Flexible Malicious Accounts Detector (FMAD) for Mining Twitter Social Network using Features and Accounts Frequent Pattern

Eman Osman
Information Systems Department
Faculty of Computers and Information
Helwan University

Mahmoud Mostafa
Assistant professor
Information Systems Department
Faculty of Computers and Information/ Helwan University
College of Computers and Information Technology/ Taif University

Sayed Abdel Gaber
Professor
Information Systems Department
Faculty of Computers and Information
Helwan University

ABSTRACT

The Online Social Networks (OSN) have a great role in increasing the communication among people. Their role never stops as they have become the way to share information and the real-time news. However, their unprecedented success has also attracted the attention of hackers, who use OSN to spread spam and malicious contents. Hackers have found a good environment, which is compatible with their goals in terms of widespread reach to the largest number of victims or even spreading large propaganda in a very short time. All this can be done using OSN. The presence of spam and malicious contents on OSN may lead to people's aversion from these sites. This research tackles this phenomenon by introducing Flexible Malicious Accounts Detector (FMAD) solution, which can detect malicious and spam accounts using predefined features. Additionally, FMDA can identify newly emerging features and classify them as either normal or abnormal. Moreover, FMDA can recognize malicious accounts campaigns. Therefore, the presented solution performs better than all previous approaches that cannot deal with new emerging features. For this purpose, FMAD uses both supervised and unsupervised machine learning techniques. The experiment shows that FMAD results in accuracy reaching 99.75 %.

General Terms

Security

Keywords

OSN; Spam; Malicious account detection; datamining; Association rules.

1. INTRODUCTION

No one can deny the importance of Online Social Networks (OSN), due to their increasing use and wide spread adoption. For these reasons, hackers try to exploit them by spearheading malicious codes and spam contents. They create automated accounts (known as bots) and group them into camping to automate the distribution of spam and malicious contents[33,11]. The number of spam and malicious accounts is increasing on OSN. This increasing reaches to a 355 % during the first half of 2013 [2]; this causes a real danger. For this reason, the security of the OSN has gained importance recently. Moreover, spam and malicious accounts do more harm to Twitter than to other social media sites such as Facebook. This is because they can easily affect real-time news and trend topics, which is the main target of twitter. For this reason, this research will focus on the study of Twitter social network.

Many researchers have tried to solve the problem of spam and malicious accounts on OSN. But most of the presented solutions lack the flexibility to identify and categorize newly emerging features to normal or abnormal ones. All conducted researches in this area assumed that the set of their discovered features are final, i.e. their features are not changing while the nature of the OSN is always rapidly changing. Spam and malicious accounts frequently change their behavior to bypass detection mechanisms; this results in the appearance of new features.

The previous approaches' assumptions lead to a gap among the real problem and their solutions. The nature of the problem is not compatible with most of the proposed solutions; because they built their systems based on specific features. Once these features are changed by the abnormal accounts or even the OSN itself, the approach becomes useless. So, this problem needs a dynamic solution that can deal with the new features as well as the known features.

The main contributions are: 1) the paper proposes a Flexible Malicious Accounts Detector (FMAD), which is the first approach to solve the problem of the new emerging features without the need to rebuild the system. The newly emerging features make any accurate approach based on supervised technique useless, because these approaches will need to be retrained and rebuilt from scratch to recognize these new features. 2) The paper provides a new taxonomy for OSN accounts' features. This taxonomy helps to reduce the processing time and enhance performance. 3) FMAD can characterize new features automatically. 4) To our knowledge FMAD is the first detector that uses association rules to detect the abnormal accounts. 5) FMAD uses a combination of supervised and unsupervised techniques. 6) In addition to its ability to detect abnormal accounts; FMAD can detect abnormal campaigns. 7) FMAD gives the best results during experimental tests.

The rest of this paper is organized as follows: Section 2 presents related work in the field of spam and malicious accounts detection. Section 3 shows background about Twitter social network and describe the methodology used to collect dataset. Then, Section 4 shows the features used to detect spam and malicious accounts. To select the most appropriate techniques and algorithms to build the system, the researchers made a comparative study among different data mining techniques; this is presented in section 5. Section 6 gives the details of the proposed solution. The experiments are presented in Section 7. Finally, Section 8 draws up the conclusion and future work.

For the rest of this research, the term malicious or abnormal will be used interchangeably to refer to spam or malicious accounts.

2. RELATED WORK

Because of the great importance of the security of social media, many researchers focused on this problem. Moreover, they investigated many directions to limit the effect of these polluters.

The study of this problem leads us to identify seven main directions that researchers used to detect abnormal accounts:

I. Identifying a specific feature and trying to detect the accounts that have this feature. In this approach, the researchers determined the most dangerous features that may cause e-crime. Gupta et al. [9], discussed the problem of using the shortening URL (Uniform Resource Locator). They showed that the massive e-crime always uses the short URL not the long URL. URL shortening is a service that enables users to map a long URL to a short URL. However, spammers use these features to obfuscate the actual URL behind a shortened link in order to be undetected. Gupta et al. [9] showed some properties of malicious short URLs and then used the classification technique for the detection. Chavoshi et al., [5]identified some important features; they produced the DeBot approach to identify the correlated user accounts. The correlated users are two accounts don't follow each other but are correlated in their activities. DeBot doesn't need labeled dataset. It consists of 4 components; collector, indexer, listener, and validator. Shehnepoor et al [23] presented the NetSpam approach that discusses the problem of spam reviews on the OSN made about products and services. They used the metapath concept in addition to a graph-based method. The general definition for the metapath method is data about data, which means data about the reviews including the account that writes the review and the demand that demands the review. The classification module of NetSpam consists of two parts: weight calculation part that determines the importance of each feature, and the labeling part that calculates the relation with the spam review.

The disadvantage of this approach is that it focuses only on a specific feature and does not consider other malicious features. Hence, they provide partial solution.

II. Collecting all available features of the abnormal account and applying different supervised data mining techniques. Some researchers tried to detect individual accounts. But most of them detected the campaigns of these accounts. It is considered the most popular approach. They depend on collecting the features of the abnormal accounts whether they are fake, spam or malicious accounts then they use the classification technique to distinguish among the accounts in their dataset. Chu et al.,[6]produced some measurements for distinguishing among the human, Cyborgs, and bots in terms of 3 groups: tweet behavior, tweet content, and account properties. They presented some features that help with their classification for example, the bots have URLs more than the human; bots tend to use third-party APIs and post at regular intervals among their tweets, this is unlike normal individual’s posts that are less on weekends and nights. Cyborgs post more tweets than the human and bots. Moreover, they used classification techniques to detect spam and malicious accounts. Another interesting approach is produced in[3]which consists of identifying a new group of features and combining them with some old features to detect the spammer and the automation on Twitter. They divided these new features into 5 categories: Bait-oriented feature, Behavioral-entropy Features, URL Features, Content-entropy Features and Profile Features. Each group contains some features to distinguish between the normal and spammers. After that, they applied the classification technique with Weka data mining tool. Then, they detected the spam campaigns by using Clustering technique (K-means algorithm).Egele et. al, [7]discovered some new features in Facebook and Twitter. They implemented their approach using COMPA tool and made the evaluation by the classification technique on the two social media platforms. Similarly, Stringhini et. al, [25]created some honey-profiles on Facebook, Twitter, and MySpace to reflect the effect of spam. After that, they extracted some features for the spammer accounts. Then they applied Random Forest algorithm from the classification technique on the Weka framework; they claimed that it is the best Classification Algorithm, which gives more accuracy with low false positive rate. Zheng et. al, [34]chose the cross-validation and Benevenuto et. al, [4] used a Support Vector Machine (SVM)which was based on spammer classification algorithm and Weka tool. Zheng applied their approach on Sina Weibo social network while Benvenuto applied on Twitter. Xiangtan criticized Benvenuto about using a lot of features that may affect the performance. Another approach created by Hua et al [13] uses content, behavioral, and graph-based data by classification algorithm. They applied a threshold (s) value. If the account has greater than S value, they assign it as spam account; else, they assign it as a normal one. They compared their approach with three other classification algorithms and found that their approach does not give the best result among them. However, this method still needs some improvement such as parallelization.

