it is even more difficult due to various constraints such as informal writing practices, data granularity and unavailability, lack of efficient multilingual natural language processing tools, etc. In the proposed approach, dynamic profile of a user u on Twitter is represented as \mathcal{D}_u and consists of four components - "textual preference: what you tweet?" as $\mathcal{D}_u(\mathcal{T})$, "interaction methods: how you tweet?" as $\mathcal{D}_u(\mathcal{I})$, "user intention and personality: why you tweet?" as $\mathcal{D}_u(\mathcal{P})$, and "network structure: whom you connect?" as $\mathcal{D}_u(\mathcal{N})$, as given in equation 2. Further details about these dynamic profile constituents are given in the following sub-sections.

$$\mathcal{D}_u = \mathcal{D}_u(\mathcal{T}) \cup \mathcal{D}_u(\mathcal{I}) \cup \mathcal{D}_u(\mathcal{P}) \cup \mathcal{D}_u(\mathcal{N}) \quad (2)$$

3.2.1 Textual Preference: What you tweet? Individuals differ with each other in terms of content usage, language preference, and writing skill. People from different walk of life have disparate writing-style and talking behaviour, e.g., journalists and bloggers generally write long sentences expressing their views about current affairs, news portals share news-links, advertisers talk about products and services, new ventures and start-ups frequently use URLs, and young people share media stuffs, and so on. *Textual preference* component of dynamic profiling captures the above discussed characteristics of individual using different features. This component has six attributes reflecting linguistic preference of a user u - *tweet*

similarity ($\mathcal{D}_u^1(\mathcal{T})$), average tweet length ($\mathcal{D}_u^2(\mathcal{T})$), tweet length variance ($\mathcal{D}_u^3(\mathcal{T})$), media ratio ($\mathcal{D}_u^4(\mathcal{T})$), advertising keyword ratio ($\mathcal{D}_u^5(\mathcal{T})$), and URL ratio ($\mathcal{D}_u^6(\mathcal{T})$). Hence, the textual preference component $\mathcal{D}_u(\mathcal{T})$ of a user dynamic profile can be modelled as a six-dimensional real-valued vector as given in equation 3.

$$\mathcal{D}_u(\mathcal{T}) = \{\mathcal{D}_u^1(\mathcal{T}), \mathcal{D}_u^2(\mathcal{T}), \dots, \mathcal{D}_u^6(\mathcal{T})\} \quad (3)$$

3.2.2 Interaction Method: How you tweet? In literature, various spammer, bot, and socialbot detection techniques have exploited interaction-based characterization of the users, e.g., some users frequently tweet and retweet, some users are always active but hardly tweet and use the Twitter as an information and news source, whereas some users generally retweet [5, 10, 26]. In a user profile, the interaction-based component translates the interaction behaviour of the user, rather than his/her access methods. As a result, the interaction-based component for a Twitter user profile consists of three features – tweet rate ($\mathcal{D}_u^1(I)$), retweet rate ($\mathcal{D}_u^2(I)$), and tweets languages count ($\mathcal{D}_u^3(I)$), and it can be represented as a 3-dimensional real-valued vector as given in equation 5

$$\mathcal{D}_u(I) = \{\mathcal{D}_u^1(I), \mathcal{D}_u^2(I), \mathcal{D}_u^3(I)\} \quad (4)$$

