Major Project Report

on

Social Networking Profile Identification using Machine Learning

Bachelor of Engineering

in

Information Technology

by

Mr.Hemant Rathod

under the guidance of

Prof. Renuka Pawar
Department of Information Technology
Bharatiya Vidya Bhavan’s
Sardar Patel Institute oTechnology
Munshi Nagar, Andheri-West, Mumbai-400058
University of Mumbai

2019-20

Certificate

This is to certify that the Project entitled ”Social Networking Profile Identification using Machine Learning”has been completed successfully by Mr.Hemant Rathod under the guidance of Prof. Renuka Pawar for the award of Degree of Bachelor of Engineering in Information Technology from University of Mumbai.

Certified by

Prof.Renuka Pawar
Guide

Prof. Varsha Hole Project
Head of Examination Department

Department of Information Technology

Bharatiya Vidya Bhavan’s Sardar
Patel Institute of Technology
Munshi Nagar, Andheri(W), Mumbai-400058
University of Mumbai
April 2014

Project Approval Certificate

This is to certify that the Project entitled ”Social Networking Profile Identification using Machine Learning”,developed by Mr.Hemant Rathod is approved for the award of Degree of Bachelor of Engineering in Information Technology from University of Mumbai.

External Examiner	**Internal Examiner**
(signature)	**(signature)**
Name:	**Name:**
Date:	**Date:**

Seal of the Institute

Abstract

In the present generation, the social life of everyone has become associated with the online social networks. Making friends and keeping in contact with them and their updates has become easier. But with their rapid growth, many problems like fake profiles, online impersonation have also grown. There are no feasible solutions exist to control these problems.In this project, we propose a model that could be used to classify an account as fake or genuine. This model uses different classification technique and can process a large dataset of accounts at once, eliminating the need to evaluate each account manually.

Contents

Chapter 1

Introduction

Social networking sites have commonly used the channel of communication between people. Users of social networking sites can share their information and daily activities which attract a number of people towards these sites. Some people create fake accounts on social networking sites. Fake accounts do not have any real identity. Basically, the person who creates fake accounts is known as Attacker. The attacker uses incorrect information or statistics about some real world person to create a fake account. Using theses fake accounts, attacker spread false information which affects other users. To protect such sensitive data of users is one of the major challenges of social networking sites. There is a range of machine learning techniques that have been developed to detect fake accounts in social networking sites. Some of these techniques are Naive Bayes, Random Forest and Support Vector Machine. In recent researches, it has been found that these techniques make available enhanced results to detect fake accounts.

1.1 Problem Statement

- Social media is growing incredibly fast these days.
- The social networks are making our social lives better but there are a lot of issues which need to be addressed. The issues related to social networking like privacy,misuse etc are most of the times used by fake Profiles on social networking sites.
- People create fake profiles On Social Media Websites like Facebook,Instagram And Twitter.
- Social networks fake profile creation considered to cause more harm than any other form of cyber crime. This crime has to be detected even before the user is notified about the fake profile creation.

1.2 Objectives

- To study the various machine learning techniques used for classification for data.
- To study the features for detecting the fake accounts.
- To identify the required and optimal techniques for desire results.
- To implement the proposed technique to detecting the fake accounts.

1.3 Scope

- Our System Will identify the Fake and Genuine Profiles On Social Media Web sites.
- We will Use Naïve Byes, Random Forest,Support Vecor Machine Classifier Algorithm For classifying the Fake and Genuine Profiles.
- There are a lot of crimes happening these days using fake profiles. Our system helps to prevent such crimes.
- Our System Protects the information of Genuine user from fake user

Chapter 2

Literature Review

2.1 Social Networks Fake Profiles Detection Using Machine Learning Algorithms (2020)

In this paper, they have identify the fake profile in social network using limited profile data.they can identify the fake profile with 99.64% correctly classified instances and only 0.35% incorrectly classified Instances.The process of fake profile detection has three levels, in the first level profile features are Extracted and then in the second level: Random Forest (RF), Naïve Bayes (NB) and Decision Tree (DT) are used to determine the fake and genuine profiles. The third level, We calculate and compare the accuracy rates across the results of both techniques.

2.2 Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP (2018)

In this paper, they have proposed machine learning algorithms along with natural language processing techniques. In this paper they took the Facebook dataset to identify the fake profiles. The NLP pre-processing techniques are used to analyze the dataset and machine learning algorithm such as SVM and Naïve Bayes are used to classify the profiles. These learning algorithms are improved the detection accuracy rate in this paper.The presented process used Facebook profile to notice false profiles. The working method of the proposed procedure includes three principal phases:

1. NLP Pre-processing
2. Principal Component Analysis(PCA)
3.Learning Algorithms

2.3 Detecting Fake Accounts in Media Application Using Machine Learning (2018)

In this Paper, they have presented a machine learning pipeline for detecting fake accounts in online social networks. Rather than making a prediction for each individual account, our system classifies clusters of fake accounts to determine whether they have been created by the same actor.The detection process starts with the selection of the profile that needs to be tested. After selection of the profile the suitable attributes ie., features are selected on which the classification algorithm is being Implemented ,the attributes extracted is passed to the trained classifier . The classifier is being trained regularly as new training data set is feed into the classifier. The classifier determines whether the profile is fake or real.