All these approaches work well and give an excellent result for a specific time period. But once the used predefined features are changed, all these approaches cannot work well. Also, they cannot work with new emerging features.

III. There are strategies and efforts that rely on visualization and do not depend on supervised techniques to detect spam. Trang et al.,[27] proposed a framework to detect Web spamming. Their strategy depends on two important steps: the first one: group similar messages into clusters. The second one: detect the clusters that violate the normal behavior of the normal user's features and marking them as compromised accounts. They made two artificially "hijacked accounts" to overcome the problem of finding the annotated dataset. Also, they made some modification to the existing COMPA tool. Another example of this approach is presented in[20]by Kaya et al. They used visualization as a step before classification to see the important features that must be used in the final classification step. They categorized their approach as an unsupervised classification. They used a self-organizing map (SOM), which is a special type of neural network model. They also used another view of SOM called heat maps.

These approaches use the Visualization technique for clustering either the accounts themselves (based on their features) or even the features to help with the classification technique. The main defect is that, abnormal accounts try to use a lot of normal features besides their abnormal features in order to look like normal ones, so many of the abnormal accounts fall into the normal cluster because of the high similarity among the features. Unfortunately, this technique has a high rate of false positives.

IV. This direction aims at enhancing the data mining algorithms used in the detection process. Miller et. al, [21]provides a modification for the two algorithms: StreamKM++ and DenStream to facilitate the problem of spam detection. They dealt with the spam problem as anomaly detection, not as a classification problem. So, the resulted model was built based on the normal behavior and the outlier is abnormal or spam accounts. This approach works well if there is a big difference between the normal and abnormal accounts concerning their features. But the abnormal accounts try to look like the normal accounts; by using a lot of normal features. So, a abnormal accounts can easily avoid detection.

V. This direction tends to design new algorithms that aim at extracting the accounts with the same purpose by their URL drove estimation. Zhang et al., [33] produced a new framework that applies the Shannon information theory. Their framework was based on three steps: the first one collects the accounts that post URL with a similar purpose. The second one extracts the campaigns that may be created for a spamming purpose. The last step defines the accounts intentions. This detection is based on two levels: message level and account level. Message-level detection technique checks every tweet or message whether it contains unwanted advertising or malicious content. But this technique needs real-time detection. Also, this type is very expensive. Account level detection examines the overall behavior of the account; it checks its tweets, URL, and properties. This approach is still suffering from the manual training because of using the classification technique, which is limited to specific predefined features. It cannot deal with newly emerging features.

VI. This direction tries to enhance the performance of the classifier by making a semi-automatic approach instead of the manual one for the training data set[26]. The researchers showed that using manual classification for training the data set is costly and takes much time. So, they tried to make a semi-automated method for labeling training dataset automatically. They used the strategy of cut-off (C) value for each feature. If the account has more than C for this feature, the account is considered as an automated one; or the account is a normal one. The disadvantage of this approach is annotating the account as an automated one if it has one automated feature. This will result in a high false positive rate. That is because many un-automated accounts may be deceived and fall to one of the automated features.

VII. The last direction produced Ensemble approach. Ensemble modeling in data mining means combing the results of applying more than two different models or algorithms into predictive analysis to enhance the accuracy of the presented model. Singh et al., [24]introduced Ensemble based Spam Detection to solve the problem of spam tweets. Their system consists of 4 classifiers. Classifier 1 checks if the tweet contains blacklist URL or not. Classifier 2 checks the duplicate tweets. Classifier 3 checks for spammy words. Classifier 4 applies 3 different models; K-Nearest Neighbor, Naïve Bayes, and Logistic Regression, Bagging and boosting techniques are used to build their dataset.

Depending on the previous survey, it is clear that these approaches have many deficiencies:

i. Using the supervised technique that needs training dataset which is expensive and costly.

ii. They lack the flexibility for dealing with the new emerging/discovered features. If a new feature appeared; this will force users to rebuild the approach from scratch to deal with this new feature.

iii. Classifying any new feature needs manual classification. So, a better solution is needed to accurately identify malicious accounts even in a dynamically changing environment.

Malicious accounts detection problem does not need a static solution that relies on specific known features. FMAD is different from all these previous works in its ability to classify new emerging features without the need for retraining or rebuilding the solution again.

3. METHODOLOGY FOR DATA COLLECTION AND ANALYSIS

Before elaborating into the details of data collection methodology, let us first introduce some important terms related to Twitter.

- "Tweet": is a short message to express what the user wants to say.
- "Retweet" is the action that the user makes if he/she wants to share another user's tweet on his/her page.
- "Favorite" action is when user admires some one's tweet; he/she marks it as Favorite.
- "Hashtag" is the most important key word for Twitter; it is a phrase which begins with "#" sign. Users use this key word if they want to collect some tweets that discuss the same topics together to make it easy for any user to know the opinions related to this topic. If there are more users speaking about

this topic, this Hashtag will become a trend.

- "Trend" is a popular Hashtag for a specific period and specific region.
- "Mention" is made by adding (@username) to Tweet or in a reply to mention this user to read this Tweet.
- "Follower" is the person who wants to know user's news.
- "Following" is the person that the user wants to know his/her news.

3.1 Data Collection and Analysis

To accurately characterize Twitter accounts features, a big, annotated, and trusted dataset in a specific format is needed. So the dataset collection of this research has been done manually. In the collection process, publicly annotated datasets such as SNAP[18]orLast.fm[35]were not used because of following reasons. First, most of the publicly annotated datasets contain annotated tweets not annotated accounts. Second, they do not show the criteria or the features that were used to make annotation decision. Third, the published dataset contains many malicious or spam accounts that have been deleted from Twitter, so it is difficult to check them now. Fourth, the published datasets did not use all features presented in section 4; so the researchers needed to build their dataset and to extract features manually.

One critical point that drew attention during the collection of the dataset was that once Twitter detects any malicious tweet or account on its site, it is removed to save other twitter users[31]. Therefore some entries in the used datasets that were manually collected from real malicious or spam accounts are not available now on Twitter.

The following six techniques were used to build their annotated dataset.

i. Around 4000 verified celebrity accounts around the world were collected. After that, the researchers checked some of their followers manually. Then accounts' features were extracted.

Spam accounts (especially which have commercial purposes such as advertising) tend to follow these celebrity accounts. By this technique, around 6,000 accounts were collected. Figure 1 shows this collection technique.