3.2.3 User Intention and Personality: Why you tweet? Twitter is used by different users for varied reasons and intentions. Some people are on Twitter for entertainment, few for connecting to their nears and dears along with their favourite celebrities, and companies have profiles to connect and update their customers. However, generally Twitter users use it for real-time update of news and events, status of politician and celebrities views on different events and so on [19]. Observing and capturing the intention of Twitter users can help in finding the answers of various unanswered questions such as why users respond or avoid follower request from socialbots or other anonymous profiles. In this dynamic profile component, we analyse the big-five personality traits and attitude of the users through content analysis of their tweets [16]. Big-five personality traits and emotional aspects of the users have been identified using *Tone Analyzer*¹, a very powerful service to assign numeric score between 0 and 1 to users’ attitude and big-five personality traits based on the contents provided by the user. Entities and topics from the tweets of each user are extracted for a maximum of 200 tweets as it is enough to reflect user’s interests. In order to track users aspiration to join the Twitter, topical and interest space analyses of the users are vital. Topics and entities are extracted using *Alchemy Language*², a very powerful service for natural language processing. Alchemy extracts topics and entities from input tweets and each topic itself is a hierarchy of sub-topics. It also assigns relevance score to each extracted topic and entity showing their importance in the tweet. Suppose, for a tweet t_i of user u , alchemy extracts topics and entities in the form of $T_u^i(\tau, \omega)$ and $E_u^i(\varepsilon, \omega)$ respectively, where $T_u^i(\tau)$ represents topic set of i^{th} tweet of user u and $T_u^i(\omega)$ represents corresponding relevance score set. $T_u^i(\tau)$ is the set of n topics as shown in equation 5, where $T_u^i(\tau_j)$ is the j^{th} topic of the i^{th} tweet of u . Further, each topic has hierarchical levels of sub-topics, e.g., hierarchical representation

of the topic $T_u^i(\tau_j)$ is shown in equation 6, where $T_u^i(\tau_j^m)$ is the m^{th} -level sub-topic of the j^{th} topic $T_u^i(\tau_j)$ of u . Whereas in case of entities $E_u^i(\varepsilon, \omega)$, everything is same as of the topics except that entities do not have hierarchical organization that is there is no sub-entity for an entity $E_u^i(\varepsilon_j)$.

$$T_u^i(\tau) = \sum_{j=1}^n T_u^i(\tau_j) \quad (5)$$

$$T_u^i(\tau_j) = T_u^i(\tau_j^1) \supset T_u^i(\tau_j^2) \supset \dots \supset T_u^i(\tau_j^m) \quad (6)$$

User intention and personality component of a user u is composed of 14 attributes – topic count ($\mathcal{D}_u^1(\mathcal{P})$), topic ratio ($\mathcal{D}_u^2(\mathcal{P})$), mean topic weightage ($\mathcal{D}_u^3(\mathcal{P})$), entity count ($\mathcal{D}_u^4(\mathcal{P})$), entity ratio ($\mathcal{D}_u^5(\mathcal{P})$), mean entity weightage ($\mathcal{D}_u^6(\mathcal{P})$), positive to negative sentiment ratio ($\mathcal{D}_u^7(\mathcal{P})$), sentiment orientation ($\mathcal{D}_u^8(\mathcal{P})$), anger ($\mathcal{D}_u^9(\mathcal{P})$), fear ($\mathcal{D}_u^{10}(\mathcal{P})$), joy ($\mathcal{D}_u^{11}(\mathcal{P})$), sadness ($\mathcal{D}_u^{12}(\mathcal{P})$), disgust ($\mathcal{D}_u^{13}(\mathcal{P})$), and dominant character ($\mathcal{D}_u^{14}(\mathcal{P})$). It is denoted as $\mathcal{D}_u(\mathcal{P})$ and represented using equation 7.

$$\mathcal{D}_u(\mathcal{P}) = \{\mathcal{D}_u^1(\mathcal{P}), \mathcal{D}_u^2(\mathcal{P}), \dots, \mathcal{D}_u^{14}(\mathcal{P})\} \quad (7)$$

3.2.4 Network Structure: Who are connected? In addition to personal and textual features, network structure is also very vital as ultimate goal of any malicious actor whether it is spammer, spam-bot, or socialbot is to maximize the infiltration space. To achieve this goal, socialbots try to gain maximum number of followers and friends using different tactics. This component aims to observe connection forming behaviour of all three categories of users. *Network structure* component for the dynamic profile of a user u is composed of four attributes – follower rate ($\mathcal{D}_u^1(\mathcal{N})$), following rate ($\mathcal{D}_u^2(\mathcal{N})$), follower granularity to following granularity ratio ($\mathcal{D}_u^3(\mathcal{N})$), and number of users mentioned in tweets ($\mathcal{D}_u^4(\mathcal{N})$), and accordingly it can be represented as a four-dimensional real-valued vector as given in equation 8.

