Detecting Fake Accounts on Social Media

Sarah Khaled
Faculty of Computers and Information
Cairo University
Cairo-Egypt
Email: Sarah.Khaled.Mostafa@gmail.com

Neamat El-Tazi
Faculty of Computers and Information
Cairo University
Cairo-Egypt
Email: n.eltazi@fci-cu.edu.eg

Hoda M. O. Mokhtar
Faculty of Computers and Information
Cairo University
Cairo-Egypt
Email: h.mokhtar@fci-cu.edu.eg

***Abstract*—In the present generation, on-Line social networks (OSNs) have become increasingly popular, people's social lives have become more associated with these sites. They use on-Line social networks (OSNs) to keep in touch with each others, share news, organize events, and even run their own e-business. The rapid growth of OSNs and the massive amount of personal data of its subscribers have attracted attackers, and imposters to steal personal data, share false news, and spread malicious activities. On the other hand, researchers have started to investigate efficient techniques to detect abnormal activities and fake accounts relying on accounts features, and classification algorithms. However, some of the account's exploited features have negative contribution in the final results or have no impact, also using standalone classification algorithms does not always achieve satisfactory results. In this paper, a new algorithm, SVM-NN, is proposed to provide efficient detection for fake Twitter accounts and bots, feature selection and dimension reduction techniques were applied. Machine learning classification algorithms were used to decide the target accounts identity real or fake, those algorithms were support vector machine (SVM), neural Network (NN), and our newly developed algorithm, SVM-NN. The proposed algorithm (SVM-NN) uses less number of features, while still being able to correctly classify about 98% of the accounts of our training dataset.**

1. Introduction

Online social networks(OSNs), such as Facebook, Twitter, RenRen, LinkedIn, Google+, and Tuenti, have become increasingly popular over the last few years. People use OSNs to keep in touch with each others, share news, organize events, and even run their own e-business. For the period between 2014 and 2018 around 2.53 million U.S. dollars have been spent on sponsoring political ads on Facebook by non-profit organizations [1].

Facebook community continues to grow with more than 2.2 billion monthly active users and 1.4 billion daily active users, with an increase of 11% on a year-over-year basis [2]. In the second quarter of 2018 alone, Facebook reported that its total revenue was $13.2 billion with $13.0 billion from ads only [2].

Similarly, in second quarter of 2018 Twitter has reported reaching about one billion of Twitter subscribers, with 335 million monthly active users [3]. In 2017 Twitter reported a steady revenue growth of 2.44 billion U.S. dollars, with 108 million U.S. dollars lower profit compared to the previous year [3].

Online Social Networks (OSNs) have also attracted the interest of researchers for mining and analyzing their massive amount of data, exploring and studying users behaviours as well as detecting their abnormal activities [4].

In [5] researchers have made a study to predict, analyze and explain customers loyalty towards a social media-based online brand community, by identifying the most effective cognitive features that predict their customers attitude.

The implications of researchers attempt may helps an OSN operator detecting fake accounts efficiently and effectively, hence, improve the experience of their users by preventing annoying spam messages and other abusive content. The OSN operator can also increase the credibility of its user metrics and enable third parties to consider its user accounts [6]. Information security and privacy are among the primary requirements of social network users, maintaining and providing those requirements increases network credibility and subsequently its revenues. As recently, banks and financial institutions in U.S. have started to analyze Twitter and Facebook accounts of loan applicants before actually granting the loan [7].

The open nature of OSNs and the massive amount of personal data for its subscribers have made them vulnerable to Sybil attacks [8]. In 2012, Facebook noticed an abuse on their platform including publishing false news, hate speech, sensational and polarizing, and others [9]. This phenomena raised the flag for the need of new techniques to detect such actions and avoid them.

In 2015 Facebook estimated that nearly 14 million of its monthly active users are in fact undesirable, representing malicious fake accounts that have been created in violation of the websites terms of service [10]. Facebook, for the first time, shared a report in the first quarter of 2018 that shows their internal guidelines used to enforce community standards covering their efforts between October 2017 to March 2018, this report illustrates the amount of undesirable content that has been removed by Facebook [11], and it

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

covers six categories:

- graphic violence
- adult nudity
- sexual activity
- terrorist propaganda
- hate speech
- spam

In addition, 837 million posts containing spam have been taken down, and about 583 million fake accounts have been disabled, Facebook also has removed around 81 million undesirable content in terms of the rest violating content types. However, even after preventing millions of fake accounts from Facebook, it was estimated that around 88 million accounts are still fake [11]. Statistics show that 40% of parents in the United States and 18% of teens have a great concern about the use of fake accounts and bots on social media to sell or influence products [12].

Another example, during the 2012 US election campaign, the Twitter account of challenger "Romney" experienced a sudden jump in the number of followers. The great majority of them were later claimed to be fake followers [13].

In December 2015, Adrian Chen, a reporter for the New Yorker, noted that he had seen a lot of the Russian accounts that he was tracking switch to pro-Trump efforts, but many of those were accounts that were better described as trolls accounts managed by real people that were meant to mimic American social media users [14].

Similarly, before the general Italian elections of February 2013, online blogs and newspapers reported statistical data over a supposed percentage of fake followers of major candidates [15]. The Jakarta gubernatorial election in 2017, has faced a fake news campaign launched by the opponents of Mr.Purnama the Christian candidate, and the Islamic opposition candidate Mr Baswedan. More than 1,000 reports about politics and the election were confirmed as hoaxes. Mr.Purnama opponents claimed that his picture shaking hands with King Salman is fake as they said "This news is hoax, because it is haram for a king to shake hands with the blasphemer of Islam". On the other hand There were also fake posters spread online saying: "If Mr Baswedan loses the election, there will be Muslim Revolution", those fake news aims to exacerbated the divisions in society, and smeared the candidates [16].