2.4 Fake Account Detection using Machine Learning and Data Science (2019)

In this Paper, they have come up with an ingenious way to detect fake accounts on OSNs By using machine learning algorithms to its full extent, they have eliminated the need for manual prediction of a fake account. In the Paper,they used a gradient boosting algorithm. Gradient boosting algorithm is like random forest algorithm which uses decision trees as its main component.They introduced new methods to find the account. The methods used are spam commenting, engagement rate and artificial activity.Following Steps used for Detecting Fake Account

1. Web Scraper
2. Cal Engagement Range
3. Artificial Activity
4. Spam Comments
5. Detect Fake Account

2.5 Fake Account Detection in Twitter Based on Minimum Weighted Feature set (2016)

This research paper aims to propose the minimum set of attributes that is able to detect the fake users with highest accuracy.The proposed method consists of two main steps, the first step is determine the main factors that influence a correct detection of fake accounts, and the second step is to apply a classification algorithm that uses the determined factors in step one on twitter accounts for discovering the fake accounts. This research paper aims to propose the minimum set of attributes that is able to detect the fake users with highest accuracy.

2.6 Detection of Fake Accounts In Instagram Using Machine Learning (2020)

In this paper, they introduced a novel approach for detecting fake user profiles on Instagram based on certain features using concepts of machine learning. They used two models for this Logistic Regression and Random Forest algorithms, achieving an accuracy of 90.8% and 92.5% respectively. The Proposed method uses the a novel approach.Following Steps used for Detecting Fake Accounts

1. Data Collection
2. Data Pre-Processing: Missing Value Treatment,Outlier Detection

3. Calssification Algorithm:Logistic Regression,Random Forest
4. Results and Outputs:Accuraccy

2.7 Machine Learning Framework for Detecting Spammer and Fake Users on Twitter (2019)

In this paper, they used Different Machine Learning algorithms like Random forest, Minimum weight and K-means. In this Paper, they detect the Fake User as well as Spammers on twitter. In this paper they are going to divide the fake users into four types are

1. Fake content,
2. URL based spam detection
3. Detecting spam in trending topics
4. Fake user identify.

2.8 Improved Model for Detecting Fake Profiles in Online Social Network: A Case Study of Twitter (2020)

In this paper,research was tailored to finding a more efficient way of detecting fake accounts in OSN. This study was carried out using datasets got from Twitter as a case SSstudy which spanned about 37 countries and contains over one hundred thousand records. PCA Algorithm was then applied on these well formatted and cleaned data for feature selection.In this Paper there are major steps to achieve the main aim to identify fake profile in online social network:
1. Data Collection
2. Feature Extraction
3. NLP Pre-processing technique
4) Dimensional Reduction Technique Using PCA

2.9 Classification Of Instagram Fake Users Using Supervised Machine Learning Algorithms(2020)

The Methodology starts with fake users and authentic users data collection. All private users were removed, because only user's metadata can be acquired from them, not media data. Available metadata on Instagram are username, full name, biography, link, profile picture, number of posts, following, followers. After data collection, these features will be extracted, and the correlation analysis will be carried out. After setting up the features to be used, machine learning algorithms will be used to classify the users.
In this paper,five supervised machine learning algorithms are used for the classification tasks, with the 17 mentioned features. The classification will be divided into 2-classes and 4-classes classification. The outcomes of each classification are the standard performance measures .i.e. accuracy, precision, recall, F-measure, ROC curve. Random Forest consistently outperforms other algorithms. Interestingly, while other algorithms struggle in the 4-classes classification, Random Forest can perform even better than the 2-classes counterpart.

2.10 Detection Of Compromised Accounts In Online Social Network(2018)

In this paper, they first build a social behavior profile of the online social network to distinguish between different users and their behavior patterns. They will consider two types of behavior for the online social users, extroversive and introversive behavior , based on these features we will be able to distinguish spam users from the legitimate owners. In this Paper,They introduce different featured to represent a users social behaviors, including extroversive and introversive behavior. A user's feature values consists of its behavior profiles. The analysis is not only conducted on user profiles and message contents, but we try to discover the users history of his social activities. Online social networks provide various features like sending messages, chatting, uploading photos, browsing content, browsing friends, uploading status, downloading pictures etc.

2.11 Recognition Of Fake Profile In Online Social Networks Using Machine Learning(2020)

In this Paper Methodology are Sybil accounts have various attributes contrasted with ordinary clients. Consequently, Researcher investigated the probability of recognizing typical and Sybil accounts utilizing grouping calculation like SVM, and NN. In this Paper, the exploration work have been done to distinguish, recognize and dispose of phony bot accounts made and cyborgs can't be utilized for separating counterfeit record made by individuals.

2.12 Twitter Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language(2020)

Bots have developed exponentially over the past few years to the point that it has become difficult to distinguish them from real accounts. Supervised machine learning models are the most popular techniques used for the detection of bots. In this Paper explains our new proposed approach, based on user attributes and ontology technologies, attempts to identify and recognize fake accounts on Twitter. The system is composed of three stages: data preprocessing and features extraction stage, ontology construction stage, and SWRL rules and reasoner as a classifier stage. In this paper, a new approach has been proposed to detect and classify fake accounts on Twitter social networks, using ontological engineering. They modeled an ontological approach of knowledge representation across the OWL language, SWRL rules, and reasoner.

2.13 Fake (Sybil) Account Detection using Machine Learning(2018)

In this paper a machine learning approach for fake profile detection on the Facebook social network is given.They have used the majorly popular machine learning framework sci-kit learn and XGBoost library for the classification task. Many machine learning algorithms have been run over the dataset and top few are listed in the results section. At the top is AdaBoost algorithm with an accuracy of 99% for both precision and recall achieving this by optimizing the parameters n_estimators to 5000 and learning_rate to 0.01 using grid search. In this Paper, a dataset is proposed of Facebook social network for fake profile detection.They have employed many machine learning approaches in the preprocessing of the dataset. Various machine learning models are evaluated over the dataset for fake profile detection along with optimization of those models. The ensemble machine learning models outperform others and increase the accuracy of predicting fake profile on Facebook social network.

