Visualization of the Social Bot’s Fingerprints

Mehmet Kaya
Electrical and Electronics Engineering Dept.
Adiyaman University
Adiyaman, Turkey
m.kaya@adyiyaman.edu.tr

Shannon Conley
Department of Software Engineering
Firat University
Elazig, Turkey
shconley@syr.edu

Asaf Varol
Department of Software Engineering
Firat University
Elazig, Turkey
varol.asaf@gmail.com

Abstract— As the number of social media users increases for platforms such as Twitter, Facebook, and Instagram, so does the number of bot or spam accounts on these platforms. Typically, these bots or spam accounts are automated programatically using the social media site's API and attempt to convey or spread a particular message. Some bots are designed for marketers trying to sell products or attract users to new sites. Other types of bots are much more malicious and disseminate misinformation that harms or tricks users. Such bots (fake accounts) may lead to serious consequences, as people’s social network has become one of the determining factors in their general decision making. Therefore, these accounts have the potential to influence people’s opinions drastically and hence real life events as well. Through different machine learning techniques, researchers have now begun to investigate ways to detect these types of malicious accounts automatically. To successfully differentiate between real accounts and bot accounts, a comprehensive analysis of the behavioral patterns of both types of accounts is required. In this paper, we investigate ways to select the best features from a data set for automated classification of different types of social media accounts (ex. bot versus real account) via visualization. To help select better feature combinations, we try to visualize which features may be more effective for classification using self-organizing maps.

Keywords— Social Bot Detection; Self Organizing Maps; Social Media Content Polluters

I. INTRODUCTION

With popular social networking services, interacting and communicating with others and sharing information or opinion within groups have become easier than ever. The way we access information and communicate with others is only one of many things that social media has changed in our lives. People are engaging in social media to such an extent that the content shared on their networks creates great impact on their lives and greatly influences their decision making. The sheer volume of users on these platforms and their power to spread content effortlessly attract many people who want to reach out to large audiences. Along with companies trying to promote services or advertise products and politicians mustering support for their campaigns, there is another key social media group attempting to influence crowds: malicious users. Global widespread usage of social media and the high density of social networks have created weak spots in these online social systems that can easily be exploited by cyber criminals [1].

A key technique for influencing large numbers of social media users is the use of multiple fake accounts or bots that masquerade as real users, but are controlled via software automation. Advances in artificial intelligence (AI) allow programmers to create automated systems (also known as bots, software robots, or social bots) that are capable of seamlessly tracking other users, interacting with them, and sharing or obtaining content based on specific criteria, e.g., a key word or user name. In other words, with the current technology, these bots are now able to exhibit human-like behavior or pose as real social media users [1]. Besides polluting the social media environment for legitimate users by generating useless content, these social bots may misuse personal information and harm users through “persuading, smearing, or deceiving.” [2] Research shows that a user's personal information such as phone number, email address, current city, gender, birth date etc. can easily be infiltrated through social bots [3]. Moreover, these bots may share misleading information that others rely on and change their opinion about something that may have far greater implication on society as a whole such as influencing the outcome of governmental elections [4, 5, 6]. Therefore, it is of critical importance to detect such accounts in order to keep social media a “secure”, "politically-neutral", "nuisance-free" place.

Automated detection systems using machine learning techniques are the most practical way to achieve this goal. Several classification/clustering algorithms have already been used to detect such accounts on different social media platforms such as Twitter and YouTube. However, there are two main factors that hinder the accuracy and performance of these algorithms: the absence of a comprehensive datasets for training/learning and the need for more effective or robust discriminative features for classification.

Recently, as a result of their comprehensive study, Lee et al have generated a dataset of Twitter content polluters (social bots) and legitimate users, the Social Honeypot Dataset, which includes data on 40,000 Twitter accounts in total. Knowing the feature sets that are relevant and discriminative has a great impact on the classification performance [7]. Therefore, before any classification process an analysis of features should be carried out to observe which ones might be more effective to obtain more accurate clustering. A self-Organizing map (SOM) [8], a special type of neural network model [9], can serve as an instrument to analyze different feature sets via visualization. Thus, this work aims to visualize the effect of different features on classification through SOM. The feature set is extracted from the Social Honeypot Dataset and analyzed to determine which one(s) should be used for final classification.
The rest of this paper is organized as follows. Chapter 2 discusses the related work including the Social Honeypot Dataset. In Chapter 3, a brief overview of SOM is given. Chapter 4 provides the results and discussions for our visualization of different input features. Finally, Chapter 5 gives some concluding remarks and lays out the future direction of this research.

II. RELATED WORK

This paper involves to two main research areas. First, it involves the automated social bot detection using machine learning. It is also related to self-organizing maps and input vector selection. We discuss briefly some of the works conducted in each one of these fields.