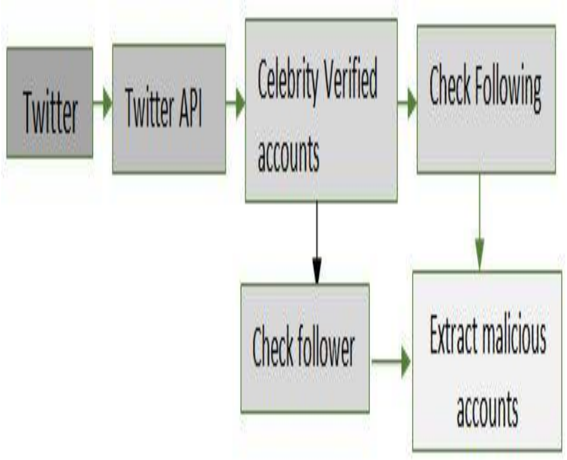

Figure 1: Tracking celebrity verified accounts.

After checking the followers and following of abnormal accounts, it was noticed that most of them were abnormal accounts from the same type; either spam or malicious accounts. That is because the abnormal accounts try to make the number of their followers close to the number of their following in order not to be noticeable for the detection. Also, they seek to increase both their follower and following to look like important normal accounts. Some of these accounts are even similar in the username. Also, if the detected abnormal accounts are developed for making a specific action like retweet for a specific account, their whole network (for the follower and following) has the same purpose.

ii. The second technique searched for the popular Hashtags (trends) around the world for a period of 7 months and checked the tweets related to these Hashtags. 5000 accounts were collected.

From the researchers' point of view, this is the best chance for the abnormal accounts to spread their content. That is because people who do not know the topic of the trend will open the Hashtag to read its tweets and access their URLs. By inspecting the tweets under these hashtags manually; the researchers found that, abnormal tweets (either malicious, spam, or fake) contain unrelated content such as (follow me, advertising material, links to unrelated or malicious sites). Figure 2 represents this technique.

iii. Collecting some accounts from Twitter blacklist, using "@Spam" indicator that the Twitter users post manually. Around 1500 accounts were collected.

iv. Investigating the scientific papers that were published in this domain or the sites that talk about this topic. Some of them publish examples of their dataset for the abnormal accounts. Moreover, by checking the follower and the following accounts. Around 500 accounts were collected.

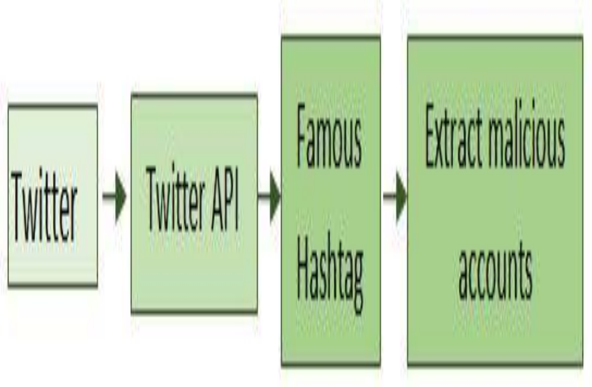

Figure 2: Tracking Popular Hashtags.

v. The fifth technique: the researchers inspected the published datasets (cresci-2017, cresci-2015) [36], and collected around 6000 accounts from these datasets.

vi. Tweets2011[37]: it is a multi- language Twitter collection which contains approximately 16 million tweets. Each tweet is identified by user ID and tweet ID. The researchers selected 2,000 English accounts and used the account ID to visit the account in Twitter. Then the researchers inspected its tweets manually and some of their followers. This yielded 5,000 accounts.

After using all these techniques, around 24,000 accounts (normal and abnormal together) with over 5M tweets and 2M URLs were collected. These accounts were labeled manually based on the features presented in section 4. However, manual annotation was a time-consuming and tedious process.

The analysis of the collected dataset gave us more insights about the problem and helped us in identifying a number of features that would be used later in the proposed solution.

4. FEATURES CLASSIFICATION

In this section the paper provides taxonomy of OSN features. As shown in figure 3, features are classified into two main categories: Predefined Features and New Unclassified Features.

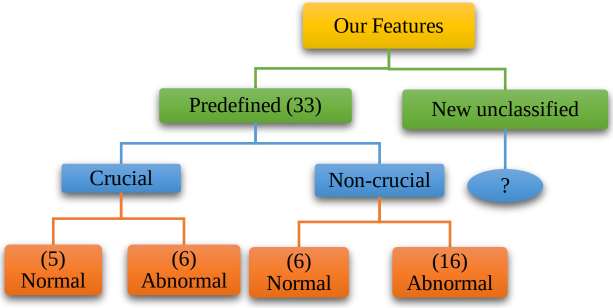

Figure 3 Features Classification

4.1 Predefined Features

These features are divided into two groups (Crucial, and Non-crucial Predefined Features). The Crucial group has the features that if one of them exists in an account; the account is marked directly as either normal or abnormal. The account cannot have normal and abnormal Crucial Predefined Feature at the same time. The Non-crucial group contains the features that cannot be used alone to judge whether the account is normal or abnormal in a crucial way. The Crucial Predefined Features facilitate the categorization process as it is an effective way to determine the account's status being malicious or not in a single classification step. This classification step is considered to be an effective genuine contribution to categorize accounts and enhance the algorithm performance.

In the following, many of the presented features were collected from the literature review (they are followed by their references) and some of them were identified by the authors; these are followed by the word (Identified). Some of the identified features may have been discussed in the literature but not used directly as a feature.

4.1.1 Crucial Normal Predefined Features

If one of the following features exists, then the account is marked as normal.

i. *Inactive accounts (Identified):* the inactive accounts cannot be malicious or an advertising one. The malicious and advertising spam accounts need to be active to spread their malware, malicious, or advertising content.

ii. *Verified accounts*[13]*:* Twitter provides this service for their users by entering their social security number (SSN) to prove that they are the right owner. So, this cannot be spam or abnormal account because hackers cannot provide valid SSN for each automated account.

iii. *The rate of tweets (Identified)*: it could be computed using equation (1).

$$Rate\ of\ Tweets = \frac{The\ number\ of\ Tweets}{The\ age\ of account{*}12} \ldots\ldots\ldots\ldots\ 1$$

If this ratio is small, this means that this account posts a small number of tweets; and that does not match with Spam and malicious behaviors. So, this cannot be abnormal account.

iv. *Registration date*[6]: bots on Twitter appeared in 2009. So, if the registration date is before 2009, this cannot be bot account.

v. *The rate of URLs(Identified):*

$$Rate\ of\ URLs = \frac{NumberofURLs}{NumberofTweets} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 2$$

If the account does not have URL or the ratio of used URL is low, it does not present abnormal account. The advertising and malicious accounts need to use URLs to publish their contents.