Chapter 3

Proposed Work

3.1 Overview

Each profile (or account) in a social network contain lots of information such as name, gender, number of friends, number of followers, number of likes, location etc. Some of these information are private and some are public. We have used information that are public to determine the fake profiles in social Network as private information is not accessible. However, if our proposed scheme is used by the social networking companies itself then they can use the private information of the profiles for detection without violating any privacy issues. We have considered these information as features of a profile for classification of fake and real profiles. The steps that we have followed for detection of fake profiles are as follows.

1. Features are selected to apply classification algorithms. Attributes are selected as features if they are not dependent on other attributes and they increase efficiency of the classification.
2. After selection of attributes, the dataset of profiles that are already classified as fake or genuine are needed for the training purpose of the classification algorithm. We have used a publicly available dataset of 1337 fake users and 1481 genuine users consisting of various attributes including name, status count, number of friends, followers count, favourites, languages known etc.

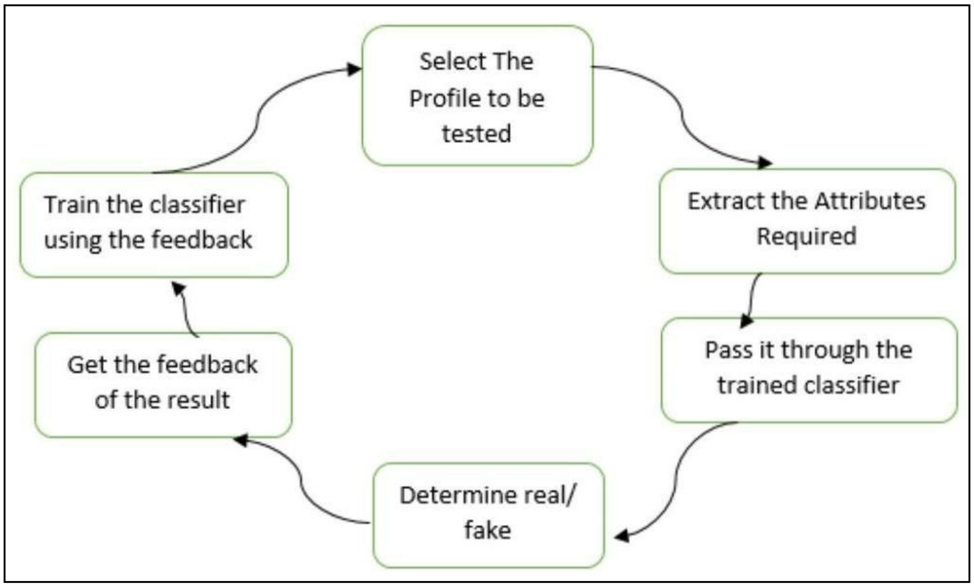
3. The selected attributes are extracted from profile for the purpose of classification.
4. After this the dataset of fake and real profiles are prepared. From this dataset, 80% of both profiles (genuine and fake) are used to prepare a training dataset and 20% of both profiles are used to prepare a testing dataset.

5. The training dataset is then fed to the classification algorithm. It learns from the training dataset and is expected to give correct class labels for the testing dataset.

6. The labels from the testing dataset are removed and are left for determination by the trained classifier. The result of classification algorithm is shown in 4.4. We have used two classification algorithms and have compared the efficiency of these algorithms.

3.2 Proposed Framework

The proposed framework in the figure 3.1 shows the sequence of processes that need to be followed for continuous detection of fake profiles with active learning from the feedback of the result given by the classification algorithm.


Select The Profile to be tested
Extract the Attributes Required
Pass it through the trained classifier
Determine real/ fake
Get the feedback of the result
Train the classifier using the feedback

Framework for detection of fake profiles

This is a framework that can easily be implemented by social networking companies as they have access to user information.

1. Classification starts from the selection of profile that needs to be classified.
2. Once the profile is selected, the useful features are extracted for the purpose of classification.
3. The extracted features are then fed to trained classifier.
4. Classifier is trained regularly as new data is fed into the classifier.
5. Classifier then determines whether the profile is genuine or fake.
6. The result of classification algorithm is then verified and feedback is fed back into the classifier.
7. As the number of training data increases the classifier becomes more and more accurate in predicting the fake profiles.

3.3 Classification

Classification is a technique of categorizing an object into a particular class based on the training data set that was used to train the classifier. We feed the classifier with data set so that we can train it to identify related objects with as best accuracy as possible. Classifier is an algorithm used for classification. In this project we have used two classifiers namely Random Forest and Support Vector Machines and have thereby compared their efficiencies.

