Available Online at www.ijcsmc.com/journal

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 11, Issue. 5, May 2022, pg.83 – 91

TWITTER ACCOUNT PREDICTION USING MACHINE LEARNING

Dr. T. A. Albinaa; Sushnitha. SE
[1]Assistant Professor, Department of Data Analytics (PG)
[1,2]PSGR Krishnammal College for Women, Coimbatore
[1] albinaathilahar894@gmail.com; [2] sushnithapk@gmail.com

DOI: https://doi.org/10.47760/ijcsmc.2022.v11i05.008

Abstract— Today's society is facing numerous cyber security related problems. In advancement of machine learning, C4.5 decision tree algorithm is applied in existing system to predict fake and clone profile in social media. It addresses cyber security needs, but it predicts clone profiles only using blank profile information. So, it achieves only 50-60 percent of accuracy, since most of the fake profiles cannot be detected using this system. Hence system has to be proposed for predicting clone profiles efficiently. This task proposes Random Forest algorithm along with Decision tree algorithm, which efficiently finds fake profile using rule based, attribute-based similarity feature, network-based feature. The system is aimed to achieve more than 80% of accuracy. Cloned profiles can also be detected using their comments and activities performed in social media. This is predicted using RF, and NN algorithm.
Keywords- Random Forest, Neural Network, MIB, Twitter Phishing.

I. INTRODUCTION

Online Social Network (OSN) [2] is a network hub where people with similar interests or real-world relationships interact. As the popularity of OSN is increasing, the security and privacy issues related to it are also rising. Fake and Clone profiles are creating dangerous security problems to social network users. Cloning of user profiles is one serious threat, where already existing user's details are stolen to create duplicate profiles and then it is misused for damaging the identity of original profile owner. They can even launch threats like phishing, stalking, spamming etc.

Fake profile is the creation of profile in the name of a person or a company which does not really exist in social media, to carry out malicious activities. In this paper, a detection method has been proposed which can detect Fake and Clone profiles in Twitter. Fake profiles are detected based on set of rules that can effectively classify fake and genuine profiles. For Profile Cloning detection two methods are used. One using Similarity Measures and the other using RF, and NN.

In Similarity Measures, two types of similarities are considered – Similarity of Attributes and Similarity of Network relationships. Algorithm detects clones by building decision tree by taking information gain into consideration. A comparison is made to check how well these two methods help in detecting clone profiles.

II. ROLE OF MACHINE LEARNING IN PREDICTION OF FAKE AND CLONE PROFILE

Many detection techniques used to classify social accounts by analysing some existing features. Some other detection techniques include ML algorithms for better classifying of accounts.

Today, Fake and Clone profiles have become a very serious threat in social networks. So, a detection method is very much necessary to find these frauds who use people's faith to gather private information and create duplicate profiles. Many authors have worked in this area and have proposed methods to identify these types of profiles in social networks. As information like phone number, email id, school or college name, company name, location etc are readily exposed in social networks, hackers can easily hack this information to create fake or clone profiles. They then try to cause various attacks like phishing, spamming, cyberbullying etc. They even try to defame the legitimate owner or the organisation. So, a detection method has been proposed which can detect both fake and clone profiles in order to make the social life of the users more secure. Unsupervised Learning and Supervised Learning are two types of machine learning methods. Input data is estimated or mapped with desired output by using the training data labelled set in supervised learning. Input data of supervised learning is called as training data and at a time it has result or known label as spam/not-spam [12].

The process considered by using some classification algorithms, and also considering the numerical features types. Moreover, other categorical features had been converted into numerical features.

As the dataset contains many attributes, we tried to test the most significant ones. Attributes that are not significant were not included in our model. This is important to apply different ML algorithms on the dataset.

III. TWITTER FAKE ACCOUNT PHISHING USING RF

Account prediction model designing for twitter uses the machine learning concept. Training and testing are two main stages in Machine learning framework.

Fig. 1 shows the block diagram of proposed detection model for Twitter account prediction using RF.