As stated above, it is compulsory to identify social bots in order to prevent possible cybercrimes or disturbances on social media sites. There is ongoing research for this purpose. In [2], previous work on automatic social both detection is grouped into three categories. These are as follows: bot detection based on social network information, crowdsourcing, and machine learning.

Based on the assumption that fake accounts have very small numbers of connection to real users, several network based approaches for the identification of fake accounts are proposed [10, 11, 12]. A crowdsourcing approach with “optimal” detection of fake accounts on online social networks like Facebook is presented in [13]. Based on statistical data on accounts’ behaviors, a web based application is developed to determine the probability of a particular account being a bot [2]. Analyzing the differences between behavioral patterns of human, bot, and cyborg (human assisted bot or bot assisted human) in terms of tweeting behavior, tweet content, and account properties, in [14] a classification algorithm is introduced. Using tweet contents and network information, in [4] a machine learning technique is used to detect accounts that disseminate misinformation.

Lee et al. have conducted a long-term project to collect information on social bots that can be used in the classification of Twitter accounts as bot or legitimate [15]. They set up sixty honeypots on Twitter – social bots designed to attract other bots and to collect information- and collected information for seven months. Their study provides not only this dataset, but also comparisons of different classification algorithms in terms of their accuracies. They concluded that their technique can detect social bots automatically with a 98.26% accuracy. We extracted some of the features from this data set to analyze via self-organizing map (SOM) in order to evaluate their power as input features to a classification algorithm.

III. THE SELF-ORGANIZING MAPS AND DATASET

A self-organizing map (SOM), a special type of neural networks, is a great tool for exploratory data analysis. It can be used to analyze the structural relationships between high dimensional multivariate data points and visualize these relationships in much lower (usually two) dimensions without supervision. Although it is closely related to artificial neural networks, SOM is mostly employed as a powerful data visualization technique rather than being used as a classification tool. SOM has a very broad application area from data exploration [16, 17], to information retrieval [18] and clustering [19]. “SOM is well suited for recognizing and classifying features in complex, multidimensional data” [19] and is an effective method to visualize high dimensional data by reducing its dimensions, while also maintaining/preserving the data’s topological relationship. Therefore, in this paper, we have chosen SOM to visually determine the features of social media account that can be used to discriminate between Twitter social bots and legitimate users.

A. Self-Organizing Map Structure

A self-organizing map is an unsupervised classification technique where the map adjusts itself to distinguish similar data from others. An example 4x4 SOM is given in Figure 2 where the input layer is fully connected to each node [20].

![Figure 1 SOM Example](image)

Each node in an SOM has a vector of weights and a topological position that represents its (x, y) coordinates on the lattice. The vector of weights for nodes has the same dimension as the input vectors. Therefore, the input vectors can be compared to the nodes to select the best matching node based on a distance measure using the nodes’ and input vectors’ weights. Once the best matching node is found, the area of the map including the best matching node and its neighbors is optimized. The optimized region therefore resembles and represents the cluster of the input data that the presented input vector is a member. The best matching unit compared to its neighbors has to be altered more so that this optimization brings the center of the selected area closer to that input vector. Over many iterations the map stabilizes with each region representing a cluster. The classification occurs based on the features (dimension for the vectors) and hence the SOM can be thought as the feature map of the input space.

The steps for SOM learning are listed as follows:

1. Nodes’ weights are initialized randomly.
2. A vector is selected randomly from the set of training data and presented to the map.
3. The best matching node is determined, the node with weights that are most like the input vector.
4. The neighborhood for the best matching node is calculated. The neighborhood usually starts large and shrinks each time-step.
5. The weights are adjusted for the nodes inside the neighborhood region (step 4) to make them resemble the
input vector. Weights of the nodes closer to the actual best matching node are adjusted more.

6. Repeat step 2 for maximum iterations.

B. Social Honeypot Dataset

Collected over a period of seven months, the Twitter account data set used in this paper comes from [15]. During their research, Lee et al. created 60 social bots to identify and collect information about other active bots’ identities and behaviors. Their bots were designed to attract other bots, while not communicating with anybody but each other. They have collected some 20,000 accounts that they thought were bots. These accounts were monitored for a while to make sure that they are indeed social bots basically relaying on the fact that they were not suspended by Twitter within a reasonably long period of time. They also used the “spam” key word that is used on Twitter to report this type of accounts and came up with their final list of accounts by comparing the identified bot accounts with these “spam-tagged” accounts. They have included some 20,000 legitimate accounts in this data set, as well. Besides providing this dataset, their paper also investigates the accuracies of some classification algorithms in WEKA for automated bot detection via the data set. And, also discuss which combination of features may give better results for these algorithms.