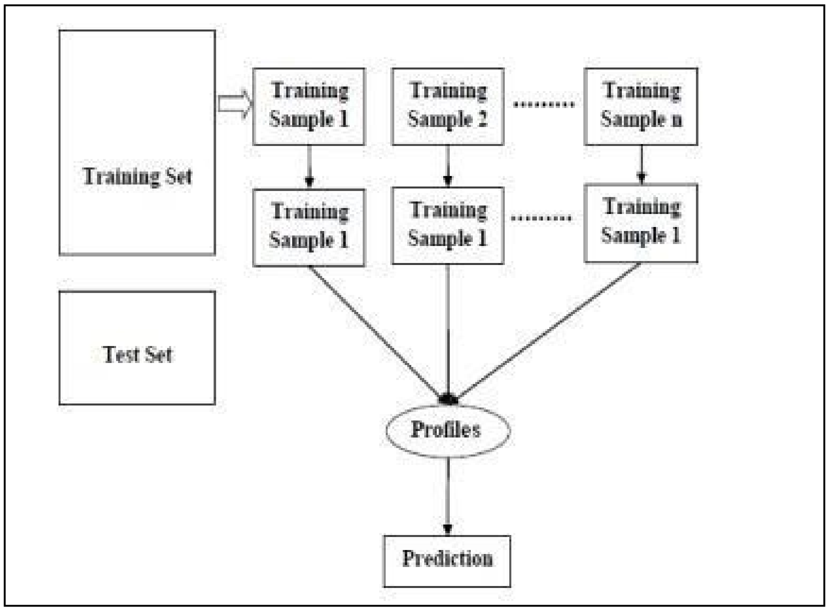
1) Random Forest

Random forest is a supervised learning algorithm that is used for both classifications as well as regression. But however, it is mainly used for classification problems.Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

We can understand the working of the Random Forest algorithm with the help of following steps:

Step 1 − First, start with the selection of random samples from a given dataset.
Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
Step 3 − In this step, voting will be performed for every predicted result.
Step 4 − At last, select the most voted prediction result as the final prediction result.

Working of Random Forest



2) Support Vector Machine

A Support Vector Machine (SVM) is a binary classifier that performs classification by finding a hyperplane that maximizes distance between two classes. It is a supervised machine learning algorithm. The algorithm outputs a hyperplane that fairly divides two classes with the help of training data and categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

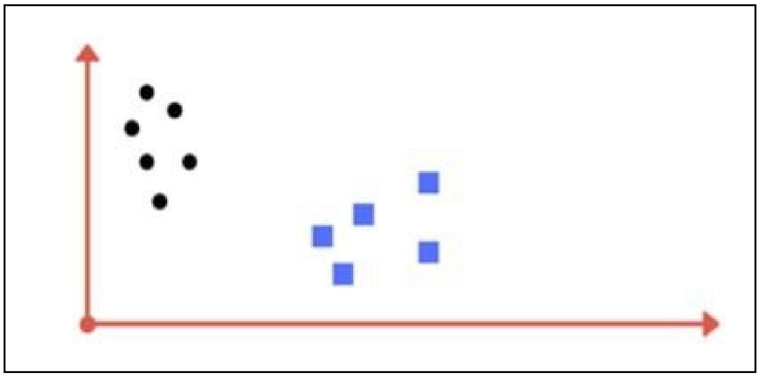


Fig3.3:Draw a line that separates black circles and blue squares

You might have come up with something similar to following image (fig. 3.3). It separates the two classes. Any point that is left of line falls into black circle class and on right falls into blue square class. Separation of classes. That’s what SVM does.

Chapter 4

Implementation and Results

4.1 Datasets

We needed dataset of fake and genuine profiles. Various attributes included in dataset are number of friends, followers, status count. Dataset is divided into training and testing data. Classification algorithms are trained using training dataset and testing dataset is used to determine efficiency of algorithm. From the dataset used, 80% of both profiles (genuine and fake) are used to prepare a training dataset and 20% of both profiles are used to prepare a testing dataset.

4.2 Attributes

1. Status Count
2. Followers count
3. Friends Count
4. Favourites Count
5. Listed Count
6. Gender
7. Language Code

4.3 Evaluation Parameters

Efficiency/Accuracy = Number of correct predictions/ total number of predictions

Percent Error = (1-Accuracy)*100

Confusion Matrix = Confusion Matrix is a technique for summarizing the performance of a classification algorithm. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

TPR- True Positive Rate
$TPR=TP/(TP+FN)$

FPR- False Positive Rate
$FPR=FP/(FP+TN)$

TNR- True Negative Rate
$TNR=TN/(FP+TN)$

FNR- False Negative Rate
$FNR=1-TPR$

Recall- How many of the true positives were recalled (found), i.e. how many of the correct hits were also found.
Recall = TP / (TP+FN)

Precision- Precision is how many of the returned hits were true positive i.e. how many of the found were correct hits. Precision = TP / (TP + FP) F1 score- F1 score is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.

ROC Curve- The Receiver Operating Characteristic is the plot of TPR versus FPR. ROC can be used to compare the performances of different classifiers.