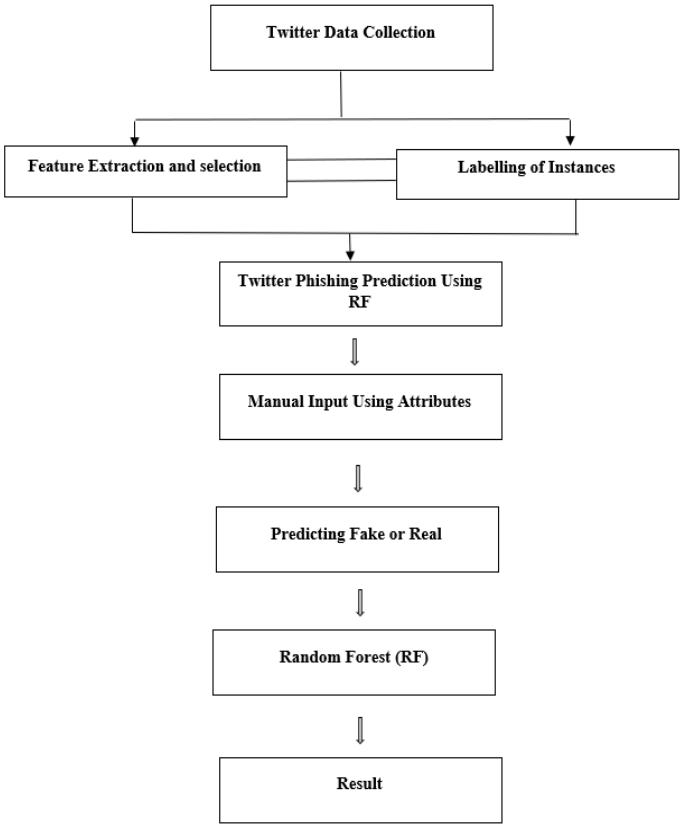


Fig. 1: Framework of Twitter Account Prediction Using RF

A. Twitter Data Collection

The dataset used in our research is the "MIB" dataset which contains 2818 accounts divided as Real Accounts and Fake Accounts.

Table 1: All Dataset Vectors of the MIB dataset.

1	Profile-link-color	18	Screen name
2	Profile-backgroundcolor	19	Protected
3	Profile-sidebar-fillcolor	20	Verified
4	Profile-backgroundtile	21	Description
5	profile_banner_url	22	Updated
6	Profile-text-color	23	Dataset
7	utc_offset	24	created_at
8	Default-profileimage	25	url
9	Default-profile	26	Lang
10	Geo-enabled	27	time_zone
11	Listed-count	28	Location
12	Favourites-count	29	profile_image_url
13	Friends-count	30	Name
14	Followers-count	31	ID
15	Statuses-count	32	profile_image_url _https
16	profile_background_ image_url_https	33	profile_backgrou nd_ image_url
17	Profile-sidebarborder-color	34	Profile-usebackgroundimage

B. Feature Selection

MIB (Management Information Base) dataset has two feature of vectors types:

- Categorical features: such as name, screenname, tweets.
- Numerical features: such as status-count, friends count, profile_text_color.

The resulted features after processing the dataset are 15 features. Those features are illustrated in Table

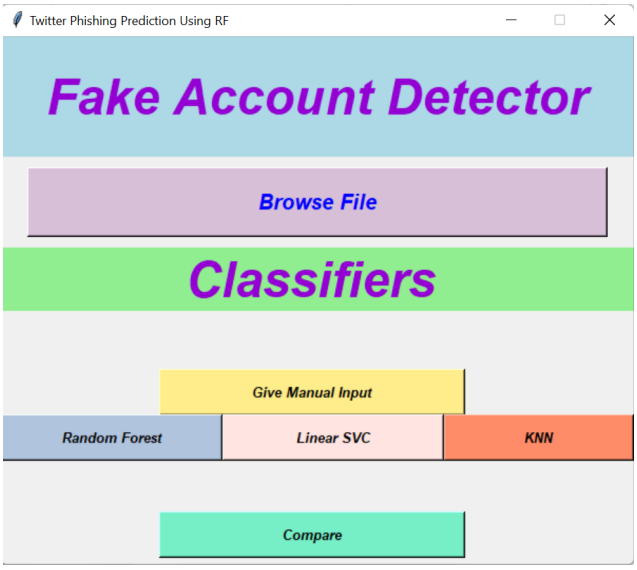
Table 2: Selected Vectors of the Dataset.