4.1.2 Crucial Abnormal Predefined Features

If one of the following features exists, then the account is marked as abnormal.

i. *The use of unregistered API*[6]*:* if the account accesses Twitter from unregistered API, this is abnormal account. Bots often developed their own API.

ii. *Performing same type of activity (Identified):* if the account only retweets or makes favorite, it refers to abnormal account. Normal accounts have a great diversity of activities.

iii. *The high similarity between two or more accounts content*[5]*:* the advertising campaigns or the malicious campaigns create more than one account to publish their materials. If there are some accounts that have the same content, this refers to abnormal accounts campaign.

iv. *The similarities between username for a group of accounts (Identified):* the accounts that belong to the same bot usually have similar usernames, and belong to the same network and follow each other.

v. *Using a lot of domain names mapping to the same IP address*[3]: the abnormal accounts try to avoid the repeated URLs and repeated domain names to avoid detection mechanisms. So, they make a lot of domain names for the same specific IP. This refers to abnormal account.

vi. *Several Blacklisted URLs (Identified)*:Google Sage Browsing[38], and Spamhaus [39] contain blacklists of malicious URLs[33]. So, if the account has a significant number of URLs from these blacklists, it is identified as an abnormal account.

4.1.3 Abnormal Non-Crucial Predefined Features

The abnormal accounts use a combination of these 6 categories.

A.Bot behavior

i. *The variance of tweet interval* [3]: A bot is developed to post tweets between predefined times. A normal user does not have defined interval between tweets; they display activities over an irregular interval. So if there is regularity in the tweet interval, this indicates abnormal account.

ii. *The number of tweets per unit of time*[3]: Some bots are developed to be active at specific time and sleep after that. So if the number of the tweets in each period is higher than other, this indicates abnormal account.

iii. *The source of the tweet* [6]: The bot system is built on a specific source such as web, mobile application or Twitter API. If the account uses the same source each time, this may indicate abnormal account because normal users can access twitter from different sources.

B. The observation of URL

This group of features discusses how the abnormal accounts use the URL to gain more access to their sites.

i. *Repeated URL* [3]: If the account repeats the same URL, it indicates abnormal accounts. The account does this behavior to attract the attention to this URL, to be sure of its widespread clicking.

ii. *Domain name instead of* URL [3]: if the account uses the URL for a specific site in a tweet and then publishes the domain name for the same location in another tweet, it refers to abnormal account. In this way, it tries to cheat the security mechanism and avoid detection.

C. Timeline content features

This group is concerned with detecting features by observing the content of the account such as language, the tweets' subjects and the similarity among them, the relation between the tweet and the URL.

i. *The language of the tweet*[3]*:* the normal user does not always use formal language. He/she writes tweets in formal and in informal language. The bot is designed for one type of language.

ii. *The similarity among tweets* [3]: the normal user posts in many topics in a specific area. Bot is designed to publish on a specific topic.

iii. *The relation between URL and tweet* [33]*:*This feature measures the similarity between the posted tweet and the content of the page which is related to the URL. If the URL content is not relevant to the tweet, this refers to abnormal account.

D. The profile properties

These features are related to the appearance of the accounts.

i. *Follower-to-following ratio*[6],[3]*:* this ratio is calculated by equation (3)

$$\textit{Follower-to-following ratio} = \frac{Followers \cap Following}{Number of following} \ldots\ldots\ldots\ldots..3$$

The abnormal accounts follow a lot of people who do not follow them to attract their attention. If the ratio is near to 1, it refers to normal accounts; otherwise, it refers to abnormal account.

ii. *Profile description* [3]: the normal accounts provide a description of their personalities. The abnormal accounts do not provide a description profile or provide a non-relevant description.

E. Cheating features

The abnormal accounts try to attract and cheat the normal accounts by using one of the following features.

i. The ratio between mentions and a total number of tweets [26] $= \frac{Total\ number\ of\ mentions}{Total\ number\ of\ Tweets} \ldots\ldots\ldots\ldots\ldots\ldots\ldots... \ 4$

If the ratio is high, it refers to abnormal accounts because the abnormal accounts need to attract attention of their followers in every tweet.

ii. *The ratio between mentions to non-follower and total mentions*[3]: this ratio shows how the abnormal accounts force the users (even if they are not among their followers) to see their tweets. If the ratio is high, it refers to abnormal accounts.

$$Number of mentions\ to\ non\_follower = \frac{number\ of\ mentions\ to\ non\_follower}{number of all\ mentions} \ldots\ldots\ldots\ldots\ldots\ldots\ldots5$$

iv. *Using trends*: it shows the intersection between trends (popular Hashtags) and the used Hashtags[3].

$$The\ used\ trend = \frac{\text{number of trends}}{\text{number of all hashtages}} \ldots\ldots\ldots \ 6$$

If the ratio is high, it refers to abnormal accounts; the abnormal accounts need to attract the attention to these posts.

v. Using Famous tweets (Identified): it illustrates the intersection between the popular tweets and the account's tweet. If the ratio is high, it refers to abnormal accounts.

$$\text{Famous Tweets ratio} = \frac{\text{number of popular tweets}}{\text{number of total tweets}} \ldots\ldots.. \ 7$$

F. Analysis of the activities

i. *The number of URL per tweet*[34]: the normal user posts one URL at most for each tweet. While some of the abnormal accounts post more than one URL for each tweet.

ii. *Making retweet and favorite for the same set of accounts* (Identified): the automated accounts are developed to make activities for a specific number of accounts to help them propagate and attract the attention. If the account always makes retweet or favorites only for a specific small set of accounts, this highly refers to abnormal account.

4.1.4 Normal Non-Crucial Predefined Features

Researchers used to identify the features of the abnormal accounts and detect them according to their features. The proposed solution uses association rules techniques to identify if the new appearing feature belongs to malicious or normal features group. The paper will present some normal features which are the opposite of the abnormal accounts features:

a) There are a variety of the activities (tweet, retweet, favorite).

b) The variety of the topics discussed in tweets.

c) Most of the mentions are for friends (from the follower and following sets).

d) The account Hashtags are related to tweets.

e) There is entropy on the posting time. (There is no specific time for posting or interval between tweets).

f) The ratio of following/followers number: $\frac{Number of following}{Number of followers} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. \ 8$

The most important property of abnormal account is that they try to gain a significant number of users to publish their advertising or their malicious content widely. So, if this ratio is small, this cannot be abnormal account because it indicates this is friendly network.

4.2 New Unclassified Features

As mentioned above OSN accounts have predefined features that help with classification; but social media are dynamically changing, and hackers try to change their behavior to bypass detection mechanisms. This yields a new set of features. The proposed solution deals with this problem to detect new features and classify them either as benign (normal) or malicious (abnormal).

5. COMPARATIVE STUDY AMONG DATA MINING TECHNIQUES

To identify which data mining technique is the most suitable to solve the problem, a comparative study of different data mining techniques and their algorithms has been made. The results help to identify several issues; first, the supervised classification fails to deal with new features. In addition, it is very expensive to rebuild the approach from scratch. Therefore, the classification technique provides partial solution.

On the other hand, the unsupervised clustering technique cannot help because it divides the accounts into groups based on the similarity of their features. The problem here is that, abnormal accounts try to use a lot of normal features to look like normal accounts. Consequently, most of these abnormal accounts will be in the normal clusters. Hence, clustering could be cheated.

Predicting others' behavior can be accomplished by considering the redundancy patterns. The main purpose for the association rules is finding the patterns that their itemset appear together. It is an unsupervised technique. Association rules approach could be used to build flexible technique that helps with annotating the new features and characterizing the accounts. Many algorithms have been introduced such as Apriori [28], Fp-growth [10], Éclat[32].Apriori and FP-growth provide only the features frequent pattern. They are not capable of detecting accounts frequent pattern. While Éclat algorithm helps to find both the Frequent Features that appear together and their corresponding frequent accounts that their features appear together. For these reasons, the researchers have chosen Éclat algorithm to be used in their proposed solution to identify all possible patterns.

