Journal of Theoretical and Applied Information Technology
15th April 2019. Vol.97. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVALUATING THE ACCURACY OF DATA MINING ALGORITHMS FOR DETECTING FAKE FACEBOOK PROFILES USING RAPIDMINER, WEKA, AND ORANGE

[1]MOHAMMED BASIL ALBAYATI1, [2]AHMAD MOUSA ALTAMIMI, [3]DIAA MOHAMMED ULIYAN

[1,2] Department of Computer Science, Applied Science Private University, Jordan

[3]Department of Computer Science, Middle East University, Amman, Jordan

E-mail: [1]mohammed.sabri@asu.edu.jo, [2]a_altamimi@asu.edu.jo, [3]duliyan@meu.edu.jo

ABSTRACT

Facebook is on constant growing, attracting more users due to the provided high-quality services for Online Socializing, Sharing Information, Communication, and alike. Facebook manages data for billions of people and is therefore be a target for attacking. As a result, sophisticated ways for infiltrating and threatening this platform have been developed. Fake profiles, for instance, are created for malicious purposes such as financial fraud, impersonate identities, Spamming, etc. Numerous studies have investigated the possibility of detecting fake profiles on Facebook with each study focusing on introducing a new set of features and employing different machine learning algorithms for countermeasure. This paper adopts a set of features from the previous studies and introduces additional features to improve classification performance in order to detect fake profiles. The performance of five supervised algorithms (Decision Tree, Support Vector Machine SVM, Naïve Bayes, Random Forrest, and k Nearest Neighbour k-NN) are evaluated across three of the common mining tools (RapidMiner, WEKA, and Orange). The experimental results showed that SVM, Naïve Bayes, and Random Forest had a stable performance with a nearly identical results across the three mining tools. However, Decision Tree outperformed other classifiers on RapidMiner and WEKA with accuracy of 0.9888 and 0.9827, respectively. Finally, we observed that k-NN showed the most significant change with an accuracy of 0.9603 for WEKA, 0.9145 for Orange, and 0.9460 for RapidMiner tool. These findings would be useful for researchers willing to develop a machine learning model to detect malicious activities on social network.

Keywords: *Fake Profiles Detection, Machine Learning, Classification, WEKA, RapidMiner, Orange*

1. INTRODUCTION

Facebook is social online platform that allows the users to interact and stay connect, locally and globally with no boundaries, with others through virtual portals of communication, entertainment, sharing information, and several interaction tools such as: posting, tagging, sharing, and many others [1-4]. As a result, Facebook is now considered as the largest social online platform, according to Facebook's administration news room, Facebook added 22 million new daily users in the second quarter of 2018 with 2.23 billion monthly active users [2].

Despite the beneficial services derived from Facebook, it became a target for the attackers to breach user's privacy. Recently, Facebook has announced that a security exploit allowed attackers to gain control of at least 50 million user accounts. Facebook says they have fixed the vulnerability and taken steps to protect other users who could have been impacted. "*We're taking this incredibly seriously*," Guy Rosen, Facebook's vice president of product management, wrote on the company's behalf [5]. To name but a few, Scammers create fake profiles to swindle users to do various kinds of suspicious activities such as cloning, spamming, deceiving, etc. [3]. That's how bad things have gotten with online security in general, and with Facebook in particular.

In this regard, Facebook urges its users through its "Statement of Rights and Responsibilities" to create safe environment for communicating and sharing information by creating profiles that represented them and use an authentic information to reflect their real identities [6]. Despite that, Fake users adopt different methods and techniques for creating their fake profiles on Facebook.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

In fact, different works have been proposed for handling this problem. However, the machine learning is the most noticeable approach utilized in the literature, especially the supervised learning. Zuckerberg's calls for artificial intelligence solutions for piracy, abuse, misinformation and other rhythms amount to magical thinking [5].

This proposition was reaffirmed by many works proposed in the literature. For instance, the work in [7] proposed a supervised machine learning approach for detecting fake/phantom profiles on social online game hosted by Facebook, which implemented by RapidMiner mining tool. In the same manner, authors of [8] proposed behavioral and community-based attributes for detecting fake/spam profiles on Facebook using Naïve Bayes, J48, k-NN, and Decision Tree classifiers. These classifiers are implemented using a WEKA mining tool. More works are discussed in detail in Section two.

In this respect, many data mining tools have been developed to implement data mining classifiers. One can consider the most common and open source tools (RapidMiner, WEKA, and Orange). Every tool has its own advantages and disadvantages [9]. The following shows brief analytics of feature for these three tools [10]:

1) RapidMiner is a statistical analysis, data mining, predictive analytics tool. Its main features are:
   - More than 20 new functions for analysis and data handling, including multiple new aggregation functions.
   - File operators to operate directly from Rapid Miner.
   - A macro viewer that shows macros and their values in real time during process execution.
   - Intuitive GUI.