$$\mathcal{D}_u(\mathcal{N}) = \{\mathcal{D}_u^1(\mathcal{N}), \mathcal{D}_u^2(\mathcal{N}), \mathcal{D}_u^3(\mathcal{N}), \mathcal{D}_u^4(\mathcal{N})\} \quad (8)$$

4 EXPERIMENTAL SETUP AND RESULTS

This section provides details of our data collection process, analysis techniques, classifiers learning, feature ranking, and performance comparison results. Further details about the experimental results are provided in the following sub-sections.

4.1 Data Collection

To retrieve data associated to all three category of users – *active*, *reactive*, and *inactive*, a live experiment was carried out in Twitter. To this end, an army of 98 socialbots related to the top-six Twitter using countries were injected in Twitter, and their all activities such as following, tweeting, etc. were programmed and performed randomly. Whole network of the socialbots was active for a period of 28 days until suspended by the Twitter. Statistical analysis of the crawled data showed some interesting results reported in [15]. Thereafter, we crawled the time-line and profile information of the followed users and trapped followers. Crawled users have been grouped into three categories – *active*, *reactive*, and *inactive*. From crawled users, we randomly selected a total of 749 users,

¹<https://www.ibm.com/watson/developercloud/tone-analyzer.html>

²<https://www.ibm.com/watson/developercloud/alchemy-language.html>

comprising 262 *active*, 261 *reactive*, and 226 *inactive* users, for establishing the efficacy of our proposed approach.

4.2 Results and Discussion

This section presents different aspects of our analyses results including topical and personality analysis, classifier performance evaluation, and dominant features ranking.

4.2.1 Topical and Interest Distribution. Why users are registered on Twitter, i.e., the intentions of users behind joining Twitter can be captured by analysing their interest space. *User intention and personality* component $\mathcal{D}_u(\mathcal{P})$ of the dynamic profile captures this aspect of a user's personality. Once topic extraction from tweets of individual users is completed, topics extracted from each users tweets are sorted in decreasing order of their relevance score and top-10 topics are selected. Each topic has a hierarchical relationship starting from abstract level of the topic to a specific topic. For example, if *shopping* is a topic then the relevant subtopics can be *footwear, e-commerce, clothing*, etc. and each subtopic may further have subtopics. During topic analysis, it is observed that active users talk more frequently about computer accessories, internet technologies, finance, banking sector, education, etc. as shown in figures 2(a) and 3(a), where vertical axis represents the topic frequency in the tweets of all three groups of users. On analysis, it is found that *active* and *reactive* users post, on average, 8 and 11 tweets per day in contrast to *inactive* users of just 3 tweets per day. *Active* and *reactive* users also show high URL ratio of 0.43 and 0.37 per tweet respectively against 0.29 per tweet by *inactive* users. Similarly *active* and *reactive* users also show higher images and videos usage rate in their tweets. Malicious users detection approaches in literature consider high rate of all these mentioned parameters as suspicious. So, it can be inferred that *active* and *reactive* users are very engaging and suspicious users. In addition, the two categories of users also show interest in obscene topics. Topical and sub-topical distribution for all three categories of users can be seen in figures 2 and 3, respectively. *Reactive* users behaviour show some deviation when critically observed at subtopic level, and it is found that *reactive* users are not as frequent as *active* users. During attitude and personality analysis, it is found that *joy, sadness, and agreeableness* are the most relevant and dominating personality factors. These parameters reflect motive behind the use of online social networks by their users. Dominance of *agreeableness* among the personality traits shows that the social etiquette prevalence among Twitter users and it is exploited by the socialbots to gain followers and favourites.