6. FLEXIBLE MALICIOUS ACCOUNTS DETECTOR (FMAD)

This section presents the proposed solution called Flexible Malicious Accounts Detector (FMAD) to detect malicious accounts and their campaigns if any, and help classifying new emerging features.

6.1 FMAD Algorithm

FMAD algorithm goes through four main phases, as shown in figure 4.

Phase1: Crucial Features Detection

The main objective of this phase is to detect accounts which contain at least one feature from the set of crucial predefined features. To achieve this objective, 11 crucial features were defined and sorted according to their precedence. This list is editable, i.e., the administrator can delete or add any specific feature as needed. Accounts that contain any crucial abnormal feature will be classified abnormal. While Accounts that contain any crucial normal feature will be classified as normal. This is done in the first phase without further processing. This clearly reduces processing overhead and enhances system performance. Accounts which do not have any crucial feature will go to phase two for further processing.

Phase2: Features' and Accounts' frequent pattern Identification

This phase uses Éclat algorithm to identify Features and Accounts frequent pattern. In general, the frequent pattern is a set of itemset that frequently appear together. Now, FMAD want to get both the Features frequent pattern (the set of features that frequently appear together) and Accounts frequent pattern (the set of accounts that their features frequently appear together) across the whole input dataset.

As an example: Given a set of accounts from A1 to A15, each contains some features from F1 to Fn. Mining these accounts by applying Éclat algorithm, the following pattern will appear (A1, A5, A8 → F2, F3, F9, F11, F12), the first part of the pattern is called Accounts frequent pattern that their features

(the second part) appear together. F2, F9, F11 are normal features, F3 is abnormal features and F12 is unspecified new feature. But so far there is not any characterization for the accounts: A1, A5, and A8 as normal or abnormal. FMAD have just certain patterns that need further analyses to reach a final decision. This is will be done in the next phases.

Additionally, in this phase, FMAD performs identification and extraction of bot-campaigns. The features of a single campaign are the same, so they will appear in the same patterns. Then FMAD can identify features frequent patterns that indicate this campaign and extract corresponding accounts easily. For example, if a campaign is intended to support a political candidate, the common features among the campaign accounts are F1, F3, and F5. FMAD will group the pattern with the second part F1, F3, and F5 in a single campaign.

Phase 3: New unclassified features annotation

In this phase, new unclassified features are annotated, in case any has been found. Considering frequent features patterns that resulted from the Éclat processing, for a given unclassified feature, one of the following scenarios can happen:

It appears with at least one abnormal feature from within a pattern resulted from phase 2; in this case, it will be marked as abnormal non-crucial. This implies that the new feature is frequently used with this abnormal feature. Orit appears with all accompanied normal features. Here, it will be marked as normal non-crucial.

Phase 4: The Decider

The decider determines if the account is normal or abnormal by applying the following procedure:

1. Create the *Benign, Suspect, Unspecific, Normal, and Abnormal* empty groups.

2. Inspect Éclat results produced from phase two. These results contain the frequent features and their corresponding frequent accounts pattern. Divide all accounts patterns into *Benign* or *Suspect* as follow:

   2.1. Check each feature in each pattern; if there is one abnormal feature in the features frequent pattern; place all its corresponding frequent accounts in the *Suspect* group. Here, the suspect group contains all accounts that have at least one abnormal feature; they have the probability to be abnormal accounts.

   2.2. For features patterns which do not contain any abnormal feature, place all their corresponding frequent accounts in the *Benign* group.

3. Get the intersection between the *Benign* and *Suspect* groups and move the results to the *Unspecific* group for further checking. This step produces accounts that have a mix of normal and abnormal features.

4. For each account in the unspecific group,
    4.1. Move the account to *Abnormal* group if it has two or more non-crucial abnormal features.
    4.2. Else, move it to the normal group.
5. Finally:
    5.1. Move the accounts that remain in the suspect group to the abnormal group as their accounts appear only with abnormal features frequent patterns.
    5.2. Move the accounts that remain in the *Benign* group to the normal group, because their accounts appear only with normal features frequent patterns.

6.2 FMAD Implementation:

The researchers implemented FMAD solution using WEKA tool [1] which refers to "Waikato Environment for Knowledge Analysis". They choose WEKA due to its popularity as a data miming tool. Moreover, it contains all data mining algorithms that the experiment needs as shown in the next section.

7. EXPERIMENT AND EVALUATION

To evaluate the effectiveness of FMAD solution, the paper did several experiments. In this section the paper presents the test dataset, test cases, algorithms under test, evaluation metrics and results analysis.

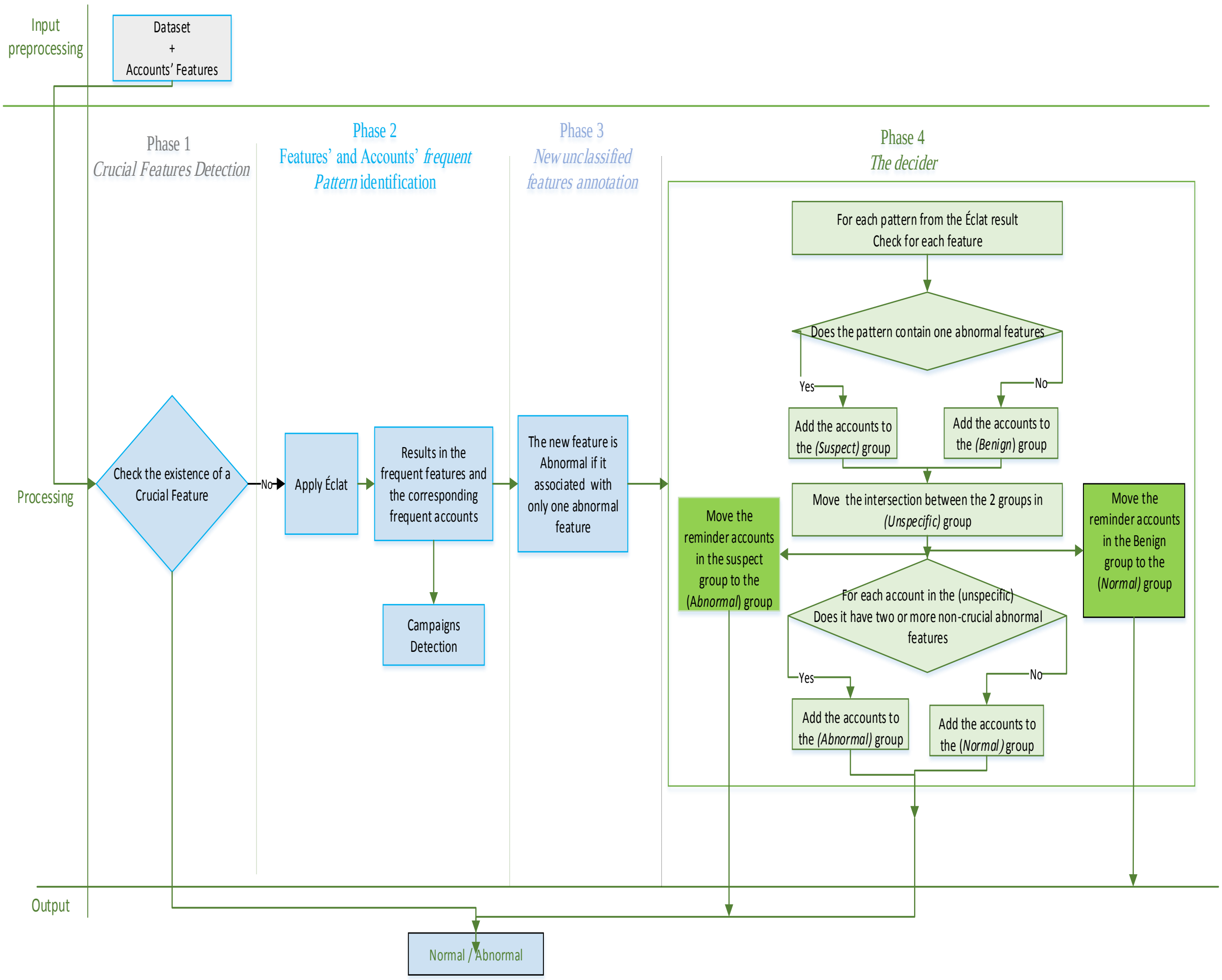
Input preprocessing
Dataset + Accounts' Features
Phase 1
Crucial Features Detection
Phase 2
Features' and Accounts' frequent Pattern identification
Phase 3
New unclassified features annotation
Phase 4
The decider
Processing
Check the existence of a Crucial Feature
No
Apply Éclat
Results in the frequent features and the corresponding frequent accounts
Campaigns Detection
The new feature is Abnormal if it associated with only one abnormal feature
For each pattern from the Éclat result Check for each feature
Does the pattern contain one abnormal features
Yes
No
Add the accounts to the (Suspect) group
Add the accounts to the (Benign) group
Move the intersection between the 2 groups in (Unspecific) group
Move the reminder accounts in the suspect group to the (Abnormal) group
Move the reminder accounts in the Benign group to the (Normal) group
For each account in the (unspecific) Does it have two or more non-crucial abnormal features
Yes
No
Add the accounts to the (Abnormal) group
Add the accounts to the (Normal) group
Output
Normal / Abnormal