2) WEKA is a machine learning tool. Its main features are:
   - Forty-nine data preprocessing tools, seventy-six classification/regression algorithms, eight clustering algorithms, fifteen attribute/subset evaluators, ten search algorithms for feature selection.
   - Three algorithms for finding association rules.
   - Three graphical user interfaces – "The Explorer" (exploratory data analysis) – "The Experimenter" (experimental environment) – "The Knowledge Flow" (new process model inspired.
   - With poor documentation.

3) ORANGE is a machine learning, Data mining, data visualization tool. Its main features are:
   - Visual Programming, Visualization
   - Interaction and Data Analytics
   - Large toolbox, Scripting interface
   - Extendable Documentation

As the number of available tools continues to grow, the choice of most suitable one becomes increasingly difficult [9]. In this work, we present a comparative analysis to evaluate the performance of five of traditional classifiers (Decision Tree, Support Vector Machine SVM, Naïve Bayes, Random Forrest, and k Nearest Neighbor k-NN) across three of the most common and open source tools (RapidMiner, WEKA, and Orange).

To ground our conceptual idea, a set of attributes that have been proposed previously are adopted here. Moreover, additional attributes (No. of Posts and No. of Tags) are proposed to conduct the experiments. Regarding the data set size, 981 profiles have been collected specifically for this study. Results showed that SVM, Naïve Bayes, and Random Forest had a close accuracy rates by implemented them using the three mining tools. Precisely, (0.9430, 0.9582, 0.9572) for the SVM, (0.9735, 0.9613, 0.9705) for Naïve Bayes, and (0.9857, 0.9929, 0.9857) for the Random Forest, respectively.

On the other hand, Decision tree showed a nearly identical results with less than %1-3 percentage change when implemented it using the three mining tools with the accuracy rates of 0.9888, 0.9827, and 0.9644 for the RapidMiner, WEKA, and Orange mining tool. Regarding the most significant change among all the mining tools, we observed that k-NN outperformed in WEKA tool over the RapidMiner and Orange with the accuracy of 0.9603 over the accuracy of 0.9460 and 0.9145, respectively.

As a final observation, the newly added attributes (No. of Posts and No. of Tags) contribute in improving the detection rates in the case of Decision Tree and k-NN when implemented using RapidMiner. For example, the accuracy rates jumped from 0.8400 to 0.9460, and from 0.9650 to 0.9888 for the k-NN and Decision Tree algorithm, respectively after adding these new attributes. In contrast, the accuracy rates of Naïve Bayes and SVM are decreased with the accuracy rate of 0.9750 to 0.9705, and from 0.9850 to 0.9572 for Naïve Bayes and SVM algorithms after adding these new attributes.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The remaining of this paper structured as follow: the literature is discussed in Section 2. The background materials are given in section 3, while section 4 explains our proposed work. Section 5 discusses the experiments and the obtained results. Finally, the conclusion of our work is presented in section 6.

2. LITERATURE REVIEW

The term of Facebook *fake profiles* represents all forms of the profiles that had been created based on fake information to perform all types of suspicious activities: spamming, deceiving, fraud, etc. [6].

To detect such profiles, many methods have been proposed in the literature. In fact, we observed that the supervised machine learning algorithms are the most utilized approach in this regard [7-8], [10-14]. These algorithms used the users' information stated on their online profiles for distinguishing the fake profiles out of the real ones and implemented using different mining tool such as: RapidMiner, Weka, and Orange tools [7-8]. One can consider for example, the work presented in [7], where the authors proposed a supervised machine learning model for detecting fake profiles in online social gaming applications hosted by Facebook. The research focused on the statistical differences among a subset of 13 attributes regarding the social activates of the users. The classification experiment focused on Support Vector Machine algorithm, implemented using the WEKA tool, and elaborates many of the machine learning techniques to improve some of the issues regarding the features' selection process of this work.

Authors of [8] proposed a framework for detecting spammers/ fake profiles on online social sites, using Facebook as a test case in a machine learning approach by exploiting a behavioral and graph-based attributes that include the structure of the nodes and some topological information of the users' profiles in the network. The framework implemented using WEKA tool as mining environment, employing ten discriminative topological attributes and a set of supervised algorithms including J48, Addtree as Decision Tree based classifiers, Naïve Bayes, and k-NN with k=5.

Another example is presented in [11] in which four classifiers (Support Vector machine, Artificial Neural Network, Naïve Bayes, and Decision Tree) were used to build a detection model. In addition, an ensemble method with a majority voting for these algorithms have been applied for increasing the accuracy of the model implemented using R language. Following the same vein, the work in [12] utilized the supervised machine learning techniques (SVM, Naïve Bayes, and Decision Tree) for detecting fake profiles on Facebook. These techniques are implemented using Python scripts to scrape the profile attributes (e.g., No. of friends, Education and work, Gender, No. of columns filled in about me, etc.).