In 2017, fake posts have shared a roamer on social media that the actor Clint Eastwood has been dead, however, the claims were proven to be false

In general, attackers follow the concept of having OSNs user accounts are "keys to walled gardens" [17], so they deceive themselves off as somebody else, by using photos and profiles that are either snatched from a real person without his/her knowledge, or are generated artificially, to spread fake news, and steal personal information. These fake accounts are generally called imposters [10], [18].

To enhance their effectiveness, these malicious accounts are often armed with stealthy automated tweeting programs, to mimic real users, known as bots [19].

OSNs are employing different detecting algorithms and mitigation approaches to address the growing threat of fake/malicious accounts. Though Sybil accounts find a way to cloak their behavior with patterns resembling real accounts [19], [20], [21], they manifest numerous profile features and activity patterns. Thus, automated Sybil detection are not always robust against adversarial attacks, and does not yield desirable accuracy. In all cases, such fake accounts have a harmful effect on users, and their motives would be anything other than good intentions as they usually flood spam messages, or steal private data [19], [22], [23].

Inspired by the importance of this problem, researchers focus on identifying fake accounts through analyzing user level activity by extracting features from recent users e.g number of posts, number of followers, profiles. They apply trained machine learning technique for real/fake accounts classification [10], [24]. Another approach is using graph level structure where the OSN is modeled as a graph essentially presented as a collection of nodes and edges. Each node represents an entity (e.g. account), and each edge represents as a relationship (e.g. friendship) [6], [25].

Though Sybil accounts find a way to cloak their behaviour with patterns resembling real accounts [19], [20], [21], they manifest numerous profile features and activity patterns. Thus, automated Sybil detection are not always robust against adversarial attacks, and does not yield desirable accuracy.

In this paper, a hybrid classification algorithm has been used by running the Neural Network (NN) [26] classification algorithm on the decision values resulting from the Support vector machine (SVM) [6], [10], [27], this algorithm uses less number of features, while still being able to correctly classify about 98% of the accounts of our training dataset. In addition, we also validated the detection performance of our classifiers over two other sets of real and fake accounts, disjoint from the original training dataset as illustrated in Section 4. The rest of this paper is organized as follows. Section 2, provides an overview about the research carried out on Twitter network and prior research on fake profile detection. In section 3, the Twitter dataset has been described. Section 4, demonstrates how the collected data has been pre-processed and used to classify the accounts into fake accounts and real accounts. In section 5, the overall accuracy rates have been discussed and compared with all other used methods. In Section 6, we present our conclusions.

2. Related work

Inspired by the importance of detecting fake accounts, researchers have recently started to investigate efficient fake accounts detection mechanisms. Most detection mechanisms attempt to predict and classify user accounts as real or fake (malicious, Sybil) by analyzing user level activities or graph-level structures. There are several data mining methodologies [4] and approaches that help detecting fake accounts that are described in the following subsections.

2.1. Feature Based detection

This approach relies on user-level activities and associated account details (user logs and profiles). Unique features are extracted from recent user activities (e.g. frequency of friend requests, fraction of accepted requests), then those features are applied to a classifier that has been trained offline using machine learning techniques [6], [28], [29], [30]. In [28], the authors used a click-stream dataset provided by RenRen, a social network used in China [31], to cluster user accounts into similar behavioral groups, corresponding to real or fake accounts. Using the METIS clustering algorithm with both session and clicks features, such as:

- Average clicks per session
- Average session length
- The percentage of clicks used to send friend requests
- Visit photos
- Share contents

The authors were able to classify the data with 3% false positive rate and 1% false negative rate.
In [30], the authors used ground-truth provided by RenRen to train an SVM classifier in order to detect fake accounts. Using simple features, such as:

- frequency of friend requests
- fraction of accepted requests

The authors were able to train a classifier with 99% true-positive rate (TPR) and 0.7% false-positive rate (FPR).
In [27], researchers used a ground truth provided by Twitter; the data have been processed using two main approaches:

- Single classification rules
- Feature sets proposed in the literature for detecting spammers

Some features have been used from previous work such as Stateofsearch.com rule set [32], and Socialbakers rule set [33]. The authors were able to correctly classify more than 95% of the accounts of the original training set.

2.2. Feature Reduction

High dimensional data could be a serious problem for many classification algorithms because of its high computational cost and memory usage. On the other hand, reducing the dimension space would remove noisy (i.e. irrelevant) and redundant features and lead to a better classification model and simple visualization technique [34]. Feature reduction techniques can be categorized into two types:

- Dimensionality reduction where the data in high-dimensional space is transformed into a space of fewer dimensions
- Feature subset selection where the original features set is disjoint into a selected subsets of features to build simpler and faster models, increases the models performance, and gain a better understanding of the data. Feature subset selection could be broken down into (filtering methods, and wrapping methods)

Among the most commonly used feature reduction techniques is the Principle Component Analysis (PCA) [35]. PCA is a technique used to identify features (dimensions) that best explain the predominant normal user behavior [30], [36]. PCA projects high-dimensional data into a low-dimensional subspace (called the normal subspace) of the top-N principal components that accounts for as much variability in the data as possible.
In [30], authors collected their data from three social networks to illustrate that normal user behavior is low dimensional. An extensive ground-truth data of Sybil behavior exhibited by fake, compromised, and colluding users were used to evaluated their Sybil detection technique. This approach achieves a detection rate of over 66 %(covering more than 94% of misbehavior) with less than 0.3% false positives.