Figure 4: FMAD algorithm

7.1 Test Dataset

There are two choices to prepare malicious test dataset. First, the developed system must be faster than Twitter in detecting these real malicious contents to get them and be able to use them in the test. Second, use simulation dataset for malicious accounts. The second choice is better because it gives a chance to make a bigger dataset. The research used a combination of these two techniques so the used test dataset contains two types of test data: real accounts and simulated accounts.

The test dataset contains 24K real accounts (8K are abnormal and 16K are normal accounts) and 81K simulated accounts (44.5K are abnormal and 36.5K are normal accounts). So, the total size of the test dataset is 105K accounts.

The accounts were annotated using features presented in section 4. However, to represent the case of new emerging undefined feature, and check FMAD ability for characterizing new features, the researchers have two choices: 1) either fabricate new features and annotate them as normal or abnormal, or 2) use known existing features as newly emerging features. The first approach is unrealistic and may produce biased results which will harm the suitability and feasibility of the solution while the second approach is more realistic and will produce unbiased results. This idea is to increase the robustness of the experiment.

For this reason, small, selected set of known features are used as new unclassified features. The sample contains 16 Abnormal Non-Crucial Predefined Features and 6 Normal

Non-Crucial Predefined Features. 5 Abnormal features and 1 Normal Feature were randomly selected to be used as new features. The final features set contain 27 Non-Crucial and 11 Crucial as predefined and 6 new unclassified features.

7.2 Algorithms Under Test

To assess the effectiveness of FMAD, it was compared with 7 different mining algorithms: 4 classification and 3 clustering algorithms. All 7 algorithms are implemented in WEKA. For Classification, Naïve Bayes [8][15] J48 [17][12] , Random Forest [19]and SVM [30] were tested. The supplied test set is used as the test option (in WEKA) for all classification algorithms. While for clustering the research used K-Means [16]with k = 2 which signifies optimal number of the wanted clusters (normal and abnormal one), DBSCAN[16] [22] with minStdDev=1.0E-6, and EM [14][29] with the number of clusters = 2 and minStdDev = 1.0E-6. For FMAD, the research used the minsup = 10%. The relation between the minsup and the dataset size is shown in fig 5. The users should reduce the minsup when the dataset size increases.

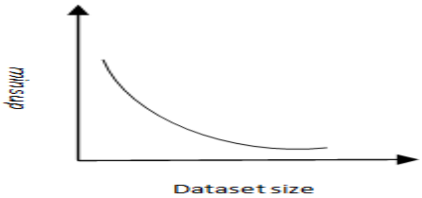

Figure 5 minsup and dataset size

7.3 Evaluation Metrics

To evaluate the proposed solution compared to other Algorithms, the following standard detection metrics are used.

True Positive (TP) describes the case of correctly identifying the abnormal account as known. False Negative (FN) describes the case of the wrong assignment of a real abnormal account as a normal account. False Positive (FP) describes the case of the wrong assignment of a real normal account as an abnormal one. True Negative (TN) describes the case of correctly judging the actual normal account as a normal one. Based on the previous four definitions the following standard metrics are calculated.

Precision (P): represents the positive prediction value; P is calculated as the ratio between the correctly detected abnormal accounts (TP) and the total predicted as abnormal accounts either they are really predicted or not. It is expressed by equation 9. Recall (R): is the sensitivity that shows the percent of the correct accounts that the proposed approach produces. It is calculated by equation 10. F-Measure (FM) is the harmonic means between the precision and the recall. It is calculated by equation 11. Accuracy is the ratio between the true result and the total data. It is calculated by equation 12.

$$Precision(\mathrm{P}) = TP/(TP+FP)\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9$$

$$\mathrm{Recall\ (R)} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})\ldots\ldots\ldots\ldots\ldots\ldots\ldots 10$$

$$\mathrm{FM} = 2(\mathrm{P}*\mathrm{R})/(\mathrm{P}+\mathrm{R})\ldots\ldots\ldots\ldots\ldots\ldots\ldots 11$$

$$\mathrm{Accuracy} = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})\ldots\ldots\ldots 12$$

7.4 Test Cases

The test dataset was divided into 4 groups to present four different test cases: group (1) consists of the accounts that all their features are known. This is a normal case. This group contains around 19K accounts. Group (2) consists of the accounts that half of their features is known and the other half is new (unclassified). To test how algorithms under test will deal with new (unclassified) features. It contains 62K accounts. Group (3) consists of the accounts that all their features are new (unclassified) except one or two features at most. This is an extreme case. Group 3 contains around 22K accounts. Finally, group (4) is a special case in which the abnormal account uses only one Non-Crucial abnormal feature and all other features are normal to avoid the detection mechanism. This group consists of 2K accounts. In all four groups, half of the accounts is normal while the other half is abnormal.

7.5 Results Analysis

In this section the paper presents the obtained results for each test case. Table 1 shows algorithms detection metrics for each group while, Figure 6 illustrate their Accuracy.