Authors of [13] Presented a Social Privacy Protector (SSP) software for detecting fake profiles on Facebook. The SSP consists of three protection layers (Friends analyzer, Privacy protector, and HTTP server). Here, the SSP software scans the user's friends list and returns a credibility score in which each friend is analyzed by machine learning algorithms, which considers the strength of the connection between the user and his/her friends. The strength of each connection is based on a set of fifteen connection features such as the number of common friends, and the number of pictures and videos the user and his friend are tagged in together. WEKA mining tool is used for implementing this work's theme by employing eight supervised algorithms (e.g. Naïve Bayes, Bagging, Random-Forest, J48, and others).

Authors of [14] proposed a model to evaluate the performance of four machine learning algorithms (Random Forest, Support Vector Machine SVM, k Nearest Neighbor k-NN, and Multilayer Perceptron MLP) for spam profile detection. These machine learning algorithms are tested across two mining tools (WEKA and RapidMiner). The conducted experiments showed that SVM, k-NN, and MLP on WEKA outperformed those algorithms on RapidMiner. However, RF achieved higher accuracy on RapidMiner compare to WEKA.

Finally, an empirical study for detecting fake Facebook profiles using supervised learning proposed is presented in [15]. Here, four mining algorithms that is SVM, Naïve Bayes, k-NN, and Decision Tree are employed in the detection model and implemented using RapidMiner mining tool with a dataset of 200 profiles, collected from the authors' profile and a honeypot page, to demonstrate the validity of the proposed model.

3. BACKGROUND MATERIAL

Five algorithms are selected for the comparative analysis, and three mining tools are utilized here. Before going into the experiments, themselves, we give a brief review about them.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.1 The Mining Tools

3.1.1 RapidMiner [9] [16]

It is an open source data science platform, developed by company of the same name. It is visual workflow for predictive analytics used for wide range of application such as: Educational, Business, industrial, and others. The tool itself is written in the Java programming language, uses GUI to design and execute mining workflows that consist of multiple components called "Operators", each one performs a single task within the workflow. RapidMiner brings artificial intelligence to the enterprise through an open and extensible data science platform contains of more than 100 learning schemes for regression classification and clustering analysis and it supports about twenty-two file formats through unifying the entire data science lifecycle from data prep to machine learning to predictive model deployment.

3.1.2 WEKA [9] [17]

Weka stands for Waikato Environment for Knowledge Analysis. The tool is a Java based open source tool, performing so many mining tasks that include pre-processing, classification, clustering, and association rule extraction. It provides User graphical interface for executing mining workflows, and simple Command-line explorer for typing commands is also provided. WEKA supports preprocessing, attribute selection, learning, visualization and much more mining methods and techniques. And it works could be extendable to be included with other java package libraries.

3.1.3 Orange [18]

It is a graphical user interface developed as component-based mining software. It provides data mining with the use of visuals. It can be data visualizations along with analysis created possible for novice users in addition to the data experts. Users can easily layout info analyses by means of visual programming and also Python scripting. It written in Python, featuring a visual programming component for Python programming and libraries for scripting. Orange tool includes a set of components for data preprocessing, scoring and filtering, modelling, evaluation, and exploration techniques. It performs an effective a mining and learning tasks in fast and simple pace.

3.2 The Mining Algorithms

3.2.1 Decision Tree [19]

Decision Tree is a predictive model takes a tree structure that generates the classification rule by breaking down the dataset into smaller and smaller subset until the decision node (class label) is met. Each node in the tree represents an attribute of the training set, however, leaf nodes hold the class label (final outcome), while the root node represents the attribute with highest information gain that determines the tree branches in which each branch represents one of the outcomes of the model. Decision tree is a recursive algorithm that calculates the attributes' information gain in each iterative and selects the attribute with higher information gain (most dominant) to split the tree, Thereafter, entropy and gain scores would be calculated again among the other attributes. Thus, the next most dominant attribute is found. this process repeated until reaching a decision. Calculating the information gain of a specific attribute is determining by the following formula:

$$Gain(A)=Info(D)-Info\ A(D)$$

$$Info(D) = -\sum_{i=1}^{m} Pi\ \log_2(Pi) \qquad (1)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (2)$$

where:
D: Dataset.
Gain (A): information gain for the attribute A.
Info (D): expected information needed to classify a tuple in D (entropy of D) [19].