2.3. Neural network (NN), and Support vector machine (SVM)

In [26], authors extracted the profile features using PCA, and then applied Neural Networks, and Support Vector machines to detect legitimate profiles. "Variance maximization" was selected as a mathematical way of deriving PCA results, which were:

- Number of Languages.
- Profile_Summary
- Number of Connections
- Number of Skills
- Number of LinkedIn Groups
- Number of Publications

Authors applied Neural Networks with resilient back propagation (Rprop) and SVM with C-support vector classification and polynomial kernel (polydot) as kernel functions. The findings of this paper showed that using PCA as a dimension reduction produce better accuracy results than using all the features without any selection.
Even though feature-based detection scales to large OSNs, it could be circumvent. Attackers used to change content and activity patterns of their actions to avoid spam detection techniques. [6]. Feature-based detection does not provide any formal security guarantees and often results in a high false positive rate in practice [24]. In the following discussion we will elaborate how different machine learning techniques along with feature reduction methods are employed to efficiently solve the problem of malicious accounts over OSNs.

3. Baseline Datasets

In this research we used "MIB" dataset [27] to perform our study, it consists of three datasets collected from Twitter as follows:

3.1. Baseline dataset of Real followers

Authors of [27] have mentioned that they had two datasets of real Twitter followers.

3.1.1. The Fake Project. The Fake Project dataset, henceforth *TFP*, was collected in a research project owned by researchers at IIT-CNR,in Pisa-Italy.It consists of 469 volunteers accounts that have been certified as human by CAPTCHA.

3.1.2. #elezioni2013 dataset. The #elezioni2013 dataset, was used to support a research initiative for a sociological study in collaboration with the University of Perugia and the Sapienza University of Rome. This study added a set of 1481 Twitter accounts with different professional background except for the following categories: political, parties, journalists, bloggers, those selectec accounts were manually labeled as a human by two sociologists from the University of Perugia, Italy. In the rest of this paper the dataset of this section is referred to as *E13*.

3.2. Baseline dataset of fake followers

In April 2013 researchers have bought around 3351 fake Twitter accounts from the following online markets: http:// fastfollowerz.com, http://interTwitter.com, and accounts from http://Twittertechnology.com, at a prices of $19, $14 and $13 respectively. They have built three datasets:

- fastfollowerz dataset contains 1169 fake accounts, labeled FSF.
- interTwitter dataset contains 1337 fake accounts, labeled INT.
- Twitter technology dataset contains 845 fake accounts, labeled TWT.

Researchers have acknowledged that their fake followers dataset is just illustrative, and not exhaustive, of all the possible existing sets of fake followers.

3.2.1. Feature set proposed by Yang et al.. Authors of [37] had make an analysis of the evasion tactics that were utilized by Twitter spammers. They observed that Twitter spammer used to change their behaviour to evade spam detection techniques, So they suggested to design a new features that would enhance detecting spammers and would be harder for them to evade. They had combined their new features in four machine learning classifiers and compared their implementation with other existing approaches.
In Table 1, eight of the highly contributed features in the detecting spammers was detailed. As the MIB dataset is different than the dataset used in [37], feature 6 was approximated to $friends/age$ as presented in [27],

1	age of the account
2	bidirectional link ratio
3	average neighbors followers
4	average neighbors tweets
5	followings to median neighbors followers
6	following rate
7	API ratio
8	API URL ratio

TABLE 1. FEATURE SET PROPOSED BY YANG ET AL. [37]

4. Proposed Algorithm

This section presents the proposed methods of predicting fake Twitter accounts. Proposed methods are divided into four main parts: data pre-processing, feature reduction, data classification, and accuracy comparison; aiming to develop a new technique that achieves a high classification accuracy results in a reasonable time as mention in Figure 1.
The data Pre-Processing model was used to make the provided data in a suitable format for classification, as the used classifiers need numerical data. Also some explicit features provided by literature called "Yang et al. feature set" as detailed in Section 3.2.1, were extracted from the original dataset in this model.
Second the pre-precessed data will be reduced in the feature reduction phase, where three feature reduction techniques were used, the ambition of applying feature reduction was to decrease the number of features, and get ride of the ineffective or miss behaving feature, which will improve the classification model results consequently.
In the classification model, three classification algorithms were used to train and test the three datasets conducted from the feature reduction model, besides the Yang et al. feature set, and the original provided dataset with no selection.
Finally the resulted accuracy from the classification phase will be compared and discussed.

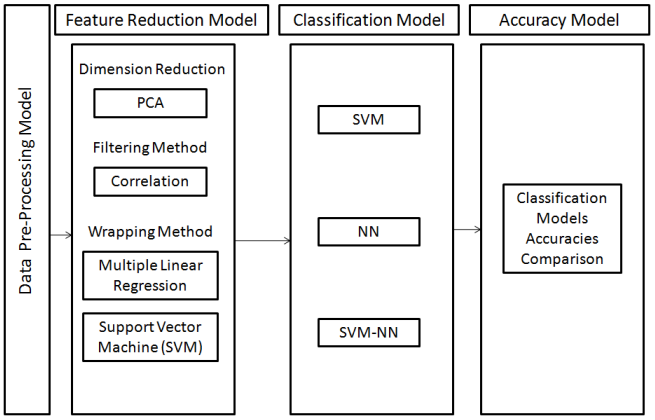

Figure 1. Design approach to calculate accuracy rates for data mining techniques

4.1. Data Pre-Processing

The "MIB" dataset feature vectors are presented in two types:

- Categorical features: for example, language, profile-sidebare-color, tweets..