For group (1) which is the casual case, almost all classification algorithms give good results. These good results achieved because this group contains known features and classification algorithms were trained to detect these features. On the other hand, there are high differences between clustering algorithms. DBSCAN gives the lowest accuracy while EM has good results and K-Means came in between. In general, clustering algorithms try to group similar accounts into same set. They are unsupervised techniques. The test reflects that EM is the most effective clustering technique that could be used for abnormal account detection while other algorithms (DBSCAN and K-Means) are not suitable enough. FMAD presents good results for this case. It uses a combination of supervised and unsupervised techniques.

Table 1: Detection Metrics for Each Test Case.

	Classification Algorithms								Clustering Algorithms							
	Naïve Bayes		J48		Random forest		SVM		DBSCAN		K-Means		EM		FMAD	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
Group (1)	100	0	100	11.5	97.1	3.8	100	0	14.8	3.8	79.5	84.6	97.1	0	98.2	0
Group (2)	97.1	12.1	100	35.9	97.1	8.3	97.1	7.7	97.1	92.4	97.1	92.4	100	0	100	0
Group (3)	100	100	100	100	100	100	100	100	100	0	95	0	100	0	100	0
Group (4)	0	0	50	0	0	0	0	0	71.4	0	71.4	0	71.4	0	95	0

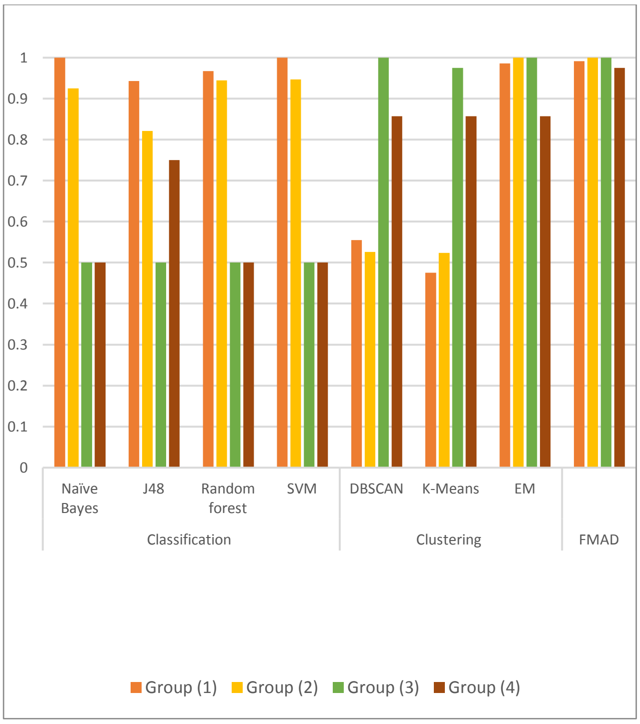

Figure 6 Accuracy for each group

For group (2), Classification algorithms accuracy was decreased because Classification algorithms are not trained to deal with new features. Regarding clustering algorithms, DBSCAN gives lowest accuracy while K-Means accuracy slightly increased. EM gives a better accuracy. FMAD gives the best results as it can deal with old and new features.

For group (3), Classification algorithms suffer dramatic decrease in their accuracy. The Classification algorithms result in high TP and high FN rate, although, the number of the used normal and abnormal accounts is equal. This means that, they classify all new features as abnormal ones either they are normal or abnormal accounts. This is very dangerous because it means that if the OSN present a new feature for their users; all these algorithms will deal with these accounts as abnormal ones. This is really a terrible problem. In contrast, all clustering algorithms have high accuracy because group (3) has high degree of differences between accounts' features and this allows clustering algorithms to work well. Also for this case, FMAD gives the best results.

For group (4) Classification algorithms still have low accuracy while clustering algorithms have higher accuracy compared to classification. However clustering algorithms accuracy in this case is less than the case of group (3). EM algorithm ability to detect abnormal accounts dramatically decreased. EM cannot deal well with group (4). In contrast, FMAD gives the best accuracy for this case. This proves its ability to avoid such evasion techniques.

Table 2: Calculated Metrics for each Algorithm.

Metrics	**Classification Algorithms**				**Clustering Algorithms**			FMAD
	Naïve Bayes	J48	Random forest	SVM	DBSCAN	K-Means	EM	
Weighted average of TP(%)	96.4	98.1	95.8	96.4	79.9	92.5	98.8	99.5
Weighted average of TN(%)	74.8	58.3	76.2	77.4	44.6	27.7	100	100
Precision(%)	79.3	70.1	80.1	81	59.1	56.1	100	100
Recall(%)	96.4	98.1	95.8	96.4	79.9	92.5	98.8	99.5
F-Measurement(%)	87.1	81.8	87.3	88.1	67.9	69.8	99.4	99.7
Overall Accuracy(%)	85.6	78.2	86	86.9	62.3	60.1	99.4	99.75

Table 2 presents weighted average for each algorithm calculated over four groups. Precision, Recall, F-measurement, and overall accuracy are calculated. It is clear that FMAD achieves the best accuracy (99.75%) among all tested algorithms. EM comes next with accuracy up to 99.4%; (it is the best algorithm for Clustering technique). Then, SVM (from Classification technique) comes with accuracy 86.9%. For the recall and precision values, FMAD comes first with values up to 100% as precision and 99.5% as recall. EM comes next with values up to 100% as precision and 98.8% as recall. Then, J48 comes with recall up to 98.1% and the Random Forest comes with precision up to 80.1%.

Based on the obtained results, the researchers are able to conclude that, the supervised technique (in general) that uses training dataset is not able to deal with rapid changes in used features. For Clustering technique, most of its algorithms do not give acceptable results except EM algorithm; but it needs a little enhancement. FMAD provides the best solution; it can deal with known and new emerging features with highest accuracy.

8. CONCLUSION

This research presented FMAD solution for the problem of spam and malicious accounts on OSNs. A set of a large number of malicious and normal accounts was manually collected from Twitter. The analysis of these accounts helped to identify the features that were used later for accounts detection and identification. The paper provided a new taxonomy for OSNs features that was used in the proposed technique to facilitate the detection process and enhance performance. In addition, the solution uses a combination of supervised and unsupervised data mining techniques. The paper may be the first who use association rules for spam and malicious account detection. The solution was implemented using WEKA. The implementation was tested against numbers of current data mining algorithms. The obtained results showed that FMAD can detect malicious and spam accounts with overall accuracy up to 99.75%. In addition, FMAD could identify and classify any new emerging features whether they are malicious or normal. Moreover, FMAD can detect bots-campaigns.

Although FMAD was implemented in the context of Twitter, it could be easily extended to any other social media site with few modifications. However, the features extraction should be done automatically; it is a point for future research.

9. REFERENCES

[1] S.S. Aksenova, WEKA Explorer Tutorial, Calif. State Univ. Sacramento. 11 (2006) 1–37. doi:10.1007/s10115-007-0114-2.

[2] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V.K. Singh, M. Alsaleh, A. Alarifi, A. Alfaris, A. Sandy, If it looks like a spammer and behaves like a spammer , it must be a spammer : analysis and detection of microblogging spam accounts, Int. J. Inf. Secur. (2016). doi:10.1007/s10207-016-0321-5.

[3] A.A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, C. Yang, CATS: Characterizing automation of Twitter spammers, 2013 5th Int. Conf. Commun. Syst. Networks, COMSNETS 2013. (2013). doi:10.1109/COMSNETS.2013.6465541.

