Detecting Fake Accounts on Social Media Networks Using Machine Learning

Mrs. J. Mala[1], E. Maria Shalini[2], B. Roshikha[3], K. Selvarani [4]
[1]Assistant Professor, Dept of IT
[2, 3, 4]Dept of Information Technology
[1, 2, 3, 4]Sri Ramakrishna Institute of Technology

Abstract- *In recent days, the majority of people utilise social networking sites on a daily basis. Many people create profiles on social networking sites each day and connect with others regardless of their location, time and others. Social networking site users not only benefit from them but also concern about the security of their personal information. We must initially identify the social network profiles of the users in order to determine who creates threats there. Based on the classification, we can tell real accounts from fake ones on social media. Several classification techniques have traditionally been used to identify fake accounts on social media, but we need to improve social media's ability to identify fraudulent profiles. Different Machine Learning algorithms try to identify fake accounts on social media platforms. The detection of fake accounts uses the categorization capabilities of the Random Forest Algorithm and Support Vector Machines.*

Keywords- Social Networking, Fake accounts, Support Vector Machine, Random Forest Algorithm

I. INTRODUCTION

Online social networks (OSNs) have grown in popularity in the current age, having an impact on people's social life and motivating them to join various social media platforms. Many activities, including promotion, communication, agenda formulation, advertisement, and news generation, have started to be done on social media platforms. It has gotten simpler to add new friends and stay in touch with them and their updates. These online social networks have been the subject of research to see how they affect people. Some fraudulent accounts are used to spread false information and further political agendas, for example. Finding a fraudulent account is important. Machine learning- based techniques employed to find bogus accounts that might lead users astray. The dataset is pre-processed using several Python libraries, and a comparison model is obtained to find an efficient solution that works with the provided dataset. Different Machine Learning algorithms are used to try to identify fake accounts on social media networks. For the purpose of identifying for accounts, Random Forest and Support Vector Machines are utilized.

1.1 LITERATURE SURVEY

Author details - P. Yana, L. Andrey.

Description - The authors of the study proposed a method for detecting fake accounts on the social network Twitter. A naïve Bayesian classifier is used to successfully compute fakes.

Author details - Alexy D. Frunzy, Aleksy A. Frolov

Description – The authors highlight key features of fake accounts and train various classifiers: random forest method, J48, support vector machine, naive Bayesian Classifier, Heffding tree, packed decision tree.

Author details - Dr.K. Sreenivasa Rao, Dr.G. Sreeram

Description - Recurrent neural networks can be utilized for the time series user data for a better detection of fake accounts and the algorithms can be applied to various social online platforms such as Instagram, LinkedIn and Twitter to detect the fake accounts.

Author details - D. M. Freeman

Description - False identities play an important role in advanced persisted threats (APT), i.e. coordinated, lasting, complex efforts at compromising targets in governmental, non-governmental, and commercial organizations.

1.2 EXISTING SYSTEM

Feature-based detection: This strategy is based on user-level actions and account information (user logs and profiles). Recent user activities are used to extract distinctive attributes (such as the frequency of friend requests and the percentage of requests that are accepted), which are then fed into a classifier that has been built using machine learning techniques offline.

Feature-Reduction: High dimensional data could be a significant challenge for many classification methods due to their high computational and memory requirements. On the other hand, a better classification model and straightforward visualization technique would result from removing noisy (i.e., irrelevant) and redundant characteristics from the dimension space.

1.3 PROPOSED SYSTEM

Proposed system is equipped with various Machine Learning tasks and the architecture followed is shown. The proposed system collects the dataset which are pre-processed by providing a framework of algorithms using which we can detect fake profiles in social media network by comparing the accuracy of two machine learning algorithms and the algorithm with very high efficiency is found for the given dataset.The machine learning algorithms which is used for the comparison of detecting fake account in social media networks is Support vector machine and Random forest Algorithm. A prediction form is created though which we can analyse the accounts based on few parameters from the dataset.