3.2.2 k-Nearest Neighbor (k-NN) [18]

are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[8][14] Random decision forests correct for decision trees' habit of overfitting to their training setIt is one of the simplest algorithms that performs similarity functions, storing all cases with a known label and classifies new data based on the similarity measures or distance function. k-NN stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique. classify new data by using k value to find the nearest case in the dataset, for example if (k =1) then simply assign the new case to the class of its first nearest neighbor, if the (k = 3) then k-NN calculate the distance of the nearest three cases and apply majority vote on the class of these cases to decide the class of the new data. The distance measures for finding the nearest

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

neighbor for the numerical data is calculated by one of the distance function. As follows:

$$Euclidian = \sqrt{\sum_{i=1}^{k} Xi - Yi} \quad (3)$$

$$Manhattan = \sum_{i=1}^{k} |Xi - Yi| \quad (4)$$

$$\text{Minkowski} = \sum_{i=1}^{k} \left(\left(|Xi - Yi| \right)^{q} \right)^{1/q} \quad (5)$$

3.2.3 Random Forest [20]

Random forest is a class of decision tree algorithms based on ensemble approach. classify instances using tree structure, in this tree, a node represents the test of an attribute value and a branch denotes the result of the test. Random forest creates an ensemble of classifiers as mentioned by constructing different decision trees using random feature selection and bagging approach at the training stage. The decision trees produce two types of nodes: the leaf node labelled as a class and the interior node associated with a feature. Different subset of training data is selected with a replacement to train each.

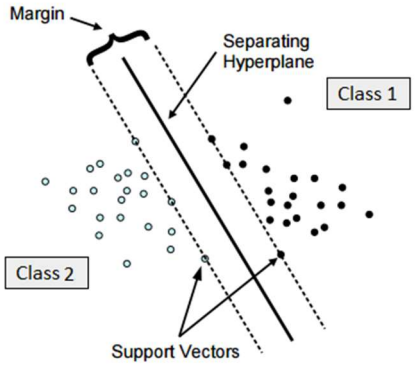

Figure 1: Support Vector Machine

3.2.4 Support Vector Machine (SVM) [19] [21]

SVM supervised machine learning algorithm which can be used for both classification or regression which defines a hyperplane that classifies the training data vectors into classes by representing n-dimensional space with the value of each feature based on a particular coordinate. Then, it performs classification by finding the hyper-plane that differentiate the two classes very well, the goal or the best choice is to find a hyperplane with the widest margin to separate the data classes. The support vectors are the data points which are closest to the hyperplane. SVM process numerical attributes only to calculate the distance of between the given object and the hyperplane that separate the class label. As illustrated in Figure 1.

3.2.5 Naïve Bayes [19] [22]

Naïve Bayes or simple Bayesian classifier is a supervised classification technique; Naïve Bayes is probabilistic algorithm depends on applying a conditional probability or based on Bayes' theorem with the independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. Bayesian theorem with naïve assumption that the occurrence of one of the attributes\predictors are independent of the occurrence of other attribute and regardless of any correlation between these attributes in the classification process. Bayes rules adopted in this algorithm stated a conditional probability of certain event based on previous knowledge about that event as follows:

$$P(C|X) = \frac{P(X|C)\,P(C)}{P(X)} \quad (6)$$

$$P(C|X) = P(X_1|C) \times P(X_2|C) \times \ldots \times P(X_n) \times P(C) \quad (7)$$

where:
P(c|x) is the posterior probability of class (target) given predictor (attribute).
P(c) is the prior probability of class.
P(x|c) is the likelihood which is the probability of predictor given class.
P(x) is the prior probability of predictor [19].

4. METHODOLOGY

The aim of this work is to conduct a comparative study to assess the performance of supervised algorithms across three mining tools, which are RapidMiner, WEKA, and Orange. To accomplish that, a dataset of 982 profiles (781 real and 201 fake) were collected and utilized by five supervised algorithms (Decision Tree, SVM, k-NN, Naïve Bayes, and Random Forest). These algorithms are implanted using the mining tools to obtain the results. It is worth to important mention here that the profiles data are extracted from 14 attributes that are listed in Table 1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 1: THE USED ATTRIBUTES

Attribute	Description	Justification
* No. of Wall Posts	Social online activities shared on the user's wall	Real users have more online activities than fake users.
* No. of Wall Tags	Linking the user with someone else on his/ her wall.	Real Users tagged more often than fake users.
Profile Picture	Visual identification of the user	Real users use their real pictures more often than fake users
Work place	Workplace or job title's information	Real users more often use their real workplace information than fake users
Education	Attended (school, college, university, etc.) information	Real users mentioned their education information in their Facebook profiles more often than fake users
Living Place	Living place address (city, town, state, etc.) information	Real users more often use their real living place information than fake users
Check In	Information for announcing user location	Real users check into places in their Facebook's profiles more often than fake users
Introduction "Bio."	Introduction information about Facebook's users	Real users are more often write something about themselves than fake users
No. of Mutual Friends	Number of the people who are Facebook friends with both users and the target profiles	Real users have more mutual friends with target profile than fake users, hence gives profile more incredibility
No. of Pages Liked	Number of pages liked	Real users usually liked more pages than fake users
No. of Groups Joined	Number of groups joined by the target profile.	Real users usually join groups more than fake users.
Family Members\ Relationship	Social relation status (married, single, etc.) information	Real users share their real social relation status than fake users.
Life Events	Information for the users to tell their stories	Real users share their life events more often than fake users.

