Social Networking Profile Identification using Machine Learning

Group No: 6

Group Member:
Hemant Rathod

Project Guide:
Prof.Renuka Pawar

Outline

- Abstract
- Problem Statement
- Litrature Survey
- Objectives of the Project
- Scope of the Project
- Analysis
- Design and Methodology
- Technology Stack
- Referances

Abstract

- In the present generation, the social life of everyone has become associated with 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.

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.

Litrature survey

Referance	Methodology	Findings
Social Networks Fake Profiles Detection Using Machine Learning Algorithms (Paper 1)	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.	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..

Referance	Methodology	Findings
Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP. (Paper 2)	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	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.

Referance	Methodology	Findings
Detecting Fake Accounts in Media Application Using Machine Learning (Paper 3)	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.	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.

Referance	Methodology	Findings
Fake Account Detection using Machine Learning and Data Science (Paper 4)	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 Accounts 1)Web Scraper 2)Cal Engagement Range 3) Artificial Activity 4)Spam Comments 5)Detect Fake Account	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

Referance	Methodology	Findings
Fake Account Detection in Twitter Based on Minimum Weighted Feature set (Paper 5)	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.	This research paper aims to propose the minimum set of attributes that is able to detect the fake users with highest accuracy.

Referance	Methodology	Findings
Detection of fake accounts in instagram using machine learning (Paper 6)	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 Logistic Regression:90.8% Random Forest:92.5%	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.

Referance	Methodology	Findings
Machine Learning Framework for Detecting Spammer and Fake Users on Twitter (Paper 7)	In this paper they are going to divide the fake users into four types are (i) fake content, (ii) URL based spam detection, (iii) detecting spam in trending topics, and (iv)fake user identify. With the help of Machine learning algorithms like Random forest, Minimum weight and K-means they using these algorithms in different stages to identify the fake users and spammer on twitter.	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.

Referance	Methodology	Findings
Improved Model for Detecting Fake Profiles in Online Social Network: A Case Study of Twitter (Paper 8)	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	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.

Referance	Methodology	Findings
Classification of instagram fake users using supervised machine learning algorithms. (Paper 9)	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.

Referance	Methodology	Findings
Detection of Compromised Accounts in Online Social Network. (Paper 10)	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.

Referance	Methodology	Findings
Recognition Of Fake Profile In Online Social Networks Using Machine Learning . (Paper 11)	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.

Referance	Methodology	Findings
Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language. (Paper 12)	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.

Referance	Methodology	Findings
Fake (Sybil) Account Detection using Machine Learning . (Paper 13)	In this paper a machine learning approach for fake profile detection on the Facebook social network is given. We 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.

