International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Fake Profile Identification Using Machine Learning Algorithm

Prof. Vivek Solvande[1],Vaishnavi Ambolkar[2], Siddhesh Birmole[3], Divya Gawas[4] , Dnyanada Juvale[5]

[1]Professor, Dept of IT Engineering, St. John College of Engineering and Management, Palghar, India
[2,3,4,5]Student, Dept of IT Engineering, St. John College of Engineering and Management, Palghar, India

---

Abstract - *Nowadays the usage of online social media has been increasing exponentially. Every day several people are creating their profiles on the social network platforms and they are interacting with others. Due to the continuous growth of data size on platforms with large data such as social media, the detection of fraudulent accounts on these platforms is becoming more difficult. Many problems like fake profiles, online impersonation have also grown. There are no feasible solutions exist to control these problems. Analyze, who are encouraging threats in social networks we need to classify the social networks profiles of the users. Traditionally, we have different classification methods for detecting the fake profiles on the social networks. But, we need to improve the accuracy rate of the fake profile detection on social media. In this paper we are proposing Machine Learning Algorithms for classification of fake and genuine profiles.*

1. INTRODUCTION

In the present generation, the social life of everyone has become associated with the online social networks. There is a tremendous increase in technologies these days. Online social networks is playing an important role in modern society. Social networking sites engage millions of users around the world. The users' interactions with these social sites, such as Twitter and Facebook have a tremendous impact and occasionally undesirable repercussions for daily life. Adding new friends and keeping in contact with them has become easier. The social networking sites are making our social lives better but nevertheless there are a lot of issues with using these social networking sites. The issues are privacy, online bullying, potential for misuse, trolling, creation of fake account etc. We will implement machine learning algorithms to predict if an account is controlled by fake user. There are two main common types of ML methods known as Supervised Learning and Unsupervised Learning. In supervised learning a labeled set of training data is used to estimate or map the input data to the desired output. In contrast, under the unsupervised learning methods no labeled examples are provided and there is no notion of the output during the learning process. In Supervised learning, input data is called training data and has a known label or result such as spam/not-spam at a time. A model is prepared through a training process where it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data. The goal of machine learning in fake profile detection is to have trained machine learning algorithm that, given the data of a particular profile like age, gender, numbers of friends etc. This data can facilitate in predicting whether a profile is fake or genuine effectively and will result in ensuring security of data on social networking sites. The machine learning algorithms that have been proposed to be used in this model are Support Vector Machine (SVM), Decision Tree (DT) and Naïve Bayes (NB). Also provide an analysis of those account activities from prediction result.

2. LITERATURE SURVEY

In 2018,Yeh-Cheng chen and Shystunfelix Wu [1] have presented Fake Buster: A Robust fake Account detection by Activity Analysis. They proposed an innovative method to detect fake account in OSNs(Online Social Networks). It is develop for accurately detecting fake account among social network users, based on various activity collection and analysis. In this research they have use Random forest, along with C$.5 and Adaptive Boosting, with decision stump as a second classifier that created behind it to focus on the instance in the training data , in case the accuracy of the first classifier is less effective. After finish training, a cluster of features for each testing account will input into models and output a prediction with rank score indicating the likelihood of being fake account.

In 2019, Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khathak, Ikram Uddin,MohsenGuizani, and MansourZuair [2] have presented in their work Spammer detection and fake user identification on social network. A review of techniques used for detecting spammers on Twitter. Spammers can be identified based on: (i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. The proposed taxonomy of spammer detection on twitter is categorized into four main classes, namely,(i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv) fake user identification. The first category (fake content) includes various techniques, such as regression prediction model, malware alerting system, and Lfun scheme approach. In the second category (URL based spam detection), the spammer is identified in URL through different machine learning algorithms. The third

category (spam in trending topics) is identified through Naïve Bayes classifier and language model divergence. The last category (fake user identification) is based on detecting fake usersthrough hybrid techniques.

Farhan, Muhammad Ibrohim, Indra Budi [3] have presented in their work Malicious Account Detection on Twitter based on Tweet Account features using Machine Learning. In this research, build a malicious account detection that can distinguish genuine accounts from malicious accounts using only tweet features of the accounts. Also managed to build a multiclass classification for the two types of malicious accounts,fake followers and spam bots using only tweet features. Lastly, found the best combination of algorithms, features, and data transformation scenario that suits best of our problem.

In 2019, Sk.Shama, K.Siva Nandini, P.Bhavya Anjali, K. Devi Manaswi [4] have presented their work Fake Profile Identification in Online Social Networks. In this project they have used two classifiers namely Neural Networks and Support Vector Machines and have thereby compared their efficiencies. First Collect Data and pre-process the data, Generate fake accounts, Data Validation to find fake and real , Create new features, Apply neural networks, random forest, Evaluate results of accuracy, recall etc parameters. They have taken the dataset of fake and genuine profiles. Various attributes to include in the dataset are number of friends, followers, status count. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm. From the dataset used, More than 80 percent of accounts are used to train the data, 20 percent of accounts to test the data. The predictions indicate that the algorithm neural network produced 93% accuracy.

3. METHODOLOGY

3.1 Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

3.2 Random forest

This classifier classifies a collection of decision trees to a subset of randomly generated training sets. Then it augments the likes from decision sub trees to known subclasses of handling objects for tests. Random forest will generate NA missing values for attributes to increase accuracy for larger sets of data. If more number of trees, it doesn't allow to trees to fit model.