1	Profile-link-color	9	Default-profile
2	Profilebackground-color	10	Geo-enabled
3	ID	11	Listed-count
4	Profilebackground-tile	12	Favourites-count

5	Profile-sidebarborder-color	13	Friends-count
6	Profile-text-color	14	Followers-count
7	Profile-usebackground-image	15	Statuses-count
8	Default-profileimage		

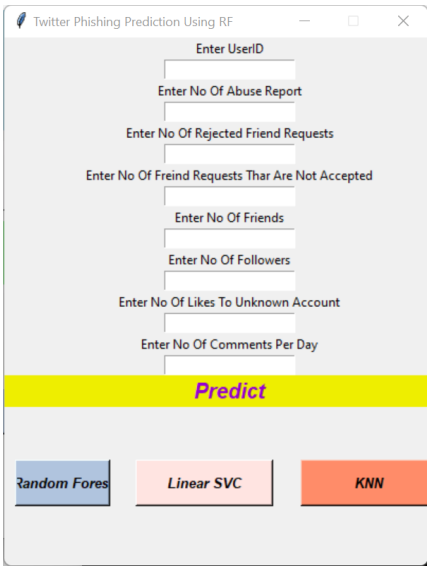
C. Twitter Phishing Detection Using RF

Using Browse File option, we input training data set into the model.



D. Manual Input Using Attributes

To test the model, we use text fields available in the application to input 8 features of the twitter account which will be used by the model to classify if the given account is Real or Fake.



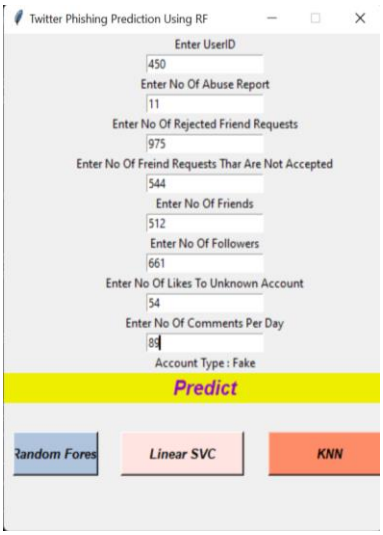
E. Predicting Fake or Real

Classification Result

After test data is inputted in the text fields, clicking any of the three buttons at bottom of screen will use the respective algorithms to provide classification result of the given test data.

Real

Fake



F. Random Forest (RF)

Random Forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Random Forest has nearly the same hyperparameters as a decision tree or a bagging classifier.

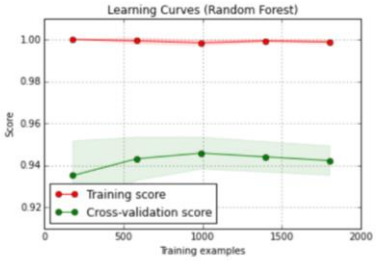
Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

IV. PERFORMANCE MEASURES

Random Forest (RF):

- **Predicting**

Learning Curves



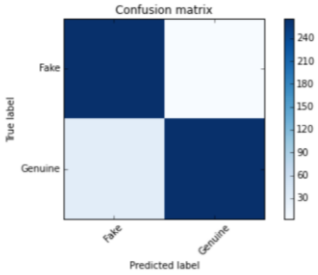
- **Visualization**

Accuracy

```
In [68]: print 'Classification Accuracy on Test dataset: ' ,accuracy_score(y_test, y_pred)

Classification Accuracy on Test dataset:  0.941489361702
```

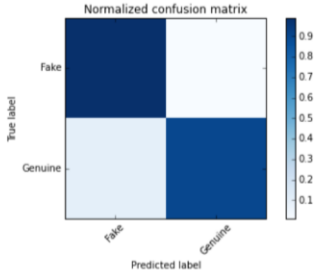
- **Confusion matrix without normalization**

```
In [70]: cm=confusion_matrix(y_test, y_pred)
print('Confusion matrix, without normalization')
print(cm)
plot_confusion_matrix(cm)

 Confusion matrix, without normalization
[[265   3]
 [ 30 266]]
```



- **Confusion matrix with normalization**

```
In [71]: cm_normalized = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
print('Normalized confusion matrix')
print(cm_normalized)
plot_confusion_matrix(cm_normalized, title='Normalized confusion matrix')

Normalized confusion matrix
[[ 0.98880597  0.01119403]
 [ 0.10135135  0.89864865]]
```



Neural Network (NN):