5. EXPERIMENTS RESULTS AND DISCUSSION

In this section, the conducted experiments we will discussed, along with the employed algorithms, methods, and the evaluation metrics.

5.1 Dataset Description

A total of 906 profiles have been collected. Out of them, 104 profiles are excluded as they are (irrelevant, duplicated, mutual friends, deactivated, or deactivated). 19 profiles founded to be fake and considered as fake in the dataset. For the fake profiles, we purchased a total of 250 profiles online, filtering them to 182 profiles as some of the profiles were deactivated, irrelevant, blocked or banned from Facebook after a while. The collecting process ended with 982 profiles (781 real, and 201 fake). Manual labelling applied to the collected dataset to addresses them as fake or real.

5.2 Performance Metrics

A group of common measurements used as a comparison and validation metrics as follow [23]:

- Accuracy: Measures the performance of the classification model and calculated as:

Accuracy= (correct predictions)/(total examples)

- Recall: true positive rate and calculated as:

Recall= (true positive predictions)/ (positive examples)

- Precision: Measure the probability that the positive predication is correct and calculated as:

Precision= (true positive predictions)/ (predicted positives)

- Specificity: true negative rates and calculated as:

Specificity= (true negative predictions)/ (negative examples)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5.3 The Experiments Environment

The employed algorithms (SVM, Naïve Bayes, Decision Tree, Random Forrest, and k-NN) are assessed according to three experiments, where in the first experiment the five classifiers implemented using RapidMiner. While in the second one the WEKA tool is used. And for the final experiment, Orange mining tool is applied. As a baseline in each experiment, a cross-validation method with 10 folds is employed, the experiments are conducted on dataset of size 982 profiles (781 real and 201 fake) with a set of 14 informative attributes.

The implementation process for the selected algorithms on each mining tool follows different configuration setting as follow:

5.3.1 RapidMiner

- Decision Tree, Random Forest, and Naïve Bayes: these algorithms implemented on default setting (straight forward manner).
- k-NN: implemented with k=1, and the other setting are applied as default.
- SVM: as some of the collected dataset had some missing values. Two operators have been applied: "Nominal to Numerical", which is an operator to map the non-numeric to a numeric data type [20]; and "Impute missing values", which is an operator used estimates values for the missing attributes [21].

5.3.2 WEKA and Orange

All algorithms in Weka and Orange mining tools are applied on default settings including k-NN applied with k=1.

5.4 The Results

5.4.1 RapidMiner

A free educational version of RapidMiner studio (Version 8.2) is used. The performance's metrics and the confusion matrixes of the applied classifiers are given in Tables (2-6).

TABLE 2: DECISION TREE IN RAPIDMINER

Actual States		Real	Fake
Predicted Sates	Real	775	5
	Fake	6	196
Accuracy	**Precision**	**Recall**	**Specificity**
0.9888	0.9936	0.9923	0.9751

TABLE 3: SVM IN RAPIDMINER

Actual States		Real	Fake
Predicted Sates	Real	756	17
	Fake	25	184
Accuracy	**Precision**	**Recall**	**Specificity**
0.9572	0.9780	0.9680	0.9154

TABLE 4: K-NN IN RAPIDMINER

Actual States		Real	Fake
Predicted Sates	Real	742	14
	Fake	39	187
Accuracy	**Precision**	**Recall**	**Specificity**
0.9460	0.9815	0.9501	0.9303

TABLE 5: NAÏVE BAYES IN RAPIDMINER

Actual States		Real	Fake
Predicted Sates	Real	761	9
	Fake	20	192
Accuracy	**Precision**	**Recall**	**Specificity**
0.9705	0.9883	0.9744	0.9552

TABLE 6: RANDOM FOREST IN RAPIDMINER

Actual States		Real	Fake
Predicted Sates	Real	779	12
	Fake	2	189
Accuracy	**Precision**	**Recall**	**Specificity**
0.9857	0.9848	0.9974	0.9403

5.4.2 WEKA

In the second experiment, WEKA tool (Version 3.8.3) is used, where the selected algorithms are applied in a default configuration as mentioned, and the obtained results are showed in tables (7-11).