4.2.2 Classifier Learning and Performance Evaluation. After extraction of static and dynamic components of users profile based on their tweets and interactions data, three machine learning classifiers namely – Naive Bayse, Reduced Error Pruning Decision Tree, and Random Forest are learned as the ternary classifiers as these are capable of handling multi-class problems. Machine learning classifiers are learned using the Weka tool³ which is very handy and implements machine learning algorithms in Java. To avoid data biasness, 10-fold cross validation is used to evaluate the performance of the classifiers. In 10-fold cross validation, the whole

dataset is divided into 10 parts out of which 9 parts are used for classifier learning and one part is used for testing purpose. This process is repeated 10 times utilizing each instance of the dataset in training as well as in testing. The performance of the classifiers is evaluated using five metrics – *True Positive Rate (TPR)*, *False Positive Rate (FPR)*, *Precision*, *Recall*, and *F-Measure*. Naive Bayes classifier under useSupervisedDiscretization=TRUE setting shows moderate accuracy with *TPR* as 56.7%, which is shown in table 1. Among the three classifiers, *Random Forest* under the default settings shows the best performance with *TPR* as 60.6%. The performance evaluation results for the all classifiers are given in the table 1. It can be observed from this table that the performance of classifiers does not seem to be appealing, which is mainly due to modeling the problem as a three-class problem as attributes' discrimination power reduces with increasing number of classes. However, in comparison to one of the existing state-of-the-arts in [25] where the same problem is modelled as a two-class problem with highest *TPR* as 68%, our result is not much discouraging. On analysis, we observed that the performance is downgraded due to poor classification of *reactive* users that have high degree of similarity with *active* users.

Therefore, in order to have a true performance comparison our proposed approach with the existing approach, we modelled the classification problem as a two-class problem, wherein *active* and *reactive* categories are merged into a single category and termed as *trapped users*. Thereafter, the same classification and evaluation process discussed earlier is applied over the modified data set. Table 2 presents the performance evaluation results for all three classifiers based on various metrics discussed earlier in this section. It can be observed from this table that the performance of classifiers is significantly improved when the problem is modelled as a 2-class problem and results are significantly better than the state-of-arts compared approach [25]. This proves the fact that there is significant similarity in the working behaviour of the *active* and *reactive* users and proposed approach is better than state-of-arts approaches. Among the classifiers, *Random Forest* again proves to be the best classifier with the true positive rate of 80.2%. As shown in table 2, the *FPR* for all classifiers is high, which is mainly due high *FPR* for *inactive* targets.

4.2.3 Features Ranking. In order to identify the dominating features in terms of the discriminative power, four feature selection algorithms – Mutual Information (MI) [21], ReliefF [18], Correlation Attribute Evaluation (CAE) [17], and Gain Ratio are considered in our experiment. Mutual information is the mutual dependence between two random variables, and it is based on joint probability distribution between two random variables. ReliefF finds the closest instances of the same and different class against the one of the picked instance. This procedure is repeated until the closest instance for same class and different class for all the instances is found. Closest same class instance is called *nearest-hit* and closest different class instance is called *nearest-miss*. Correlation attribute feature selection algorithm is based on Pearson's correlation coefficient between attributes and class labels. Finally, gain ratio is the ratio of information gain to the split information value. It overcomes the biasness towards multi-valued attributes as in the case of using information gain. Table 3 presents the list of top-10 highly ranked features. It can be observed from this table that *following*