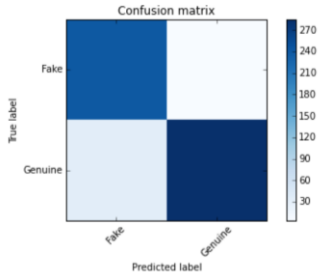
- **Visualization**

Accuracy

```
In [11]: print 'Classification Accuracy on Test dataset: ' ,accuracy_score(y_test, y_pred)
         Classification Accuracy on Test dataset:  0.934280639432

In [12]: print 'Percent Error on Test dataset: ' ,percentError(y_pred,y_test)
         Percent Error on Test dataset:  6.57193605684
```

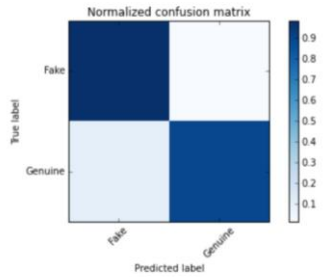
- **Confusion matrix without normalization**

```
In [13]: cm=confusion_matrix(y_test, y_pred)
         print('Confusion matrix, without normalization')
         print(cm)
         plot_confusion_matrix(cm)

         Confusion matrix, without normalization
         [[241   4]
          [ 33 285]]
```



- **Confusion matrix with normalization**

```
In [14]: cm_normalized = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
         print('Normalized confusion matrix')
         print(cm_normalized)
         plot_confusion_matrix(cm_normalized, title='Normalized confusion matrix')

         Normalized confusion matrix
         [[ 0.98367347  0.01632653]
          [ 0.10377358  0.89622642]]
```



V. RESULT EVALUATION

Once the proposed RF classifier had been trained, their effectiveness was evaluated. Using confusion matrix, the performance of the classification model is being described. The most fundamental terms used with a confusion matrix for a binary classifier are:

True-positive (TP): the number of accounts correctly identified as Faked.
False-positive (FP): the number of accounts incorrectly identified as Faked.
True-negative (TN): the number of accounts correctly identified as Trusted.
False-negative (FN): the number of accounts incorrectly identified as Trusted These can be further used to find following metrics to determine the effectiveness of each model:

Precision: Precision is the ratio of true positives to the values predicted correctly. It is defined and, given as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall: Recall is the ratio of true positives to the total number of positives. It is defined and, given as follows:

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Accuracy: the correct identification of accounts from corpus are determined by using the parameter accuracy.

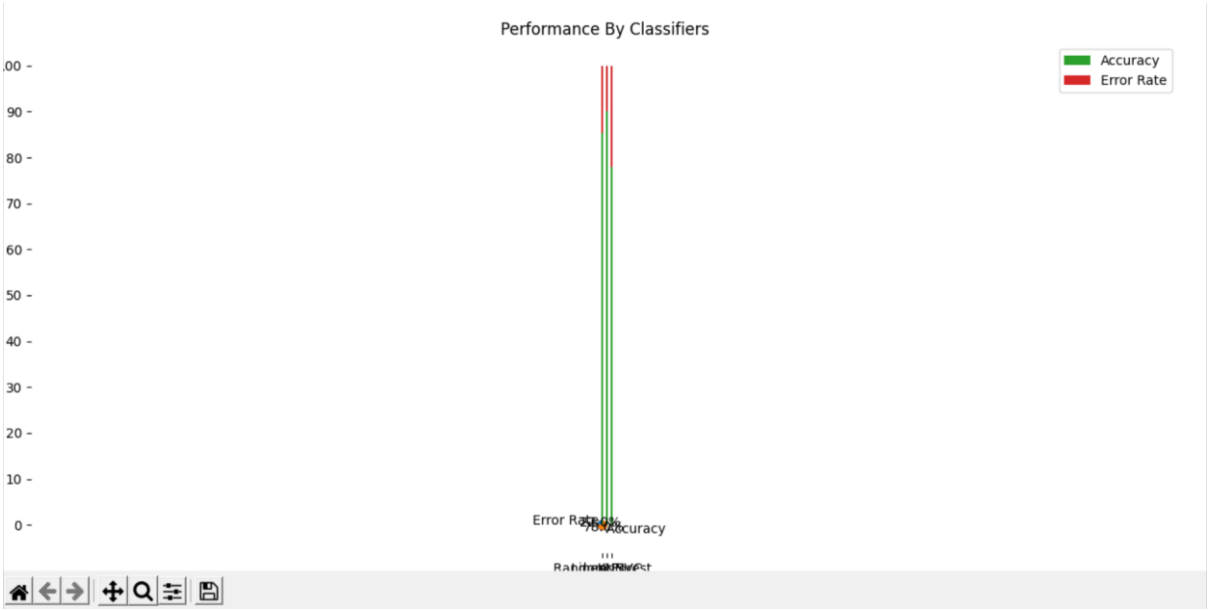
$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