- Numerical features: for example, friends-count, followers-count, default-profile, profile-use-background-image,..

We have converted six categorical features into numerical so we could apply classification algorithms on them. Feature label was added to distinguish between real and fake accounts. The pre-processing step resulted 16 numerical feature vectors that describe users behaviours on Twitter as listed in Table 2.

1	Statuses-count	9	Default-profile-banner-url
2	Followers-count	10	Profile-use-background-image
3	Friends-count	11	Profile-text-color
4	Favourites-count	12	Profile-sidebar-border-color
5	Listed-count	13	Profile-background-tile
6	Geo-enabled	14	Profile-sidebar-fill-color
7	Default-profile	15	Profile-background-color
8	Default-profile-image	16	Profile-link-color

TABLE 2. MIB DATASET FEATURE VECTORS

4.2. Feature Reduction

In feature reduction phase, four data reduction techniques were applied to guide the process of deciding the most promising feature patterns to be used in the mining process as shown in figure 1.

- PCA
- SpearmansRank-Order Correlation
- Wrapper Feature Selection using SVM
- Multiple Linear Regression

Those techniques are discussed in the following subsections.

4.2.1. Principal Component Analysis "PCA". As we mentioned in section 2.2, PCA is a dimension reduction technique that is used to reduce feature vector dimensions, it finds the top N features that best describe the data and covers as much variance of it, stripping out the unnecessary features by assigning them a lower weight so they would not impact the mining process. In this work 10 PCAs have been selected out of 16 PCAs, those 10 PCAs cover around 92% of the data.

4.2.2. Spearman's Rank-Order Correlation. Spearman's Rank-Order Correlation is one of the feature selection filtering methods [38]. It measures the strength and direction of the monotonic relationship between two quantitative variables X and Y. Each of these correlation measures is exactly zero when X and Y are independent, and have values that range between -1 and +1 to indicate the level and direction of the correlation. The output of this algorithm is a Table containing the correlation coefficients between each variable and the other variables.

4.2.3. Relevance and Redundancy Analysis Technique. Relevance and Redundancy analysis technique is used for feature selection [39]. In relevance analysis step, Spearman's Rank-Order correlation was used to eliminate all pairs of features whose level of correlation to the class variable is below our estimated threshold which was "0.8". The output of this step was 11 sets of pairs of correlated features shown in Table 3.
The selected features in the relevance analysis step are used as an input to the redundancy analysis step. It is widely accepted that two features are redundant to each other if their values are completely correlated but in reality it is not so straightforward to determine feature redundancy when a feature is correlated with a set of features. Hence, Markov Blanket technique was employed to eliminate redundant features [40].

#	feature 1	feature 2
set1	2	1
set 2	4	1
set 3	10	7
set 4	14	11
set 5	16	11
set 6	14	12
set 7	15	12
set 8	16	12
set 9	15	14
set 10	16	14
set 11	16	15

TABLE 3. PAIRS OF CORRELATED FEATURES THAT HAVE CORRELATION COEFFICIENT >0.8

4.2.4. Markov Blanket Technique. The Markov Blanket for a node A, MB(A), in a Bayesian network is the set of nodes composed of A's parents, its children, and the other parents of its children [39], [40], [41].
In a Markov random field, the Markov Blanket of a node is its set of neighboring nodes. After applying Markov blanket on the pairs of correlated features MB(Fi) and MB(Fj) we had two versions of two output sets of non-redundant features shown in Figure 2. The output feature sets from the redundancy step will be used later in the classification phase section 4.3.

#	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16
Set 1	1	0	0	0	0	0	1	0	0	0	1	1	0	1	1	0
Set 2	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1	1

Figure 2. Two selected feature sets where the contributing features represented by 1 and ignored features represented by 0 overall 16 features.

4.2.5. Wrapper Feature Selection using SVM. One of the well known feature selection methods is wrapper methods [38], where different feature subsets are selected and qualified by a learning model. The features subset with highest predictive performance would be selected. All subsets of a set can be found using bit manipulation, there will be 2^n subsets for a given set, where n is the number of features F, in a set. For example, there will be 2^3 subsets for the set $\{1, 2, 3\}$ as shown in Figure 3.
This method usually provides the best performing feature set for that particular learning model, but for large feature

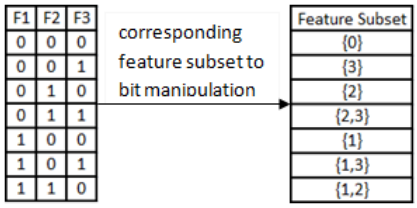

Figure 3. All available subsets for the set {1,2,3} presented using bit manipulation, Each 1 in the binary representation indicate an element in that position.

space it might need intensive computational requirements. In our baseline dataset we had 16 feature vectors which means 2^{16}-1=65,535 feature subsets without the empty subset. The baseline dataset has been splited into 70% training and 30% testing. Then, all of the 65,535 feature subsets have been trained and tested using Support vector machine "SVM" as illustrated in Figure 4. Feature subsets that provided