Objectives Of the Project

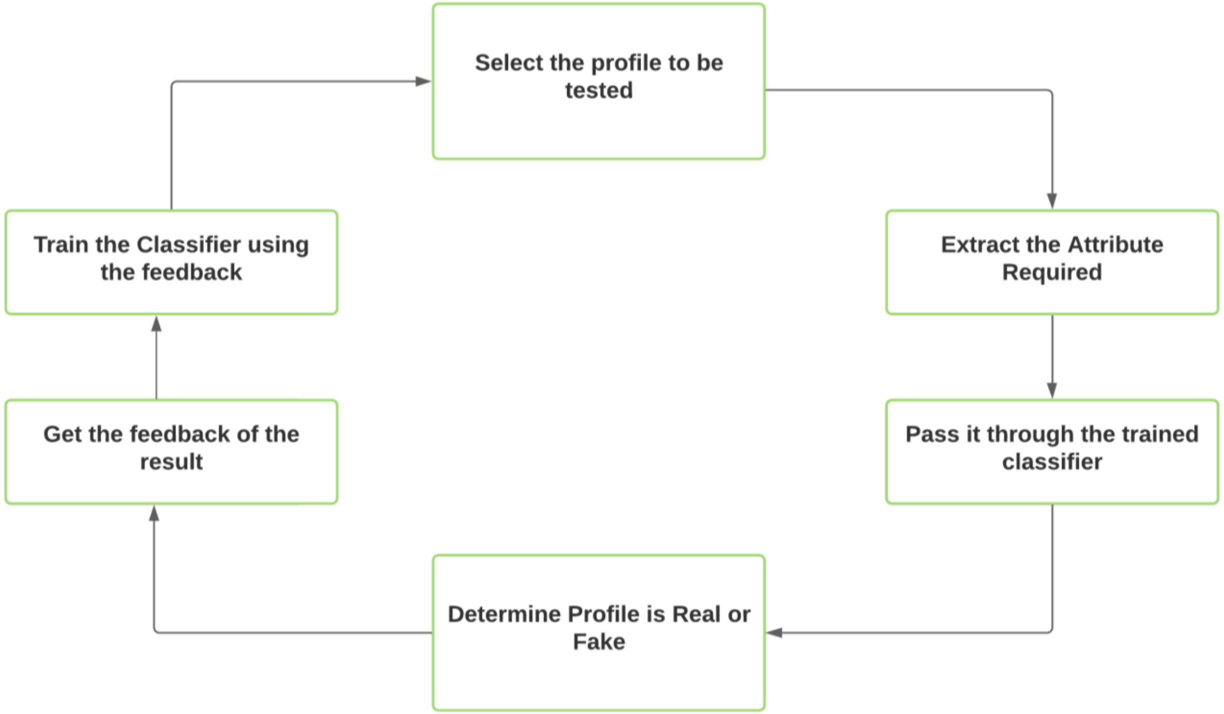
- To study the various machine learning techniques used for classification for data.
- To study the features for detecting the fake accounts.
- To identify the required and optimal techniques for desire results.
- To implement the proposed technique to detecting the fake accounts.

Scope of the Project

- Our System Will identify the Fake and Genuine Profiles On Social Media Web sites.
- We will Use 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.

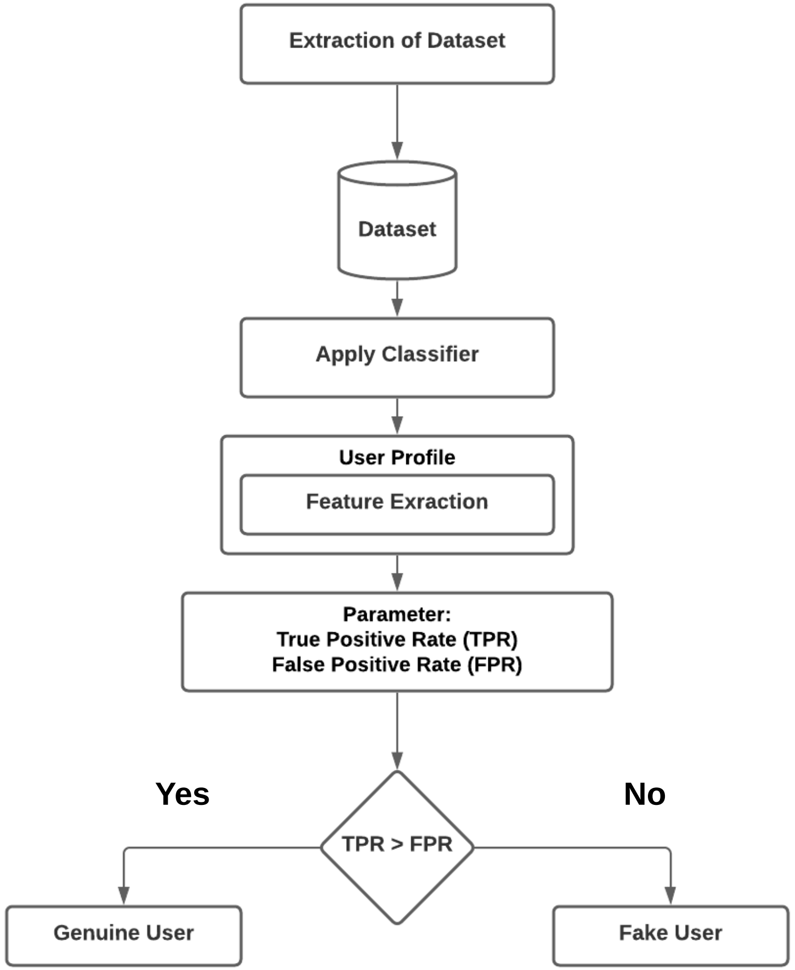
Analysis

i)System Architechture


Select the profile to be tested
Extract the Attribute Required
Pass it through the trained classifier
Determine Profile is Real or Fake
Get the feedback of the result
Train the Classifier using the feedback