The data set includes 2,818 Twitter accounts. This dataset is divided into two parts as training set with 70% of data and testing set with 30% of data. Total feature subsets in the dataset are then trained and also tested with the use of the proposed RF technique.

Table 1: A Comparative Analysis of Different Classified Techniques has been done as Follows

Parameter	Precision	Recall
RF	97.58%	84.71%
KNN	90.97%	97.15%
SVM	82.70%	92.82%

From the above table, Random Forest is been classified as the best technique with 97.58% precision.



VI. CONCLUSION

The techniques of machine learning modules usage are increasing day to day. The usage of datasets with fake and clone profiles efficiently eliminates the difficulty in finding fake profiles. Our work is based on MIB dataset. The data pre-processing and reduction phases of our model were designed to make the dataset applicable for the classification process. Different classification phases have been determined to find the best accurate method among different Machine Learning algorithms. It has been found that, RF is the best method compared to the others.

ACKNOWLEDGEMENT

My sincere thanks to all my staff of Department of Data Analytics (PG) for their timely support and encouragement. Finally, I place on record my deep sense of gratitude to my beloved parents and to my friends for their timely support in completing this project work.

REFERENCES

[1]. Alsaleh, M., Alarif, A., Al-Salman, A., AlFayez, M., Almuhaysin, A, TSD: Detecting Sybil Accounts in Twitter. 13th International Conference on Machine Learning and Applications, pp.462-469, 2014.

[2]. Aridas, C., Karlos, S. Kanas, V. Fazakis, N. Kotsiantis, S, Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers Under Imbalanced Data Sets. IEEE Access, 2019.

[3]. Awasthi, S., Shanmugam, R., Soumya, J., Atul, S, Review of Techniques to Prevent Fake Accounts on Social Media. International Journal of Advanced Science and Technology, 2020.

[4]. Ebtihal A. Hassan, Farid Meziane, "A Survey on Automatic Fake News Identification Techniques for Online and Socially Produced Data",*International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE),* 2019.

[5]. Fatih Cagatay Akyon, M. Esat Kalfaoglu, "Instagram Fake and Automated Account Detection", *Innovations in Intelligent Systems and Applications Conference (ASYU),* 2019.

[6]. Majed Alrubaian, Muhammad Al-Qurishi, Mohammad Mehedi Hassan, and Atif Alamri, "A Credibility Analysis System for Assessing Information on Twitter", IEEE Exchanges on Reliable and Secure Processing ,Volume: 15, Issue: 4, July-Aug. 1, 2018.

[7]. Md. Arafatur Rahman, Vitaliy Mezhuyev, Md Zakirul Alam Bhuiyan, S. M. Nazmus Sadat, Siti Aishah Binti Zakaria, Nadia Refat, "Reliable Decision Making of Accepting Friend Request on Online Social Networks", IEEE Access, Volume:6, 2018.

[8]. Muhammad Adil, Rahim Khan, M. Ahmad Nawaz Ul Ghani, "Preventive Techniques of Phishing Attacks in Networks", *3rd International Conference on Advancements in Computational Sciences (ICACS),* 2020.

[9]. Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Chaudhury, "Detection of Fake profile in Online Social Network using Machine Learning," in International Conference on Advances in Computing and Communication Engineering (ICACCE), International Conference on. IEEE, pp. 231-234, 2018.

[10]. Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning", *International Conference on Advances in Computing and Communication Engineering (ICACCE),* 2018.

[11]. Singh, N., Sharma, T., Thakral, A., Choudhury, T, Detection of Fake Profile in Online Social Networks Using Machine Learning. pp.231-234, 2018.

[12]. Yubao Zhang, Xin Ruan, Haining Wang, Hui Wang, and Su He "Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending" IEEE Exchanges on Data Crime scene investigation and Security, Volume: 12, Issue: 1, Jan. 2017.