3.3 Support Vector Machine

Support Vector Machine is a binary classification algorithm that finds the maximum separation hyper plane between two classes. It is a supervised learning algorithm that gives enough training examples, divides two classes fairly well and classifies new examples .It offers a principle approach to machine learning problems because of their mathematical foundation in statistical learning theory. SVM constructs their solution as a weighted sum of SVs , which are only a subset of the training input .It is effective in cases where the number of dimensions is greater than the number of samples given.

Table -1: Algorithm Accuracy Table

Algorithm	Accuracy
Decision Tree	87%
Support Vector Machine	72%
Random Forest	85%

4. BLOCK DIAGRAM

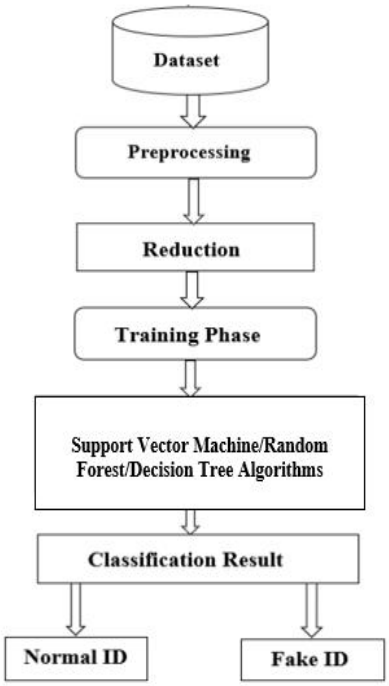

fig 1: Block Diagram

5. PROPOSED SYSTEM

In this project we have used different classification technique. Classification predicts the result based on specified input. Classification is two step process. The first step is learning in which classification algorithm analyzed the training data. The second step is a classification in which test data is used to calculate the accuracy of data. This proposed work uses the techniques like Decision tree, Random forest and Support Vector Machine. The accuracy of three algorithm is calculated for better results. Decision tree algorithm provides the better accuracy of 87%. The Decision tree algorithm provides the accurate results.

1. Collect the data
2. Pre-process the collected data
3. Reduction of the feature
4. Training the data
5. Apply the machine learning algorithm
6. Evaluate the classification result into fake and real

6. RESULT

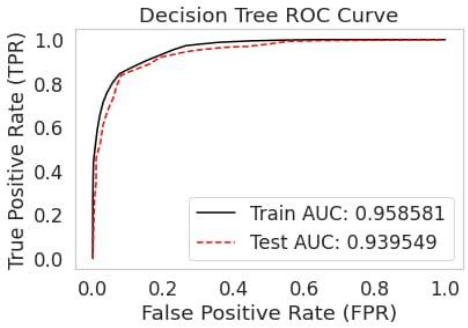

Chart -1: Decision Tree ROC curve

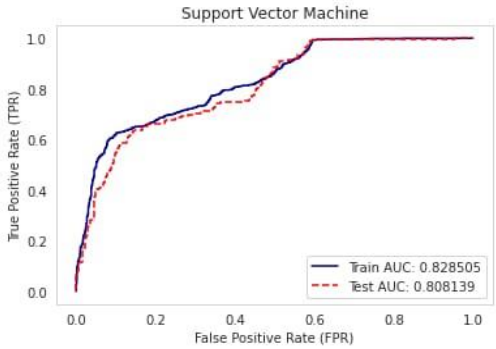

Chart -2: Support Vector Machine ROC curve

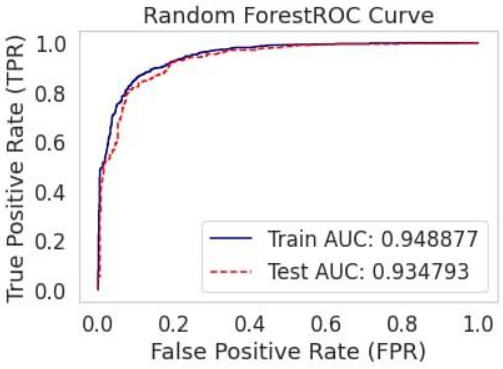

Chart -3: Random Forest ROC curve

7. CONCLUSION

Thus, we can conclude that a theoretical machine learning model has been proposed for prediction of fake profiles on online social networks. Based on the analysis in this research work it was concluded as there is no such model being used for detection of fake as well genuine profiles. Therefore, a combination of two or more machine learning algorithms can be used for detection of fake as well as genuine profiles on social networking sites.

8. FUTURE WORK

Main problem is that a person can have multiple accounts which makes them an advantage of creating fake profiles and accounts in online social networks. The idea is to attach an Aadhar card number when signing up an account so that we can restrict to creation of multiple account.

REFERENCES

[1] Yeh-Cheng chen and ShystunfelixWu,Fake Buster: A Robust fake Account detection by Activity Analysis,2018

[2] Sk.Shama, K.Siva Nandini, P.Bhavya Anjali, K. Devi Manaswi, Fake Profile.Identification in Online Social Network, 2019

[3] Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khathak,Ikram Uddin,MohsenGuizani, and Mansour Zuair, Spammer Detection and fake Profile Identification on Social Network, 2019.

[4] Malicious Account Detection on Twitter Based on Tweet Account Features using Machine Learning.