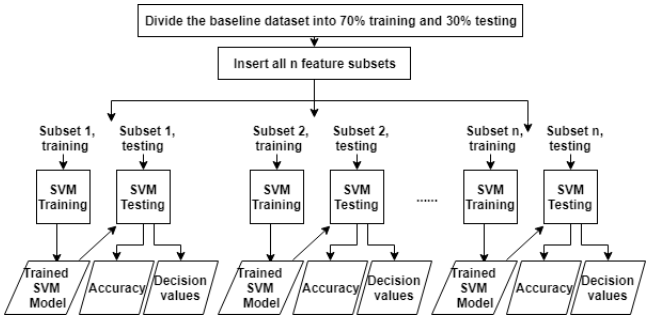

Figure 4. Training and testing all feature subsets using SVM

a predictive model accuracy greater than 95% have been selected as shown in Figure 5. In Table 4, the top 5 selected

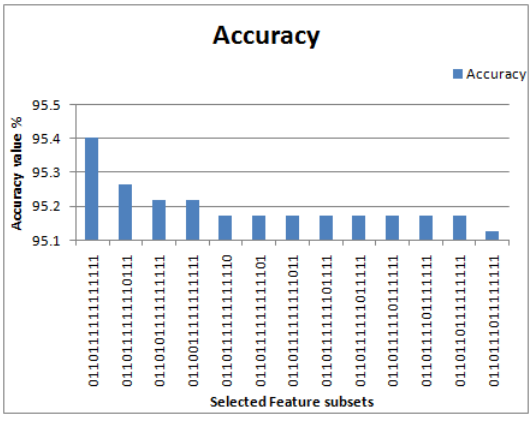

Figure 5. 14 feature subsets with performance accuracy >95%

feature subsets were listed. It is clear that feature 1 and feature 4 were not used by any feature subset.

4.2.6. Wrapper Feature Selection Using Multiple Linear Regression. Linear Regression models describe the relationship between a response output, dependent variable, and one

#	set1	set2	set3	set4	set5
F1	0	0	0	0	0
F2	1	1	1	1	1
F3	1	1	1	1	1
F4	0	0	0	0	0
F5	1	1	1	0	1
F6	1	1	0	1	1
F7	1	1	1	1	1
F8	1	1	1	1	1
F9	1	1	1	1	1
F10	1	1	1	1	1
F11	1	1	1	1	1
F12	1	1	1	1	1
F13	1	0	1	1	1
F14	1	1	1	1	1
F15	1	1	1	1	1
F16	1	1	1	1	0

TABLE 4. TOP 5 SELECTED FEATURE SUBSETS USING WRAPPER TECHNIQUE WITH SVM

or more predictor input, independent variables. In a simple linear regression, there are two variables x and y, where y depends on x or influenced by x as shown in Equation 1.

$$y = a + bx \tag{1}$$

Where, a is a constant, b is the regression coefficient.
In multiple linear regression, two or more independent variables are used to predict the value of a dependent variable as shown in Equation 2.

$$y = a + bx1 + cx2 + dx3 \tag{2}$$

Multiple linear regression was used, as MIB dataset have 16 dependent variables and one independent variable. The problem with multiple linear regression is **multicollinearity**, where the model includes multiple factors that are correlated not just to the response variable, but also to each other.
Multicollinearity increases the standard errors of the coefficients, which makes some variables statistically insignificant where otherwise, they should be significant. Removing redundant variables resulted in 12 predictors instead of 16 predictors.
As discussed earlier in Section 4.2.5, all the possible feature subsets were extracted using the formula $(2^{12} - 1)$, and applied to regression model.
Regression model performance could be measured according to two attributes, R^2, and P-value [42], [43].
R^2: is a statistical measure of how close the data are to the fitted regression line. In other words the higher the R-squared, the better the data would be fitted by the regression model. R^2 was specified to be greater than 0.57%.
P-value: indicates the null hypothesis of the regression model, the smaller the p-value, the better the model fits the data, it was assigned to be less than 0.01. The output of this step was 106 qualified subsets, the top 5 feature subsets are listed in Table 5. As explained in this section each feature reduction technique deals with the data from different perspective result in diverse datasets. If those resulted datasets combined together the output will be the original proposed feature set with out reduction, which will turn us back to the original starting point.

#	set1	set2	set3	set4	set5
F1	0	0	0	0	0
F2	1	1	1	1	1
F3	1	1	1	1	1
F4	1	1	1	1	1
F5	1	1	1	1	1
F6	1	1	1	1	1
F7	0	0	0	0	0
F8	1	1	1	1	1
F9	1	1	1	1	1
F10	1	1	0	1	0
F11	1	0	0	0	0
F12	0	0	0	0	0
F13	1	0	1	1	1
F14	1	1	0	0	1
F15	1	1	1	1	1
F16	0	1	1	1	1

TABLE 5. TOP 5 SELECTED FEATURE SUBSETS USING WRAPPER TECHNIQUE WITH MULTIPLE LINEAR REGRESSION

4.3. Data Classification

The five feature sets that resulted after applying our feature reduction techniques in section 4.2, were trained and tested using SVM-NN newly developed classification algorithm as shown in Figure 1. A cross validation of 10 and 8-folds were used for estimating the performance for each classifier.

4.3.1. SVM-NN classification. As proposed in literature Sybil accounts have different characteristics compared to normal users. Hence, researcher explored the possibility of distinguishing normal and Sybil accounts using some classification algorithm like SVM, and NN as mentioned in Section 2.

As a potential for improving the classification accuracy, a new algorithm named SVM-NN has been developed, where the SVM trained model decision values were used to train a NN model, and SVM testing decision values were used to test the NN model as shown in Figure 6. In other words, a hybrid classification algorithm was used, by running the Neural Network classification algorithm on the decision values resulting from the SVM classification algorithm as shown in Algorithm 1.