1.4 ARCHITECTURE DIAGRAM

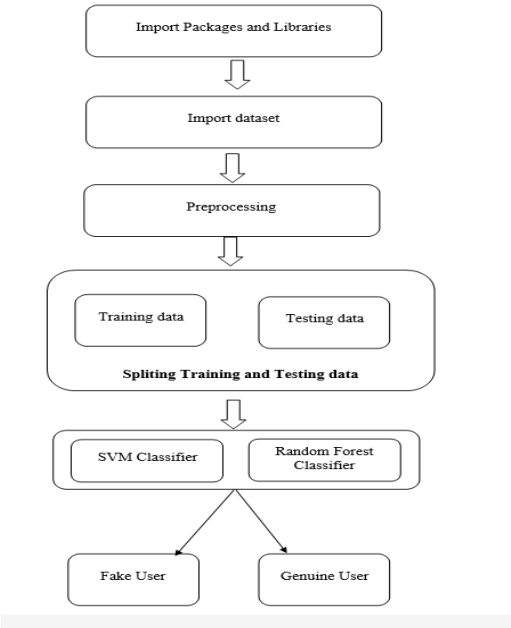

Fig 1: Architecture

1.5 DATASET

A dataset is a collection of data set corresponds to one or more database table where every column of a table represents a particular variable. Kaggle Dataset is used in this project.Kaggle is an online community platform for data scientist and machine learning enthusiasts. Users of Kaggle can work together, access and share datasets, use notebooks with GPU integration, and compete with other data scientists to solve data science problems.“Instagram fake spammer genuine accounts” is used from Kaggle Dataset. Test set is of half spammer and half non-spammer.

II. MODULES

Importing the packages

Our main tools for this project will be Pandas to work with data, NumPy to work with arrays, and scikit-learn to partition the data and develop and test classification models. Importing all of our essential packages will create a Python environment.

Splitting of data

We will define the independent variables (X) and dependent variables (Y) in this process. We will divide the data into a training set and a testing set using the given variables, which will then be utilized for modelling and evaluation. The Python algorithm "**train_test_split**" makes it simple to divide the data.

Modeling of data

Specifically, Decision Tree and Logistic Regression classification models will be constructed in this step. These are the most common models used to solve classification problems, even though there are many more that we can employ. Using the algorithms offered by the scikit-learn package, all of these models are amenable to construction.

Classification using SVM

A supervised machine learning method known as SVM bases its operation on the idea of decision planes, which specify decision boundaries. The difference between an object belonging to one class and another is determined by a decision boundary. The data points that are closest to the hyper-plane are known as support vectors. By translating input into a higher dimensional space, the kernel function is used to separate non-linear data.

Classification using Random Forest Algorithm

To categorize the fake account detection data, a Random Forest Classifier is used. An approach for classifying data called the random forest which uses several decision trees. It attempts to produce an uncorrelated forest of trees

whose forecast by committee is more accurate than that of any individual tree by using bagging and feature randomness when generating each individual tree.

2.1 SPECIFICATIONS

Hardware Requirements

RAM: 4GB (min)
Processor: I3/Intel Processor.
Hard Disk: 128 GB

Software Requirements

Operating System: Windows 7+
Server-side Script: Python 3.6+
IDE: Anaconda
Libraries Used: Pandas, Numpy, SKlearn

III. RESULTS

While comparing results, we consider many factors to ensure an accurate value in the results. The algorithms are taken into account and differences are used to calculate the better algorithm with better accuracy. The Support Vector Machine algorithm shows less accuracy than the Random Forest Algorithm in Machine Learning. The Random Forest algorithm contains large number of decision trees and it takes average to improve the prediction accuracy of the dataset.

An evaluation tool for machine learning models is a table called a confusion matrix. It gives an overview of the number of accurate and inaccurate predictions the algorithm model made. The matrix may be shown as a grid of four cells, each of which corresponds to one of four potential outcomes that is False positives, true positives, false negatives, and true negatives.

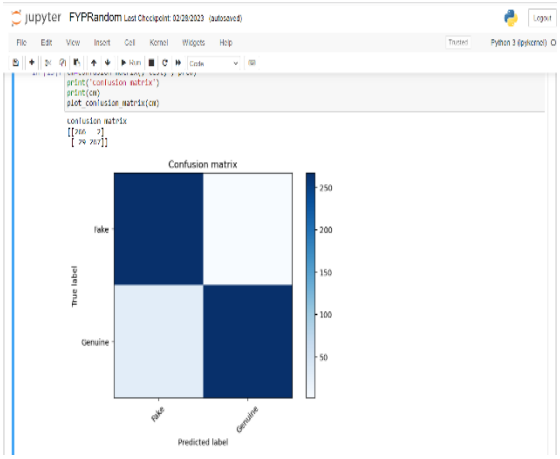

Fig 2: Confusion Matrix of Random Forest Algorithm

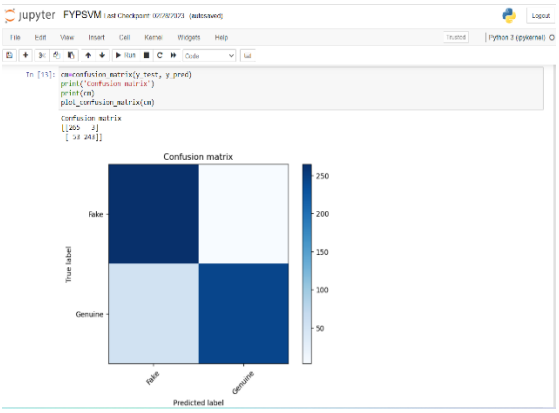

Fig 3: Confusion Matrix of Support Vector Machine

A prediction form for a machine learning project normally consists of an output field that shows the expected outcome and a collection of input fields where the user can enter the data for which they want a forecast