In this paper, we are using nine easily extractable features from this data set to visualize their possible effects on classification. We show that SOM can effectively be used for analysis of input feature selection. We do not discuss the final classification of social bots and legitimate users as this has already been done in [15]. We use the following features extracted from the dataset: Following (number of Twitter accounts a user is following), Follower (number of Twitter accounts following the user), #ofTweets (number of tweets generated by the user), AverageLengthOfTweets (average character length of a user's tweet), LengthOfScreenName (character length of the user's Twitter or screen name), LengthOfProfileDescription (character length of a user's profile description), PercentageOf@ (percentage of '@' characters used in the user's tweets), PercentageOf# (percentage of '#' character's used in the user's tweets), PercentageOfLink (percentage of hyperlinks found in the user's tweets).

IV. VISUALIZATION OF FEATURES WITH SOM

We have used the features described above in different combinations. Due to the page limit, we provide only some of them in this paper. We first analyzed the extreme case where only the Following and Follower information is available. For our analysis, we used R language and the kohenan package, which lets us observe the features and the SOM in several different ways and views. Figure 2 shows the training process for the Following-Follower feature combination. It shows the training process and how the SOM settled over many iterations (Fig. 2-A) and the distributions of data among nodes on the lattice (Fig. 2-B). The distribution of features in different feature-zones on the lattice (Fig. 2-C) is also provided.

Other than the views shown in Figure 2, this package provides another SOM view called heat maps. With this view we can observe the effect of a particular feature on different zones. This is the main view that helps us to determine which feature can classify the input data more clearly. The heat maps for Follower-Following feature combination is given in Figure 3.
The SOM heat map view basically computes for each node the average value of that particular feature from all the data points in the input space that were mapped to that particular node. Therefore, from a heat map, we can clearly see the feature zone for a particular feature where the data points in that zone exhibit a higher possession of that feature. This enables us to see which features are good at classifying the input data set and should be chosen for the final classification. For example, in Figure 3-B, we see that the Following feature can distinguish the data points in the lower-left corner of the map, while the heat map for the follower is more local and shows a smaller feature-zone (Figure 3-A). From the heat maps for these two features, one can conclude that Following is a feature with higher discriminative power.

Next, we included all nine features and run the SOM training again. Figure 4 shows the distribution of the number of points in each node and the influence of features on different zones as well as how the map has come to a settlement layout. From this figure (especially Figure 4-C), we can see where each feature is more dominant. This gives us a chance to see roughly which features should be visualized in more detail with the SOM heat map view. In other words, we could conjecture from Figure 4-C that four features may indicate a clearer separation/classification on the heat map. These four features are Following, AvrLinkCount (Percentage of Links in Tweets), AvrTweetLength, Following and LengthOfDescription and due to page limit we share heat maps only for these four features.

From the heat maps in Figure 5, we can easily see clusters with higher and lower values for a particular feature. AvrLinkCount, for instance, indicates clusters more explicitly than others (Figure 5-A). Therefore, based on these views we conclude that this feature should be considered as an input feature for the classification of social bots and legitimate users.
Finally, we have eliminated the features that we did not observe significant discriminative power. With the remaining four features in Figure 4, we have trained the SOM again and show the heat map. The new training process, node counts and the feature zones are given in Figure 6-A, Figure 6-B, and Figure 6-C, respectively. As we see from the Figure 6-A, the refined feature produces a more stable map. We can now see clusters directly from node counts and feature-zones as well.

As seen in Figure 6, we can now observe clearly how each one of these four features influences different zones on the map. For example, Figure 6-C shows that data with higher AvrLinkCount feature affects the upper left corner, while AvrTweetLength affects the upper-middle part of the map.

Here we should note that the Social Honeypot dataset includes tweets that we used to compute the average tweet lengths for accounts. Some of the tweets were in foreign languages which made our data extraction program count more characters than 140, which is the limit for tweets. Therefore, for some accounts average tweet lengths were determined to be greater than 140 (Figure 5-B).

V. CONCLUSION

With the widespread use of social networking sites, malicious activities on these platforms have increased, especially due to the advanced capabilities of social bots. Identifying and suspending these type of activities and accounts have turned into a major cyber security activity because of the potential harm that they can induce on the society by utilizing the strength of social media. Research gives promising results for detecting social bots automatically based on their observed behavior. Yet, with the advance of technology, these bots also improve their ability to mimic legitimate user behaviors and can easily go undetected. Therefore, automated identification of social bots does and will remain as an active research area and new algorithms and techniques will always be needed. In this paper, we have investigated possible input features in order to determine which ones should be selected to obtain a more accurate automated social bot detection system. We have shown the effectiveness of using SOM and different views for this analysis. For our future work, we plan to expand the feature set that we used for this study in order to come up with a more general approach that includes quantitative classification results as well. Although these features produce distinct clusters, an investigation is needed to determine what type of features are able achieve the desired discrimination or categorization (i.e. average user versus a spam bot versus a malicious bot).

REFERENCES

Visualization of the Social Bot’s Fingerprints