TABLE 7 DECISION TREE IN WEKA

Actual States		Real	Fake
Predicted Sates	Real	770	6
	Fake	11	195
Accuracy	**Precision**	**Recall**	**Specificity**
0.9827	0.9923	0.9859	0.9701

TABLE 8 SVM IN WEKA

Actual States		Real	Fake
Predicted Sates	Real	760	20
	Fake	21	181
Accuracy	**Precision**	**Recall**	**Specificity**
0.9582	0.9744	0.9731	0.9005

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 9 K-NN IN WEKA

Actual States		Real	Fake
Predicted Sates	Real	755	15
	Fake	26	186
Accuracy	**Precision**	**Recall**	**Specificity**
0.9603	0.9781	0.9718	0.9154

TABLE 10 NAÏVE BAYES IN WEKA

Actual States		Real	Fake
Predicted Sates	Real	747	4
	Fake	34	197
Accuracy	**Precision**	**Recall**	**Specificity**
0.9613	0.9947	0.9565	0.9801

TABLE 11 RANDOM FOREST IN WEKA

Actual States		Real	Fake
Predicted Sates	Real	779	5
	Fake	2	196
Accuracy	**Precision**	**Recall**	**Specificity**
0.9929	0.9936	0.9974	0. 9751

5.4.3 Orange

The final experiment conducted in Orange (Version 3.14) mining tool. The results of applied classifiers are shown in tables (12-16).

TABLE 12 DECISION TREE IN ORANGE

Actual States		Real	Fake
Predicted Sates	Real	767	21
	Fake	14	180
Accuracy	**Precision**	**Recall**	**Specificity**
0.9644	0.9734	0.9821	0.8955

TABLE 13 SVM IN ORANGE

Actual States		Real	Fake
Predicted Sates	Real	733	8
	Fake	48	193
Accuracy	**Precision**	**Recall**	**Specificity**
0.9430	0.9892	0.9385	0.9602

TABLE 14 K-NN IN ORANGE

Actual States		Real	Fake
Predicted Sates	Real	735	38
	Fake	46	163
Accuracy	**Precision**	**Recall**	**Specificity**
0.9145	0.9508	0.9411	0.8109

TABLE 15 NAÏVE BAYES IN ORANGE

Actual States		Real	Fake
Predicted Sates	Real	763	8
	Fake	18	193
Accuracy	**Precision**	**Recall**	**Specificity**
0.9735	0.9896	0.9770	0.9602

TABLE 16 RANDOM FOREST IN ORANGE

Actual States		Real	Fake
Predicted Sates	Real	772	5
	Fake	9	196
Accuracy	**Precision**	**Recall**	**Specificity**
0.9857	0.9936	0.9885	0.9751

The conducted experiments showed variant results, a discussion in in more details will be presented in the next sub-sections. Figures 2-6 show the performance charts across the three mining tools, while Figures 7-9 show the ROC curves of the algorithms in each mining tool.

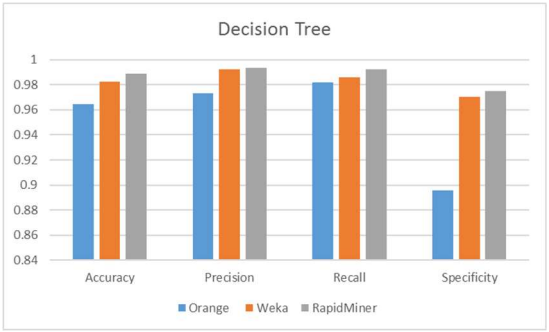

Figure 2: Decision Tree's Performance

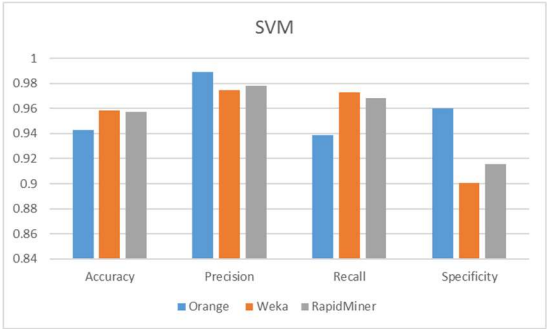

Figure 3: SVM's Performance

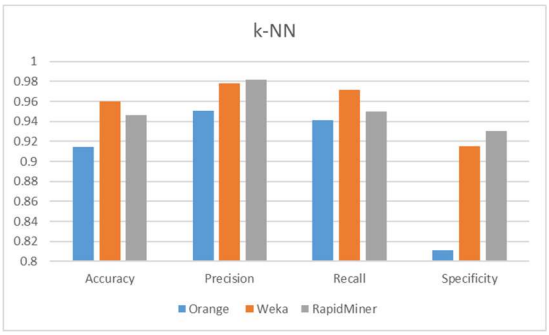

Figure 4: k-NN's Performance

Journal of Theoretical and Applied Information Technology
15th April 2019. Vol.97. No 7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