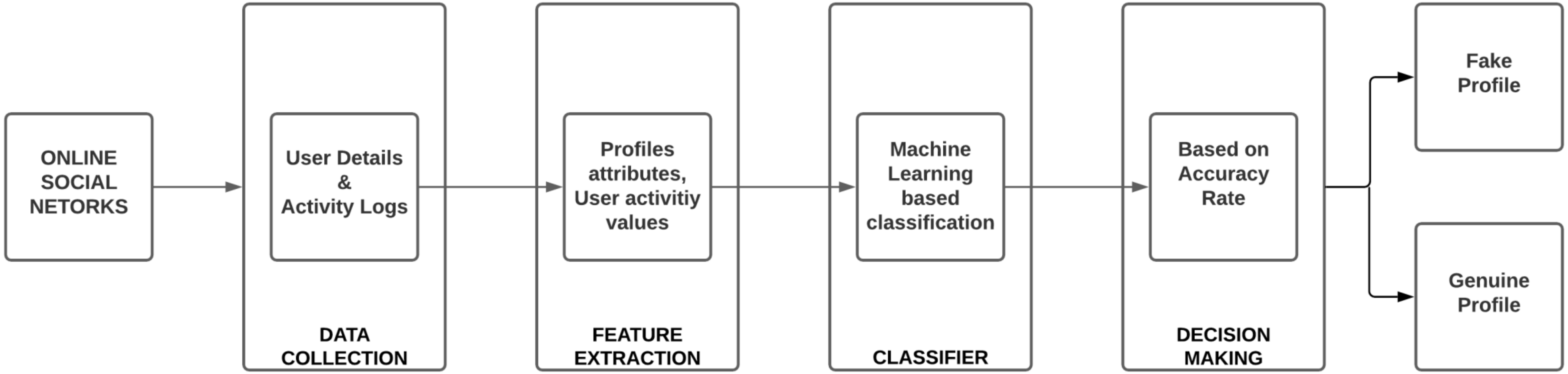
SYSTEM ARCHITECHTURE

ii) Flowchart



FLOWCHART DIAGRAM OF PROPOSED SYSTEM

iii)Flow Diagram



FLOW DIAGRAM OF PROPOSED SYSTEM

Design and Methodology

i)Proposed System

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

- The detection process starts with the selection of the profile that needs to be tested.
- After the selection of the profile, the suitable attributes (i.e.Profile ID, Profile Name,Status Count,Friends Count,Followers Count,Gender,Favorites Count,Language Code) are selected on which the classification algorithm is implemented.
- The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is feed into the classifier.
- The classifier determines whether the profile is fake or genuine.

- The classifier may not be 100% accurate in classifying the profile so; the feedback of the result is given back to the classifier.
- This process repeats and as the time proceeds, the no. of training data increases and the classifier becomes more and more accurate in predicting the fake profiles.