4.4 Results

RANDOM FOREST

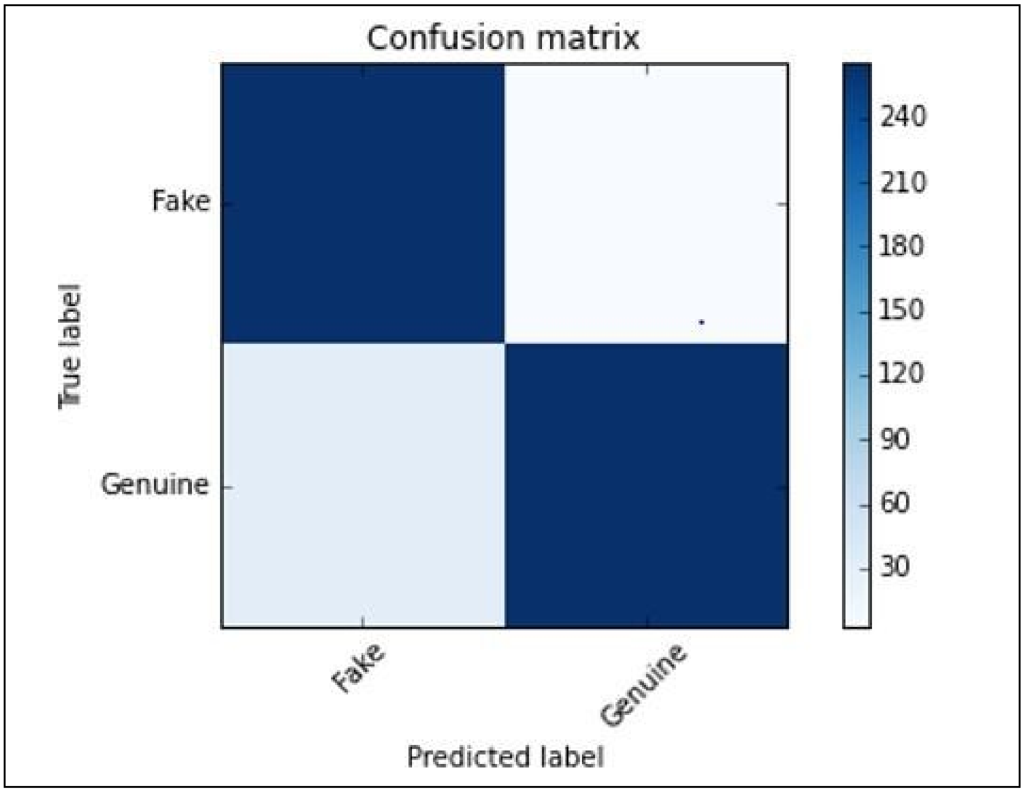


Fig 4.1- Confusion Matrix

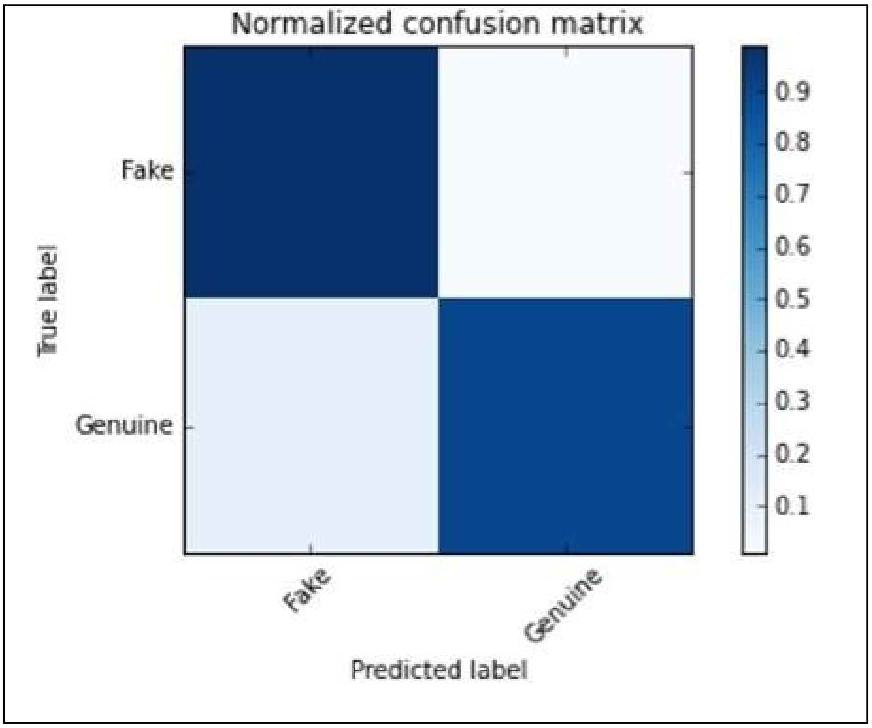
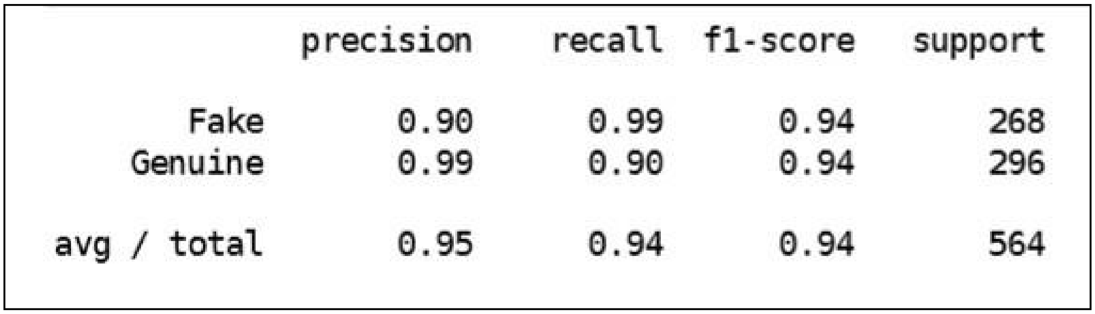


Fig 4.2- Normalized Confusion Matrix



```
             precision    recall  f1-score   support

       Fake       0.90      0.99      0.94       268
    Genuine       0.99      0.90      0.94       296

avg / total       0.95      0.94      0.94       564
```