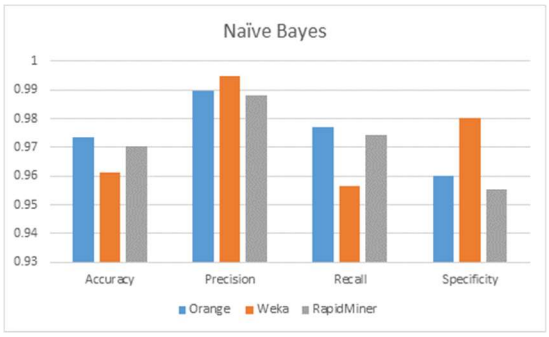

Figure 5: Naïve Bayes's Performance

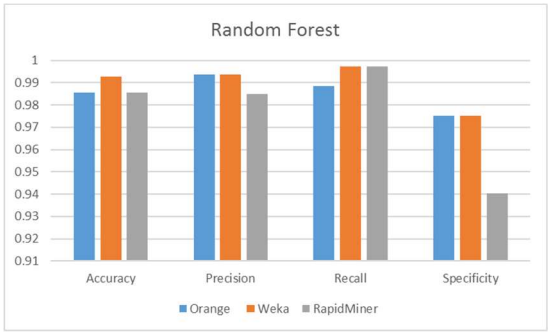

Figure 6: Random Forrest's Performance

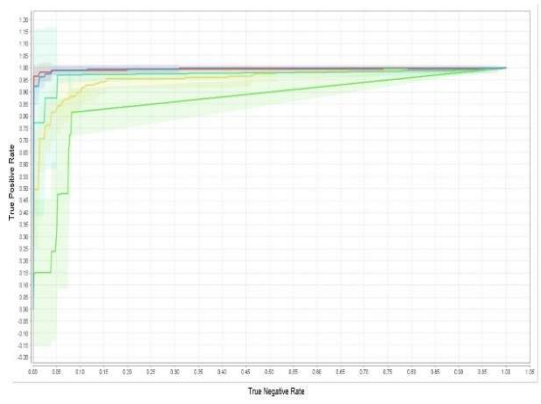

Figure 7: ROC Curve in RapidMiner

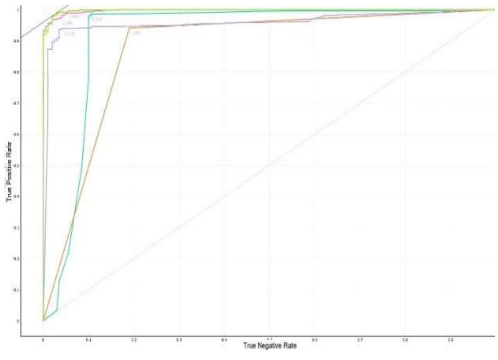

Figure 8: ROC Curve in Orange

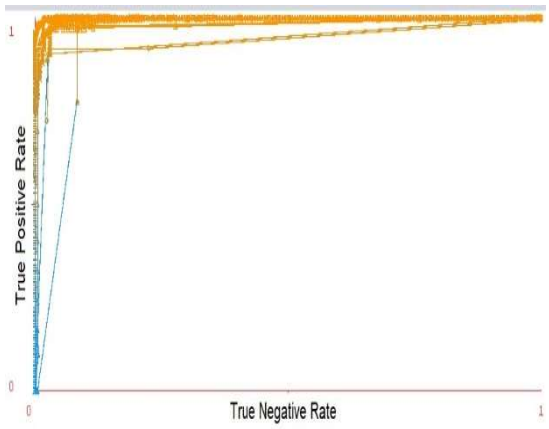

Figure 9: ROC Curve in WEKA

5.5 The Discussion

The main goal of the conducted experiments presented in this work is to give an insight about the performance of the employed algorithms across the different mining tools and attempting to support the decision-making process for choosing the suitable one. Moreover, the work aims to assess the impact of the newly added attributes on the model's detection performance. Next, we will review the obtained results and discuss them based of the following points:

- The Model's detection rates after adding the newly attributes: The detection rates for the (Decision Tree and k-NN) are improved, especially in case of the k-NN. We conducted two experiments to measure the difference (with the new attributes and without). Here, the accuracy jumped from 0.8400 to 0.9460. In contrast, SVM and Naïve Bayes are decreased with tiny proportion (less than %1) for the Naïve Bayes and about (%3) for SVM. It is noteworthy that this comparison is conducted using the RapidMiner tool. Table 17 summarizes the obtained results.

Table 17 Comparison of Accuracies

Algorithm	Accuracy with new attributes	Accuracy without
Decision	0.9888	0.9650
SVM	0.9572	0.9850
k-NN	0.9460	0.8400
Naïve Bayes	0.9705	0.9750

- The algorithms' performance across the three mining tools: The employed algorithms (SVM,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Naïve Bayes, and Random Forest) showed a stable performance in the three experiments with a slightly change in the accuracy rates (about %1 for the three algorithms), where the SVM in WEKA outperformed the one in RapidMiner and Orange. For the Naïve Bayes, the obtained results from Orange showed higher accuracy than RapidMiner and WEKA. In case of the Random Forrest, the results showed WEKA had the higher accuracy rates, and equal percentage for both RapidMiner and Orange. More details showed in Tables (3, 5, 6, 8, 10, 11, 13, 15, and 16) and Figures (5 and 6).