³<http://www.cs.waikato.ac.nz/ml/weka/>

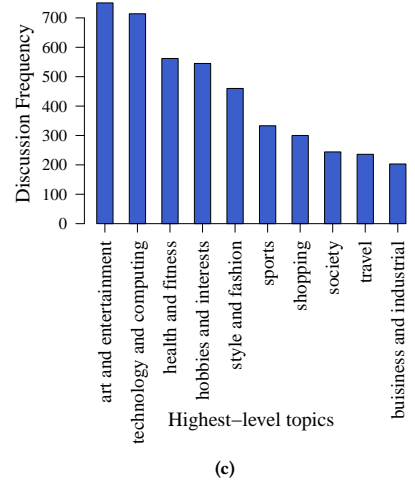
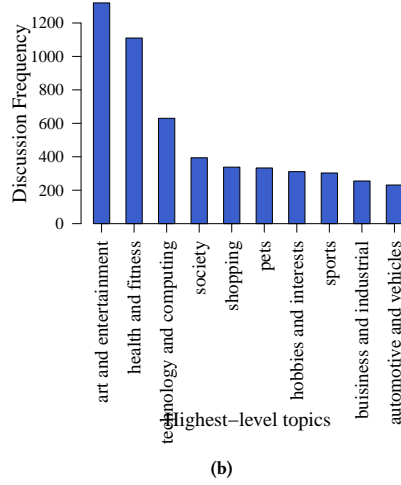
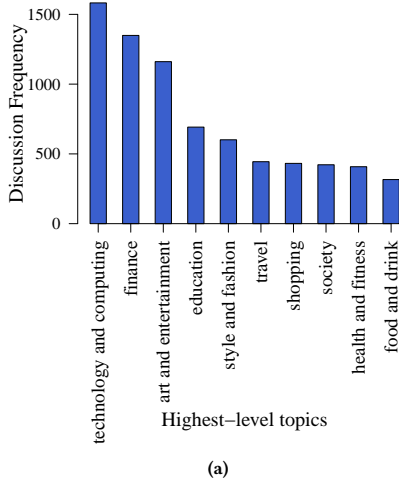


Figure 2: Frequency distribution of top-10 highest-level topics discussed by all three groups of users – (a) active users, (b) reactive users, and (c) inactive users

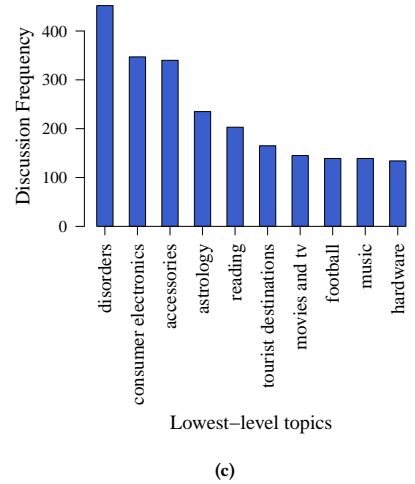
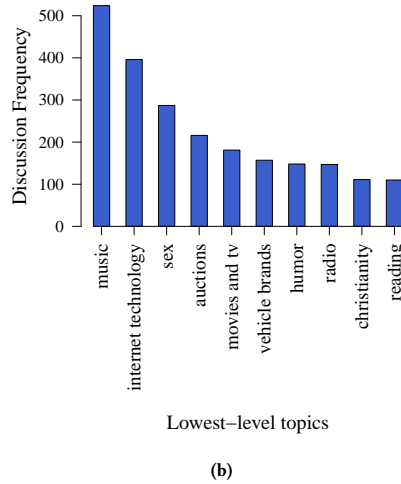
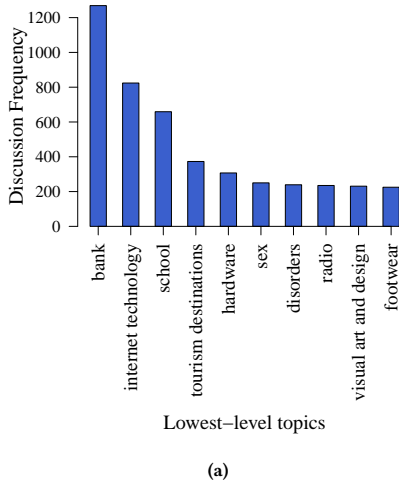


Figure 3: Frequency distribution of top-10 lowest-level topics discussed by all three groups of users – (a) active users, (b) reactive users, and (c) inactive users

Table 1: Performance evaluation results of the classifiers for 3-class problem

Classifier	TPR	FPR	Precision	Recall	F-Measure
Naive Bayes	0.567	0.214	0.564	0.567	0.564
Reduced Error Pruning Decision Tree	0.562	0.213	0.559	0.562	0.558
Random Forest	0.606	0.200	0.606	0.606	0.605

rate and follower rate are the most dominating features. Moreover, it is obvious from the results shown in this table that static profile component does not play much significant role in discriminating

different categories of users, whereas topical features are found to be relevant.