The five feature sets presented in Section 4.2, were used as an input to train, and test the SVM-NN classifier.

Radial Basis Function (RBF) was exploited as SVM classifier kernel, and it was trained using libSVM machine learning algorithm [44]. For NN feed-forward back propagation has been selected as the base algorithm, the NN trained using 7 neurons and one hidden layer.

In each one of the feature sets, it was noticed that there is a feature subset that provides a better classification accuracy compared with the other subsets. As in Spearman's rank-order Correlation best feature subset was (00101000001000111), Multiple linear Regression best feature subset was (0011110110001111), and in Wrapper-SVM best feature subset was (00110111111110111).

Algorithm 1: SVM-NN

Result: feature subsets classification accuracy

Identify list of reduced features using PCA, Correlation, Regression, SVM;
Set feature subsets to *s*;
Split your data into testing and training using 8 cross validation;
Set the training identifying labels to *rLable* Set the testing identifying labels to *sLable*
for *each s* **do**
    Use SVM classification algorithm to Train the model using the training set, and the identifying labels *rLable*.
    Predict the output using the SVM trained model, and set the output decision-values to *decisionV*
    Train NN model using *decisionV*, and the identifying labels *rLable* .
    Predict the testing set output using the **SVM trained model**, and set the output decision-values to *testingDecisionV*
    Test NN using the *testingDecisionV*, and **NN trained model**, set the output to *nnPredicted*
    Calculate NN prediction for each s accuracy using the *sLable*, and *nnPredicted*
end
calculate the average accuracy for each fold

5. Performance and Evaluation

In this section we present the results of the proposed algorithm and discuss them.

Initially, three different classification algorithms have been trained and tested using divergent four feature sets.

Neural network classification algorithm and SVM classification algorithm were used as the principles mining techniques in many social network researches, so they have been applied on the feature sets mentiond in Section 4.2 and compared with the proposed SVN-NN algorithm.

5.1. SVM classification

As proposed in related work Section 2, researchers used SVM classification algorithm to distinguish between Sybil accounts, and real accounts. Hence, SVM were applied on the provided dataset and compared with NN and, SVM-NN. Radial Basis Function (RBF) was exploited as SVM classifier kernel, and it was trained using libSVM machine learning algorithm [44]. It was noticed that there is a feature subset that has the maximum prediction accuracy results compared with the other subsets. As in Spearmans rank-order correlation best pattern was (1 0 0 0 0 0 1 0 0 0 1 1 0 1 1 0), Multiple linear Regression best pattern was (0011110111001111), and Wrapper-SVM best pattern was (0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1). The detailed accuracy results of this experiment are presented in Figure 7.

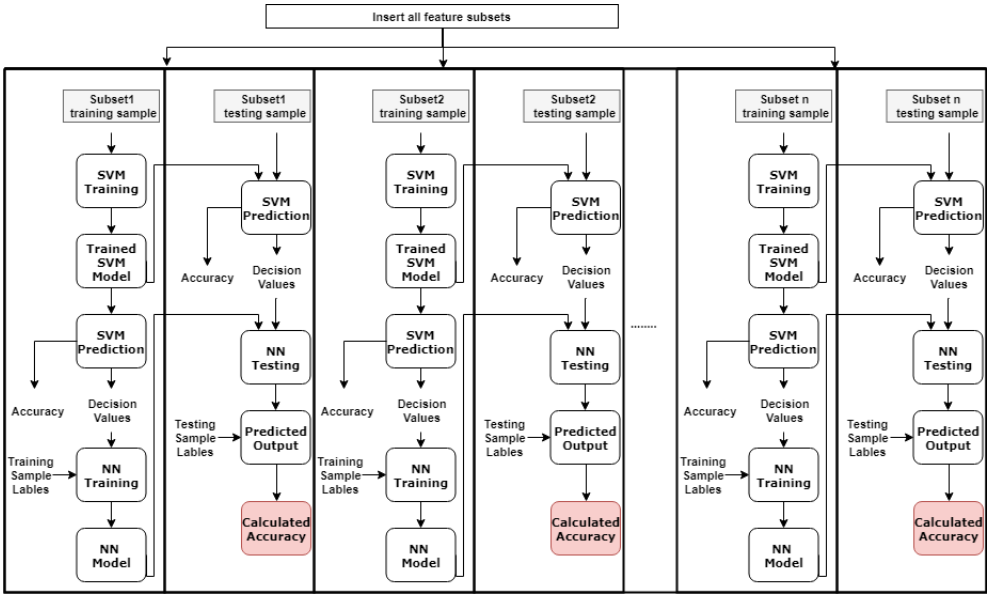

Figure 6. SVM-NN algorithm

Feature Set	SVM			NN			SVM-NN		
	accuracy	False-Positive	False-Negative	accuracy	False-Positive	False-Negative	accuracy	False-Positive	False-Negative
Yang et al.	0.886	0.111	0.001	0.737	0.059	0.203	**0.912**	0.086	0.001
PCA	0.914	0.039	0.046	0.653	0.278	0.067	**0.922**	0.033	0.043
Correlation	0.923	0.036	0.046	0.822	0.079	0.097	**0.983**	0.013	0.003
Regression	0.947	0.035	0.016	0.888	0.04	0.071	**0.96**	0.027	0.011
Wrapper-SVM	0.956	0.039	0.004	0.833	0.052	0.114	**0.965**	0.027	0.007

Figure 7. Accuracy results of applying SVM, NN, and SVM-NN on the proposed feature sets