6. CONCLUSION

In this paper, we studied Fake profile detection on Facebook, and focusing on performance comparison of five selected classifiers (Decision Tree, SVM, Naïve Bayes, k-NN, Random Forest) implemented across three popular Data mining tools (RapidMiner, WEKA, and Orange). A set of existing features that have been proposed in the literature are adopted. In addition, new features are introduced in this paper to improve the performance of the selected classifiers.

The obtained results are promised based on four performance metrics: Accuracy, Recall, Precision, and Specificity. Specifically, the implemented classifiers on RapidMiner tool demonstrated good performance in majority of the cases. However, Random Forest in WEKA produced higher accuracy when compared with Random Forest on RapidMiner and Orange. On the other hand, k-NN showed a lowest accuracy rates in Orange followed by RapidMiner then WEKA tool. In addition, k-NN showed the most significant changes among the algorithms as it showed different accuracy rates across the three miners. The findings of this research can be useful for other researchers willing to develop machine learning models to detect malicious profiles on social online networks.

ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman-Jordan, for the full financial support granted to cover the publication fee of this research article.

REFRENCES

[1] Romero, Daniel M., Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. "Influence and passivity in social media." *In Proceedings of the 20th international conference companion on World Wide Web*, ACM, pp. 113-114,2011.

[2] https://newsroom.fb.com/company-info/ (21st September 2018).

[3] Obar, Jonathan A., and Steven S. Wildman. "Social media definition and the governance challenge: An introduction to the special issue." 2015.

[4] Kaplan, Andreas M., and Michael Haenlein. "Users of the world, unite! The challenges and opportunities of Social Media." Business Horizons 53, no. 1: 59-68, 2010.

[5] https://newsroom.fb.com/news/2018/09/security-update/ Sep. 2018.

[6] Wani, Mudasir Ahmad, Suraiya Jabin, and Nehaluddin Ahmad. "A sneak into the Devil's Colony-Fake Profiles in Online Social Networks." preprint arXiv:1705.09929 ,2017.

[7] Nazir, Atif, Saqib Raza, Chen-Nee Chuah, Burkhard Schipper, and C. A. Davis. "Ghostbusting Facebook: Detecting and Characterizing Phantom Profiles in Online Social Gaming Applications." In WOSN. 2010.

[8] Bhat, Sajid Yousuf, and Muhammad Abulaish. "Community-based features for identifying spammers in online social networks." In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on, pp. 100-107. IEEE, 2013.

[9] Mulik, S. R., and S. G. Gulawani. "Performance Comparison of Data Mining Tools in Mining Association Rules." International Journal of Research in IT, Management and Engineering (IJRIME), Volume1Issue3 ISSN: 2249-1619.Hajirnis, Aditi. "Social media networking: Parent guidance required." The Brown University Child and Adolescent Behavior Letter 31, no. 12: 1-7, 2015.

[10] Rangra, Kalpana and Bansal, KL. "Comparative study of data mining tools," International journal of advanced research in computer science and software engineering, Vol. 4, No. 6, 2014.

[11] Wani, Suheel Yousuf, Mudasir M. Kirmani, and Syed Imamul Ansarulla. "Prediction of Fake Profiles on Facebook using Supervised Machine

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Learning Techniques-A Theoretical Model." (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 7 (4), 1735-1738, 2016.

[12] Reddy, R. Nithin, and Nitesh Kumar. "Automatic detection of fake profiles in online social networks.", 2012.

[13] Fire, Michael, Dima Kagan, Aviad Elyashar, and Yuval Elovici. "Friend or foe? Fake profile identification in online social networks." Social Network Analysis and Mining 4, no. 1 (2014).

[14] Hanif, Mohamad Hazim Md, Kayode Sakariyah Adewole, Nor Badrul Anuar, and Amirrudin Kamsin. "Performance Evaluation of Machine Learning Algorithms for Spam Profile Detection on Twitter Using WEKA and RapidMiner." Advanced Science Letters 24, no. 2, 2018: 1043-1046.

[15] Albayati, M. and Altamimi, A. "An Empirical Study for Detecting Fake Facebook Profiles Using Supervised Mining Techniques". *International Journal of Computing and Informatics,Vol. 42, No. 3, 2018.*

[16] Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. "Yale: Rapid prototyping for complex data mining tasks." In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 935-940. ACM, 2006.

[17] Witten, I.H., Frank, E.: "Data Mining: Practical machine Learning tools and techniques", 2nd addition, Morgan Kaufmann, San Francisco, 2005

[18] http://orange.biolab.si/features/. Sep. 2018.

[19] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[20] Narudin, F. A., Feizollah, A., Anuar, N. B., & Gani, A. (2014). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, 1-15.

[21] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing, 14*(3), 199-222.

[22] Cook, Diane J., and Lawrence B. Holder, eds. Mining graph data. John Wiley & Sons, 2006.

[23] Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." 2011.