Table 2: Performance evaluation results of the classifiers for 2-class problem

Classifier	TPR	FPR	Precision	Recall	F-Measure
Naive Bayes	0.764	0.313	0.766	0.764	0.765
Reduced Error Pruning Decision Tree	0.752	0.414	0.738	0.752	0.740
Random Forest	0.802	0.367	0.797	0.802	0.789

Table 3: Top-10 features selected by three different feature ranking algorithms

Rank	Ranking algorithm			
	Mutual Information	Relieff	Correlation Attribute Evaluation	Gain ratio
1	Following rate	Retweet ratio	Following rate	Follower rate
2	Follower rate	Media ratio	Mean entity relevance	Following rate
3	Tweets similarity	Prodesc_length	Media ratio	Handle_spechar_ratio
4	Fol_flg ratio	Sent_orientation	Handle_spechar ratio	Tweet rate
5	Media ratio	Anger	Entity ratio	Tweets similarity
6	Twitter age	Handle_spechar_ratio	Prodesc_length	Tweet age
7	Mean entity relevance	Twitter age	Tweets similarity	Media ratio
8	Anger	Entity ratio	Follower rate	Topic count
9	Entity count	Tweet_lang_count	URL ratio	Entity count
10	Entity ratio	Mean_entity_rele	Entity count	Fol_flg ratio

5 CONCLUSION AND FUTURE WORKS

In this paper, we have presented a supervised machine learning approach to classify socialbots' targets into three categories – *active*, *reactive*, and *inactive* targets, based on their interaction behaviour with socialbots. The classification problem is also modelled and studied as a two class problem, wherein *active* and *reactive* users are merged into a single category, termed as *trapped users*, due to similarity in their working behaviour. We have also presented a user profiling approach where profile of a user is generated using static and dynamic components representing their identical and behavioural aspects, respectively. As a future work, the proposed approach can be evaluated over larger real-life datasets from different online social networks. Moreover, analysing temporal and topical evolution of users also seems to be one of the interesting future directions of work.

REFERENCES

- [1] Gediminas Adomavicius and Alexander Tuzhilin. 1999. User Profiling in Personalization Applications through Rule Discovery and Validation. In *Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining*. ACM, San Diego, USA, 377–381.
- [2] Faraz Ahmad and Muhammad Abulaish. June 25-27, 2012. An MCL-Based Approach for Spam Profile Detection in Online Social Networks. In *Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE-TrustCom)*. IEEE Computer Society, Liverpool, UK, 602–608.
- [3] Faraz Ahmed and Muhammad Abulaish. 2013. A Generic Statistical Approach for Spam Detection in Online Social Networks. *Computer Communications* 36, 10-11 (2013), 1120–1129.
- [4] Luca Maria Aiello, Martina Deplano, Rossano Schifanella, and Giancarlo Ruffo. 2012. People are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. In *Proceedings of the 6th International Conference on Weblogs and Social Media*. AAAI Press, Dublin, Ireland, 10–17.
- [5] Nutan Reddy Amit A Amleshwaram, Suneel Yadav, Guofei Gu, and Chao Yang. 2013. CATS: Characterizing Automation of Twitter Spammers. In *Proceedings of the 5th International Conference on Communication Systems and Networks (COMSNETS)*. IEEE Computer Society, Bangalore, India, 1–10.
- [6] Sajid Y. Bhat and Muhammad Abulaish. 2014. Communities against Deception in Online Social Networks. *Computer Fraud and Security* 2014, 2 (2014), 8–16.
- [7] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. The Socialbot Network: When Bots Socialize for Fame and Money. In *Proceedings of the 27th Annual Computer Security Applications Conference*. ACM, Orlando, Florida USA, 93–102.
- [8] Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2013. Design and Analysis of Social Botnet. *Computer Networks* 57, 2 (2013), 556–578.
- [9] Yazan Boshmaf, Matei Ripeanu, Konstantin Beznosov, and Elizeu Santos-Neto. 2015. Thwarting Fake OSN Accounts by Predicting their Victims. In *Proceedings of the 8th Workshop on Artificial Intelligence and Security*. ACM, Denver, USA, 81–89.
- [10] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. Who is Tweeting on Twitter: Human, Bot, or Cyborg?. In *Proceedings of the 26th Annual Computer Security Applications Conference*. ACM, Austin, Texas, USA, 21–30.
- [11] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2012. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? *IEEE Transactions on Dependable and Secure Computing* 9, 6 (2012), 811–824.
- [12] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2016. DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. *IEEE Intelligent System* 31, 5 (2016), 58–64.
- [13] Aviad Elyashar, Michael Fire, Dima Kagan, and Yuval Elovici. 2013. Homing Socialbots: Intrusion on a Specific Organization's Employee using Socialbots. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*. IEEE Computer Society/ACM, Niagara Falls, Canada, 1358–1365.
- [14] Sandra Garcia Esparza, Michael P. OfiMahony, and Barry Smyth. 2013. CatStream: Categorising Tweets for User Profiling and Stream Filtering. In *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, Santa Monica, CA, USA, 25–36.
- [15] Mohd Fazil and Muhammad Abulaish. 2017. Why a Socialbot is Effective in Twitter? A Statistical Insight. In *Proceedings of the 9th International Conference on Communication Systems and Networks (COMSNETS), Social Networking Workshop*. IEEE Computer Society, Bengaluru, India, 562–567.