5.2. Neural Networks

Currently, there are many neural network algorithms used to train models and predict results based on the previously trained models. The same model structure with the SVM-NN in Section 4.3.1, applied here as Feed-forward back propagation algorithm has been selected as the base algorithm with one hidden layer, and 7 neurons. The predicted results have been compared with the actual legitimate values (i.e. whether the account is real or fake). Unlike the SVM algorithm, NN dose not calculate the prediction accuracy implicitly, so the prediction accuracy has to be calculated separately using the following formula:

$$\%Accuracy = \frac{All\ correctly\ identified\ accounts}{Total\ number\ of\ accounts} \times 100$$

As mentioned above the feature subsets with highest accuracy was highlighted, as following:
spearmans rank-order Correlation best pattern was (1000001000110110), Multiple linear Regression best pattern was (0110110111001111), Wrapper-SVM best pattern was (0110111111011111). NN accuracy results are illustrated in Figure 7.

As shown in Figure 7, the results show that SVM classifier has the highest accuracy while using Wrapper-SVM feature set and the lowest accuracy was with Yang et al. feature set. while the accuracy results for NN classifier were lower than their counterparts using SVM classifier, with highest accuracy 0.888 from regression feature set and lowest accuracy using PCA feature set.
By comparing the accuracy results of all the three classification algorithms, it was illuminated that SVM-NN classification algorithm has the highest classification accuracy results on all the feature subsets compared with the other two previous classifiers as in Figure 8, with highest accuracy 0.983.

6. Conclusions and Future Work

In this paper a new classification algorithm was proposed to improve detecting fake accounts on social networks, where the SVM trained model decision values were used to train a NN model, and SVM testing decision values were used to test the NN model.
To reach our goal we used ”MIB” baseline dataset from [27] and run it into pre-processing phase where different feature

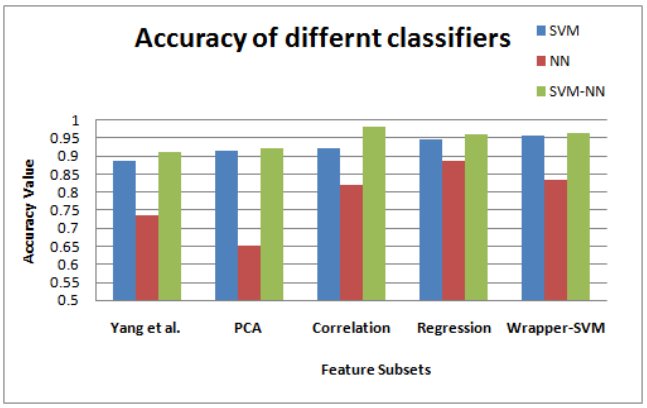

Figure 8. SVM-NN classifier has the highest classification accuracy results on all the feature subsets compared with the SVM, NN classifiers

reduction techniques were used to reduce the feature vector. In the classification phase learning algorithms were used.

The results of the analyses showed that "SVM-NN" has archived better accuracy results with all feature sets comparing with the other two classifiers, with classification accuracy around 98%. It was noticed that the NN algorithm has the lowest classification accuracy compared with SVM, and SVM-NN. This occurred because the SVM algorithm reaches the global minimum of the optimized function [45], while the NN used the gradient descent technique, and may reach the local minimum, not global minimum like SVM [46].

It was also noticed that using the feature set provided by PCA, encountered a very low classification accuracy, while the correlation feature set achieves high classification accuracy. This happened because PCA performs dimension reduction and generate a new features base on linear combination of original features. But the correlation approach, and other feature selection techniques select the best set of original features, not linear combination of all features.

On other words, feature selection selects the most effective original features, but PCA performs a linear combination of the original features event they are not effective.

The correlation feature set records a remarkable accuracy among the other feature selection technique sets, because correlation technique not only select the best features, but also removes the redundancy.

References

[1] (2018) Political advertising spending on facebook between 2014 and 2018. Internet draft. [Online]. Available: https://www.statista.com/statistics/891327/political-advertising-spending-facebook-by-sponsor-category/

[2] (2018) Quarterly earning reports. Internet draft. [Online]. Available: https://investor.fb.com/home/default.aspx

[3] (2018) Statista.twitter: number of monthly active users 2010-2018. Internet draft. [Online]. Available: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[4] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egyptian informatics journal*, vol. 17, no. 2, pp. 199–216, 2016.

[5] L. M. Potgieter and R. Naidoo, "Factors explaining user loyalty in a social media-based brand community," *South African Journal of Information Management*, vol. 19, no. 1, pp. 1–9, 2017.

[6] Y. Boshmaf, D. Logothetis, G. Siganos, J. Lería, J. Lorenzo, M. Ripeanu, K. Beznosov, and H. Halawa, "Íntegro: Leveraging victim prediction for robust fake account detection in large scale osns," *Computers & Security*, vol. 61, pp. 142–168, 2016.

[7] (2013) Banque populaire dis-moi combien damis tu as sur facebook, je te dirai si ta banque va taccorder un prłt. Internet draft. [Online]. Available: http://bigbrowser.blog.lemonde.fr/2013/09/19/popularite-dis-moi-combien-damis-tu-as-sur-facebook-je-te-dirai-si-ta-banque-va-taccorder-un-pret/

[8] J. R. Douceur, "The sybil attack," in *International workshop on peer-to-peer systems*. Springer, 2002, pp. 251–260.