Fig 4.3- Classification Report

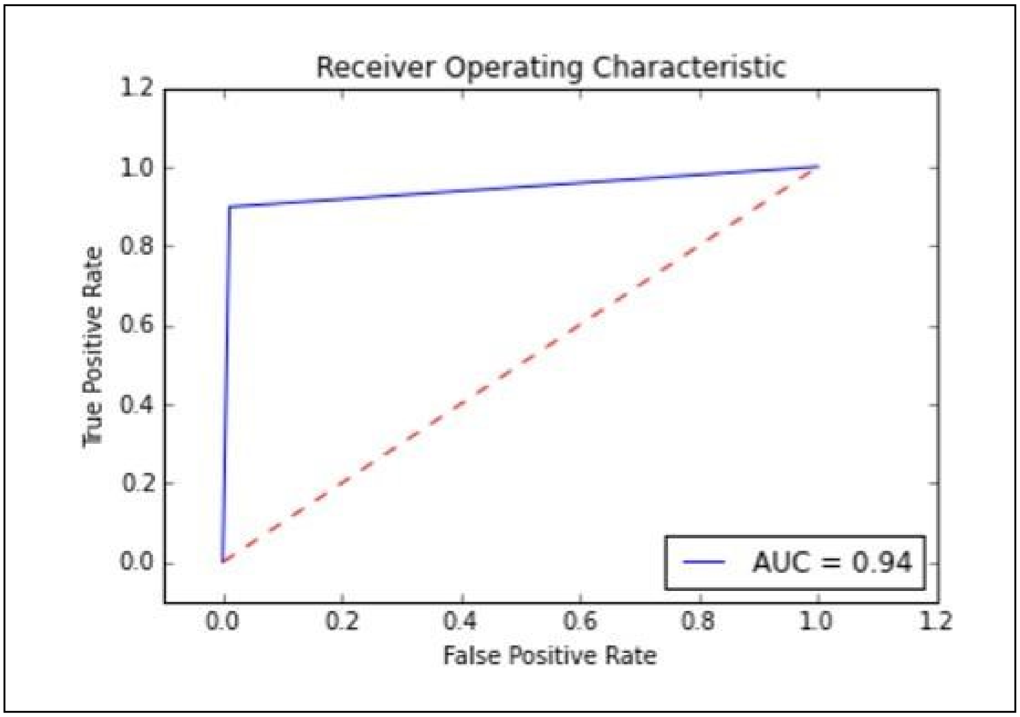


Fig 4.4- ROC Curve

Efficiency of Random Forest in classifying data is 94 %. We have taken 80% of data for the purpose of training and 20% for classification.