- [16] Lewis R Goldberg. 1993. The Structure of Phenotypic Personality Traits. *American Psychologist* 48, 1 (1993), 26–34.
- [17] Mark A. Hall. 1999. *Correlation-based Feature Selection for Machine Learning*. Ph.D. Dissertation. The University of Waikato, New Zealand.
- [18] Igor Kononenko. 1994. Estimating Attributes: Analysis and Extensions of RELIEF. In *Proceedings of the European Conference on Machine Learning*. Springer, Berlin, Heidelberg, Italy, 171–182.
- [19] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a Social Network or a News Media?. In *Proceedings of the 19th International Conference on World Wide Web*. ACM, Raleigh, North Carolina, USA, 591–600.
- [20] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. ACM, Santa Monica, CA, USA, 185–192.
- [21] Tom Mitchell. 1997. *Machine Learning*. McGraw Hill.
- [22] Richard J. Oentaryo, Arinto Murdopo, Philips K. Prasetyo, and Ee-Peng Lim. 2016. On Profiling Bots in Social Media. In *Proceedings of the International Conference on Social Informatics*. Springer, Bellevue, WA, USA, 92–109.
- [23] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the 5th International Conference on Weblogs and Social Media*. AAAI Press, Barcelona, Spain, 281–288.
- [24] Muhammad Z. Rafique and Muhammad Abulaish. August 27–31, 2012. Graph-Based Learning Model for Detection of SMS Spam on Smart Phones. In *Proceedings of the 8th International Wireless Communications and Mobile Computing Conference (IWCMC'12) fi Trust, Privacy and Security Symposium*. IEEE Computer Society, Limasol, Cyprus, 27–31.
- [25] Randall Wald, Taghi M. Khoshgoftaar, Amri Napolitano, and Chris Sumner. 2013. Which Users Reply to and Interact with Twitter Social Bots?. In *Proceedings of the 25th International Conference on Tools with Artificial Intelligence*. IEEE Computer Society, Herndon, VA, USA, 135–144.
- [26] Chao Yang, Robert Harkreader, and Guofei Gu. 2013. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Transactions on Information Forensics and Security* 8, 8 (2013), 1280–1293.
- [27] Jinxue Zhang, Rui Zhang, Yanchao Zhang, and Guanhua Yan. 2012. On the Impact of Social Botnets for Spam Distribution and Digital Influence Manipulation. In *Proceedings of the 6th International Conference on Communications and Network Security*. IEEE Communications Society, National Harbor, MD, USA, 46–54.