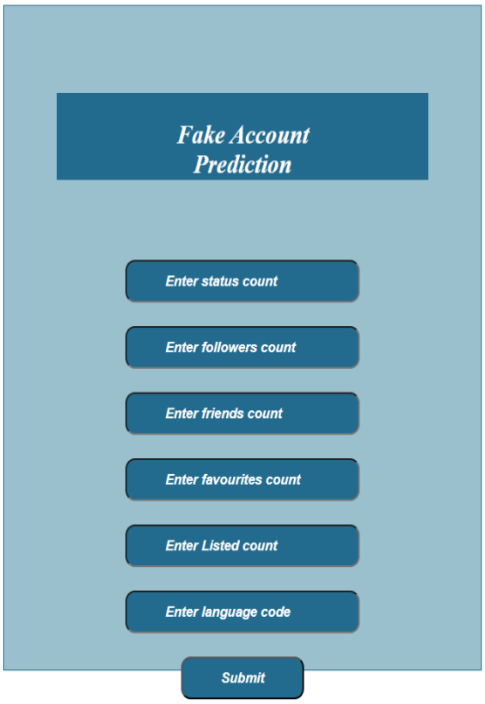

Fig 4: Prediction Form

The Random Forest Algorithm shows an accuracy of 94.50% whereas the Support Vector machine shows the accuracy of 90.07%.

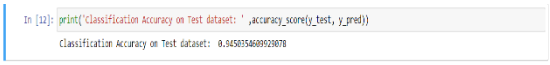

Fig 5: Accuracy of Random Forest Algorithm

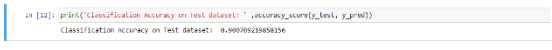

Fig 6: Accuracy of SVM Algorithm

IV. CONCLUSION

This research aims to utilise several dataset elements that have not been thoroughly studied in literature and to find a good approach of detecting automated and fake accounts by using various machine learning algorithms.In this paper, Machine Learning algorithms are used for detecting fake accounts in online social networks. Rather than making a prediction using one single algorithm, our system uses two different classification algorithms to determine whether or not an account in the provided dataset is a fake account or not. Our evaluation using Support Vector Machine and Random Forest showed strong performance, and the comparison of the accuracy of prediction seemed to be higher using Random Forest Algorithm for the given dataset. The Accuracy of detecting fake accounts is found to be higher using Random Forest Algorithm followed by SVM Algorithm for a given dataset. As a future work, recurrent neural networks can be utilized for the time series user data for a better detection of fake accounts and the algorithms can be applied to various social online platforms such as Instagram, LinkedIn and Twitter to detect the fake accounts.

REFERENCES

[1] Alexy D. Frunzy, Aleksy A. Frolov (2018) RBF, MLP, naive Bayesian Classifier, Heffding tree, packed decision tree.“IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering”

[2] Dr. K. Sreenivasa Rao, Dr. G. Sreeram, Detection of Fake profiles (2019),” International Journal of Control and Automation”, presented that Recurrent neural networks can be utilize.

[3] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy", Journal of machine learning research, vol. 5, pp. 1205-1224, Oct 2021.

[4] N. B. Karayiannis, "Reformulated radial basis neural networks trained by gradient descent", IEEE transactions on neural networks, vol. 10, no. 3, pp. 657-671, 2020

[5] S. Fong, Y. Zhuang and J. He, "Not every friend on a social network can be trusted: Classifying imposters using decision trees", Future Generation Communication Technology (FGCT) 2012 International Conference on, pp. 58-63, 2018.

[6] Y.Boshmaf, I. Muslukhov, K. Beznosov and M. Ripeanu, "The socialbot network: when bots socialize for fame and money", Proceedings of the 27th annual computer security applications conference, pp. 93-102, 2019.

[7] Yasyn Elyusufi, Zakaria Elyusufi, and M’hamed Ait Kbir, “Social Network Fake Profile Detection using Machine Learning Algorithms”, Springer Nature Switzerland AG 2020, LNITI, pp 30-40, 2020.