SUPPORT VECTOR MACHINE

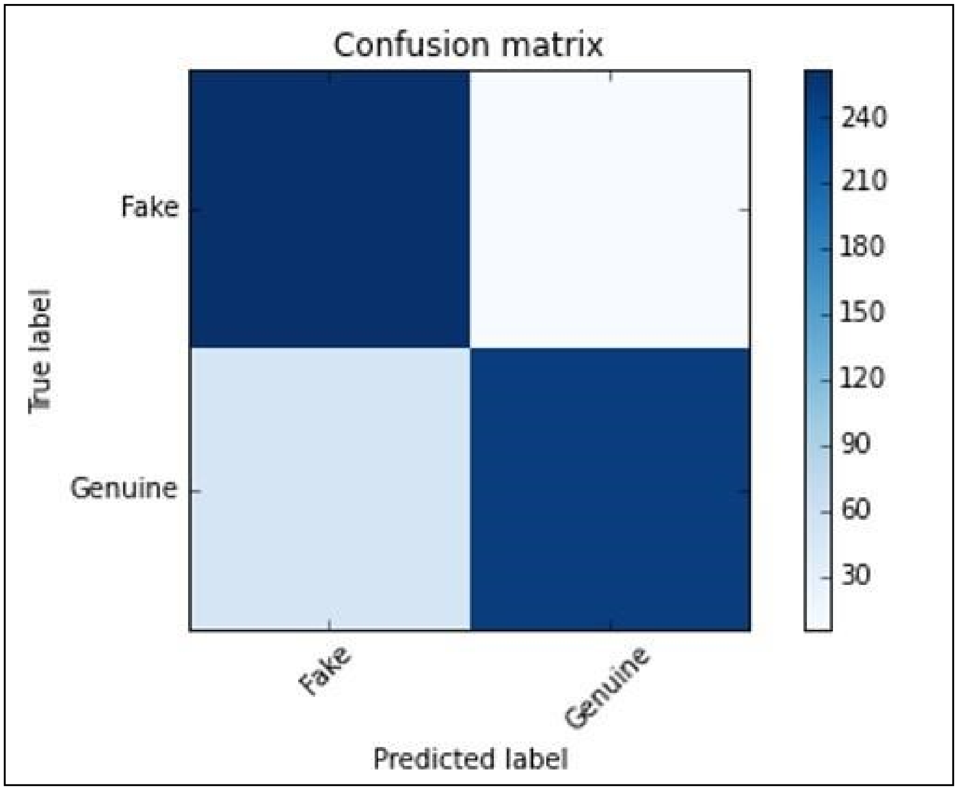


Fig 4.5- Confusion Matrix

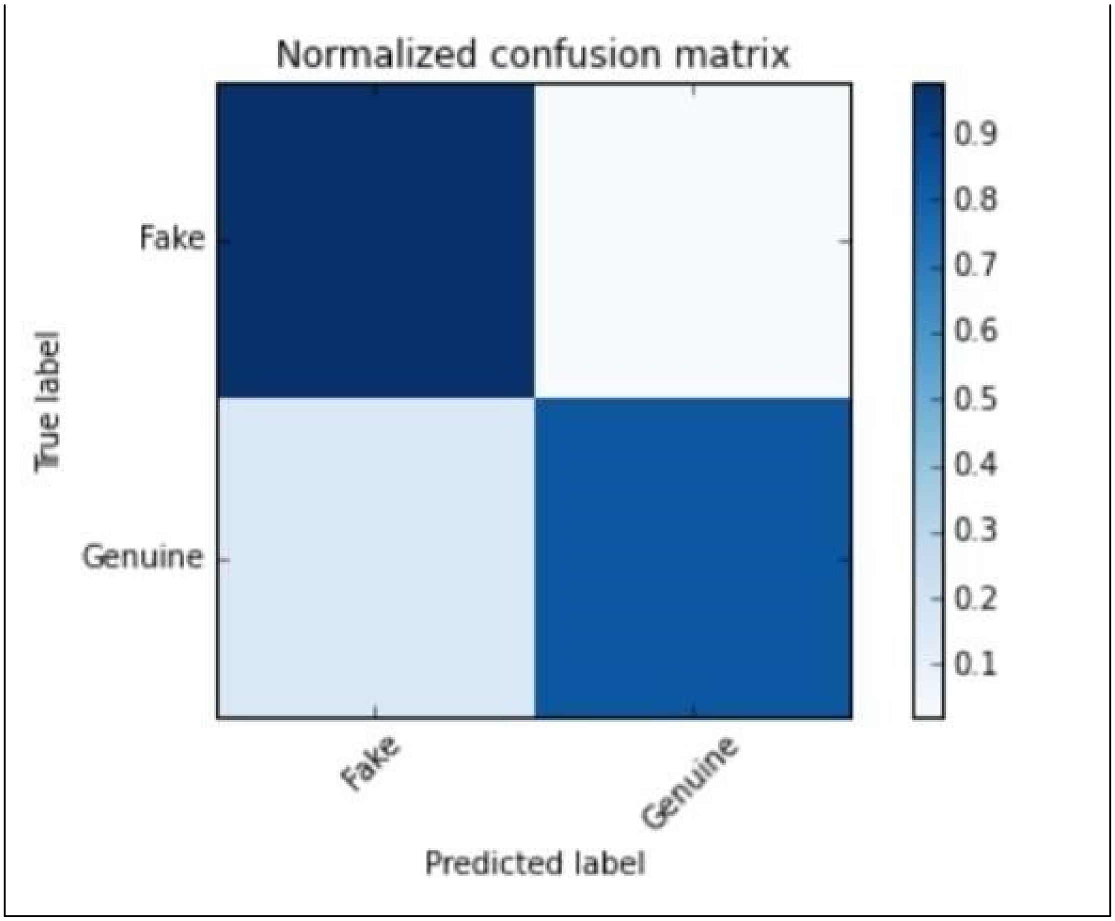
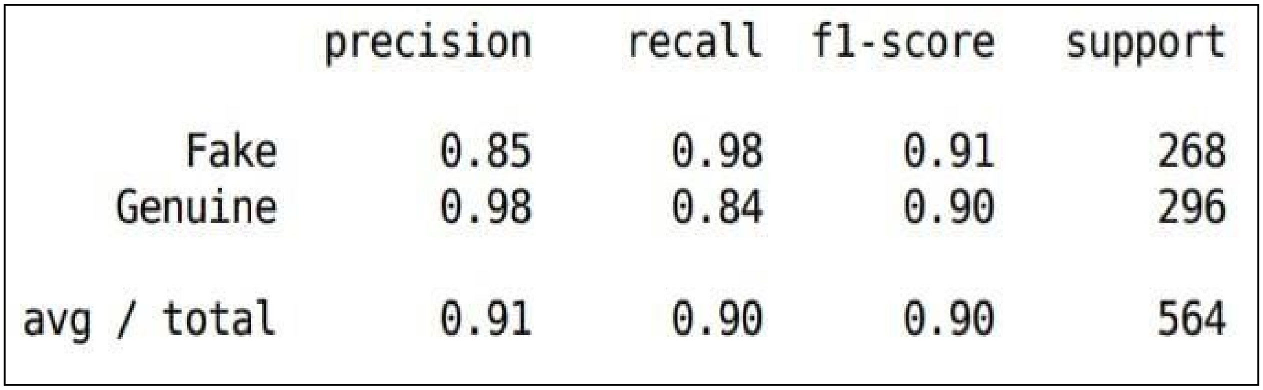


Fig 4.6- Normalized Confusion Matrix



```
             precision    recall  f1-score   support

       Fake       0.85      0.98      0.91       268
    Genuine       0.98      0.84      0.90       296

avg / total       0.91      0.90      0.90       564
```