[4] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on twitter, Collab. Electron. Messag. Anti-Abuse Spam Conf. 6 (2010) 12. doi:10.1.1.297.5340.

[5] N. Chavoshi, H. Hamooni, A. Mueen, DeBot: Twitter bot detection via warped correlation, Proc. - IEEE Int. Conf. Data Mining, ICDM. (2017) 817–822. doi:10.1109/ICDM.2016.86.

[6] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Who is Tweeting on Twitter: Human, Bot, or Cyborg?, ACSAC. (2010) 21. doi:10.1145/1920261.1920265.

[7] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, COMPA: Detecting Compromised Accounts on Social Networks, Symp. Netw. Distrib. Syst. Secur. . (2013). http://www.people.vcu.edu/~cfung/bib/compromised_accounts_detection-ndss13.pdf (accessed April 4, 2018).

[8] A. Goyal, R. Mehta, Performance comparison of Naïve Bayes and J48 classification algorithms, Int. J. Appl. Eng. Res. 7 (2012) 1389–1393.

[9] N. Gupta, A. Aggarwal, P. Kumaraguru, Bit.ly/malicious: Deep dive into short URL based e-crime detection, eCrime Res. Summit, eCrime. 2014–Janua (2014) 14–24. doi:10.1109/ECRIME.2014.6963161.

[10] J. Han, J. Pei, Y. Yin, Mining Frequent P atterns without Candidate Generation, in: Conf. Manag. Data (SIGMOD'00, Dallas, TX), New York, NY, USA, 2000. https://www.cs.sfu.ca/~jpei/publications/sigmod00.pdf (accessed April 6, 2018).

[11] P.N. Howard, B. Kollanyi, Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum, (n.d.). http://ssrn.com/abstract=2798311 (accessed April 4, 2018).

[12] B. HSSINA, A. MERBOUHA, H. EZZIKOURI, M. ERRITALI, A comparative study of decision tree ID3 and C4.5, Int. J. Adv. Comput. Sci. Appl. 4 (2014) 13–19. doi:10.14569/SpecialIssue.2014.040203.

[13] W. Hua, Y. Zhang, Threshold and associative based classification for social spam profile detection on twitter, Proc. - 2013 9th Int. Conf. Semant. Knowl. Grids, SKG 2013. (2013) 113–120. doi:10.1109/SKG.2013.15.

[14] S. Huang, P. Adviser-Rastgoufard, A comparative study of clustering and classification algorithms, (2007) 170–178.

[15] S.D. Jadhav, H.P. Channe, Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques, Int. J. Sci. Res. 14611 (2013) 2319–7064. www.ijsr.net.

[16] P. Kakkar, A. Parashar, Comparison of Different Clustering Algorithms using WEKA Tool, Int. J. Adv. Res. Technol. Eng. Sci. 1 (2014) 20–22.

[17] H. Kaur, H. Kaur, Classification of data using New Enhanced Decision Tree Algorithm (NEDTA), Int. J. Emerg. Technol. Comput. Appl. Sci. (IJETCAS). (2014) 147–152.

[18] J. Leskovec, Stanford Large Network Dataset Collection, (n.d.). https://snap.stanford.edu/data/ (accessed April 5, 2018).

[19] Y. Liu, Random forest algorithm in big data environment, Comput. Model. NEW Technol. 18 (2014) 147–151.

[20] Kaya Mehmet, S. Conley, A. Varol, Visualization of the Social Bot's Fingerprints, 4th Int. Symp. Digit. Forensics Secur. IEEE. (2016) 161–166.

[21] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, A.H. Wang, Twitter spammer detection using data stream clustering, Inf. Sci. (Ny). 260 (2014) 64–73. doi:10.1016/j.ins.2013.11.016.

[22] K. Mumtaz, M. Studies, T. Nadu, An Analysis on Density Based Clustering of Multi Dimensional Spatial Data, Indian J. Comput. Sci. Eng. 1 (2010) 8–12.

[23] S. Shehnepoor, M. Salehi, R. Farahbakhsh, N. Crespi, NetSpam: A Network-Based Spam Detection Framework for Reviews in Online Social Media, IEEE Trans. Inf. Forensics Secur. 12 (2017) 1585–1595. doi:10.1109/TIFS.2017.2675361.

[24] A. Singh, S. Batra, Ensemble based spam detection in social IoT using probabilistic data structures, Futur. Gener. Comput. Syst. 81 (2018) 359–371. doi:10.1016/j.future.2017.09.072.

[25] G. Stringhini, C. Kruegel, G. Vigna, Detecting Spammers on Social Networks, ACSAC. (2010) 1–9. http://www.cse.fau.edu/~xqzhu/courses/Resources/GSC.acsac10-socialnets.pdf.

[26] C. Teljstedt, M. Rosell, F. Johansson, A Semi-automatic Approach for Labeling Large Amounts of Automated and Non-automated Social Media User Accounts, Proc. - 2nd Eur. Netw. Intell. Conf. ENIC 2015. (2015) 155–159. doi:10.1109/ENIC.2015.31.

[27] D. Trang, F. Johansson, M. Rosell, Evaluating Algorithms for Detection of Compromised Social Media User Accounts, Proc. - 2nd Eur. Netw. Intell. Conf. ENIC 2015. (2015) 75–82. doi:10.1109/ENIC.2015.19.

[28] I. Tudor, Association Rule Mining as a Data Mining Technique, Univ. Pet. Din Ploiesti. LX (2008) 49–56.

[29] B. Umale, Overview of K-means and Expectation Maximization Algorithm for Document Clustering, (2014) 5–8.

[30] V.N. Vapnik, Statistical Learning Theory, Adapt. Learn.

Syst. Signal Process. Commun. Control. 2 (1998) 1–740. doi:10.2307/1271368.

[31] R. Venkatesh, J.K. Rout, S.K. Jena, Malicious account detection based on short URLs in twitter, Master Technol. Natl. Inst. Technol. Rourkela. 395 (2015) 243–251. doi:10.1007/978-81-322-3592-7_24.

[32] M.J. Zaki, S. Parthasarathy, M. Ogihara, W. Li, New Algorithms for Fast Discovery of Association Rules, 3rd Intl Conf Knowl. Discov. Data Min. 20 (1997) 283–286. doi:10.1.1.42.5143.

[33] X. Zhang, S. Zhu, W. Liang, Detecting spam and promoting campaigns in the Twitter social network, Proc. - IEEE Int. Conf. Data Mining, ICDM. (2012) 1194–1199. doi:10.1109/ICDM.2012.28.

[34] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, C. Rong, Detecting spammers on social networks, Neurocomputing. 159 (2015) 27–34. doi:10.1016/j.neucom.2015.02.047.

[35] Last.fm dataset 360K - MTG - Music Technology Group (UPF), (2010). https://www.upf.edu/web/mtg/lastfm360k (accessed April 5, 2018).

[36] Datasets, (2017). https://botometer.iuni.iu.edu/bot-repository/datasets.html.

[37] Tweets2011, Tweets2011, (2011).

[38] Google Transparency Report, Google Transparency Report, (2010). https://transparencyreport.google.com/.

[39] Spamhaus, https://www.spamhaus.org/lookup/.