[9] (2012) Cbc.facebook shares drop on news of fake accounts. Internet draft. [Online]. Available: http://www.cbc.ca/news/technology/facebook-shares-drop-on-news-of-fake-accounts-1.1177067

[10] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake osn accounts by predicting their victims," in *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security*. ACM, 2015, pp. 81–89.

[11] (2018) Facebook publishes enforcement numbers for the first time. Internet draft. [Online]. Available: https://newsroom.fb.com/news/2018/05/enforcement-numbers/

[12] (2018) How concerned are you that there are fake accounts and bots on social media platforms that are used to try to sell you things or influence you? Internet draft. [Online]. Available: https://www.statista.com/statistics/881017/fake-social-media-accounts-bots-influencing-selling-purchases-usa/

[13] (2012) Buying their way to twitter fame. Internet draft. [Online]. Available: www.nytimes.com/2012/08/23/fashion/twitter-followers-for-sale.html?smid=pl-share

[14] (2017) Welcome to the era of the bot as political boogeyman. Internet draft. [Online]. Available: https://www.washingtonpost.com/news/politics/wp/2017/06/12/welcome-to-the-era-of-the-bot-as-political-boogeyman

[15] (2018) Human or 'bot'? doubts over italian comic beppe grillo's twitter followers. Internet draft. [Online]. Available: https://www.telegraph.co.uk/technology/twitter/9421072/Human-or-bot-Doubts-over-Italian-comic-Beppe-Grillos-Twitter-followers.html

[16] (2017) How fake news and hoaxes have tried to derail jakarta's election. Internet draft. [Online]. Available: https://www.bbc.com/news/world-asia-39176350

[17] S.-T. Sun, Y. Boshmaf, K. Hawkey, and K. Beznosov, "A billion keys, but few locks: the crisis of web single sign-on," in *Proceedings of the 2010 New Security Paradigms Workshop*. ACM, 2010, pp. 61–72.

[18] S. Fong, Y. Zhuang, and J. He, "Not every friend on a social network can be trusted: Classifying imposters using decision trees," in *Future Generation Communication Technology (FGCT), 2012 International Conference on*. IEEE, 2012, pp. 58–63.

[19] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in *Proceedings of the 27th annual computer security applications conference*. ACM, 2011, pp. 93–102.

[20] P. Patel, K. Kannoorpatti, B. Shanmugam, S. Azam, and K. C. Yeo, "A theoretical review of social media usage by cyber-criminals," in *Computer Communication and Informatics (ICCCI), 2017 International Conference on*. IEEE, 2017, pp. 1–6.

[21] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 8, pp. 1311–1321, 2014.

[22] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended accounts in retrospect: an analysis of twitter spam," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 243–258.

[23] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: mapping the spread of astroturf in microblog streams," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 249–252.

[24] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 15–15.

[25] L. Alvisi, A. Clement, A. Epasto, S. Lattanzi, and A. Panconesi, "Sok: The evolution of sybil defense via social networks," in *Security and Privacy (SP), 2013 IEEE Symposium on*. IEEE, 2013, pp. 382–396.

[26] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin." in *PACIS*, 2014, p. 278.

[27] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[28] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao, "You are how you click: Clickstream analysis for sybil detection." in *USENIX Security Symposium*, vol. 9, 2013, pp. 1–008.

[29] S. Fong, Y. Zhuang, and J. He, "Not every friend on a social network can be trusted: Classifying imposters using decision trees," in *Future Generation Communication Technology (FGCT), 2012 International Conference on*. IEEE, 2012, pp. 58–63.

[30] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks." in *USENIX Security Symposium*, 2014, pp. 223–238.

[31] a social network used in china. Internet draft. [Online]. Available: http://www.renren-inc.com/en/

[32] (2012) How to recognize twitter bots: 7 signals to look out for. Internet draft. [Online]. Available: http://www.stateofdigital.com/how-to-recognize-twitter-bots-6-signals-to-look-out-for/

[33] (last check 2018) Fake followers check: A new free tool from socialbakers. Internet draft. [Online]. Available: https://www.socialbakers.com/blog/1099-fake-followers-check-a-new-free-tool-from-socialbakers?showMoreList-page=1

[34] A. S. M. Salih and A. Abraham, "Novel ensemble decision support and health care monitoring system," *Journal of Network and Innovative Computing*, vol. 2, no. 2014, pp. 041–051, 2014.

[35] A. Mackiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers and Geosciences*, vol. 19, pp. 303–342, 1993.

[36] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, no. 4. ACM, 2004, pp. 219–230.

[37] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.

[38] E. P. Xing, M. I. Jordan, R. M. Karp *et al.*, "Feature selection for high-dimensional genomic microarray data," in *ICML*, vol. 1. Citeseer, 2001, pp. 601–608.

[39] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research*, vol. 5, no. Oct, pp. 1205–1224, 2004.

[40] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery." in *FLAIRS conference*, vol. 2, 2003, pp. 376–380.

[41] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[42] L. E. Eberly, "Multiple linear regression," in *Topics in Biostatistics*. Springer, 2007, pp. 165–187.

[43] L. S. Aiken, S. G. West, and S. C. Pitts, "Multiple linear regression," *Handbook of psychology*, pp. 481–507, 2003.

[44] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[45] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for svm and the training of non-psd kernels by smo-type methods," *submitted to Neural Computation*, vol. 3, pp. 1–32, 2003.

[46] N. B. Karayiannis, "Reformulated radial basis neural networks trained by gradient descent," *IEEE transactions on neural networks*, vol. 10, no. 3, pp. 657–671, 1999.