Fig 4.7- Classification Report

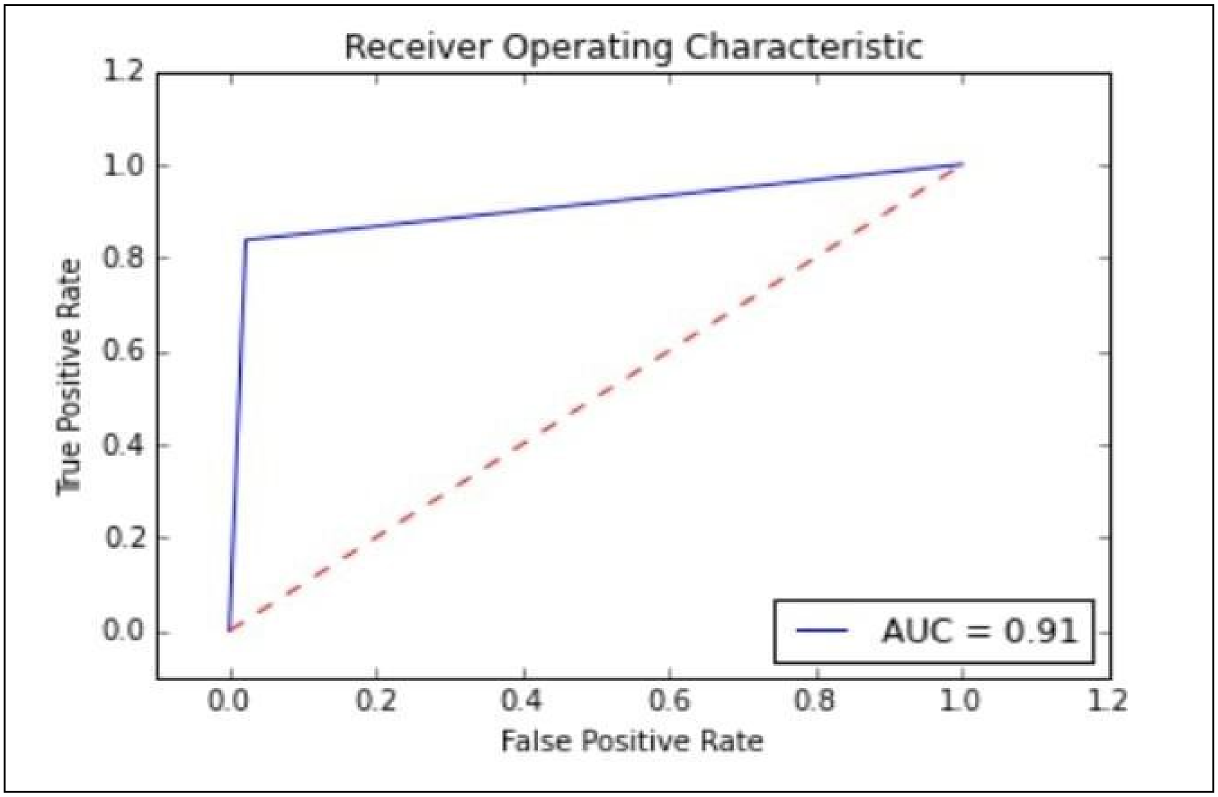


Fig 4.8- ROC Curve

Efficiency of SVM in classifying data is 91%. We have taken 80% of data for training SVM classifier and 20% for classification.

4.5 Conclusion and Future Work

We have given a framework which suggests that binary classification through Random Forest is more efficient than through Support Vector Machine. Using Random forest we have achieved efficiency of 94%. In the future we wish to classify profiles by taking profile pictures as one of the features.

Chapter 5

Technology Stack

5.1 Libraries:

- Numpy : Its a fundamental package for array processing and performs numerical operation on statistical data.

- Pandas : it is built on the numpy package and its key data structure is called as dataframe.dataframe allows to you store and manipulate tabular data in rows of observations and columns of variables.

- Matplotlib : Matplotlib is a Python programming language plotting library and its numpy Numerical Mathematics extension.

- Pip : Pip is Package Manager For Python Languages.

- Scikit-learn : Scikit-learn is Python 's free machine-learning library. It features various algorithms such as support vector machine, random forests, and k-neighbours, and also supports numerical and scientific libraries such as NumPy and SciPy in Python.

- SexMachine : For getting the Gender Attribute of the Profile

5.2 Referances:

[1] Yasyn Elyusufi (&), Zakaria Elyusufi, and M'hamed Ait Kbir: "Social Networks Fake Profiles Detection Using Machine Learning Algorithms**"**, 2020.

[2] P. Srinivas Rao, Dr. Jayadev Gyani, Dr.G.Narsimha: "Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP*", International Journal of Applied Engineering Research,* 2018.

[3] Gayathri A, Radhika S, Mrs. Jayalakshmi S.L:"Detecting Fake Accounts in Media Application Using Machine Learning***",** Int. Jnl. Of Advanced Networking & Applications (IJANA),* 2018.

[4] S. P. Maniraj, Harie Krishnan G, Surya T, Pranav R:"Fake Account Detection using Machine Learning and Data Science"**,** International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2019.

[5] Ahmed El Azab, Amira M. Idrees, Mahmoud A. Mahmoud, Hesham Hefny:"Fake Account Detection in Twitter Based on Minimum Weighted Feature set", *World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 2016.

[6] Ananya Dey1, Hamsashree Reddy2 ,Manjistha Dey3 and Niharika Sinha4 :"Detection of Fake Accounts in Instagram Using Machine Learning"**,** *International Journal of Computer Science & Information Technology (IJCSIT)* , October 2019

[7] Akshatha T M, Dr. M. N Veena :"Machine Learning Framework for Detecting Spammer and Fake Users on Twitter",International journal of engineering research & technology (IJERT), 2020.

[8] K. Ojo, A.: "Improved Model for Detecting Fake Profiles in Online Social Network: A Case Study of Twitter, *Journal of Advances in Mathematics and Computer Science*,2019.

[9] Kristo Radion Purba, David Asirvatham, Raja Kumar Murugesan.: "Classification Of Instagram Fake Users Using Supervised Machine Learning Algorithms", *International Journal of Electrical and Computer Engineering (IJECE)*,2020.

[10] Sneha Rane, Megha Ainapurkar, Ameya Wadekar.: "Detection of Compromised Accounts in Online Social Network", *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*,2018.

[11] S.Sandeep Bhat, M. Vishnu Priya.: "Recognition Of Fake Profile In Online Social Networks Using Machine Learning", *International journal of engineering research & technology (IJERT)*,2020.

[12] Mohammed Jabardi, Imohammed Jabard.: "Twitter Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language", *Karbala International Journal of Modern Science*,2020.

[13] Yeshwant Singh, Subhasish Banerjee.: "Fake (Sybil) Account Detection using Machine Learning", *National Institute of Technology*,2018.