ii)Proposed Algorithms

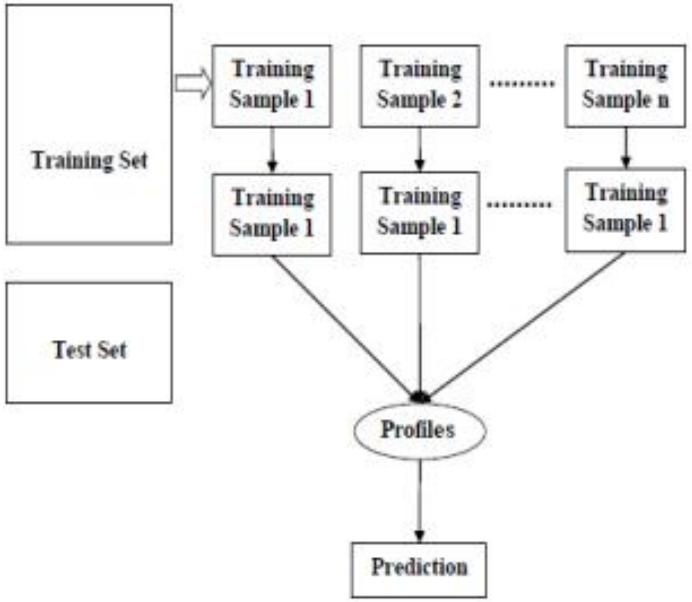
RANDOM FOREST

- 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:

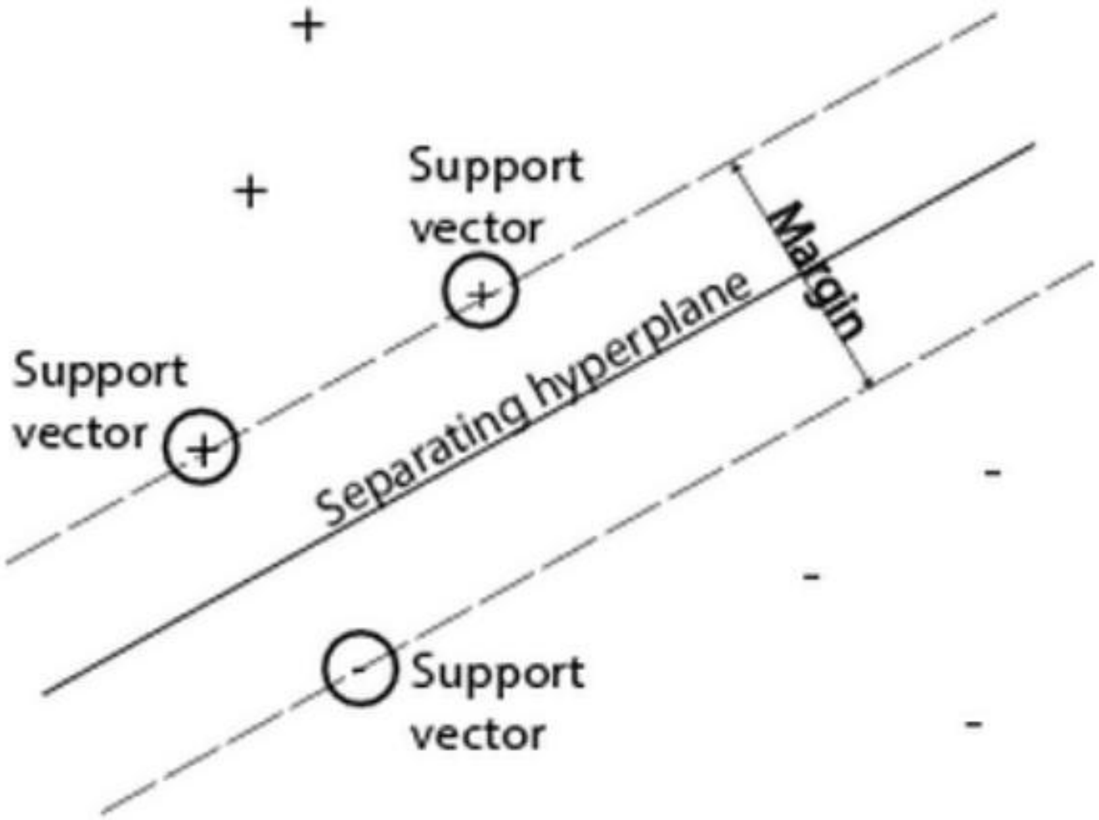
- Step 1 − First, start with the selection of random samples from a given dataset.
- Step 2 − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- Step 3 − In this step, voting will be performed for every predicted result.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

WORKING OF RANDOM FOREST :



SUPPORT VECTOR MACHINE

- Support Vector Machine is an elegant and robust technique for classification on a large data set not unlike the data sets of Social Network with several millions of profiles.
- It is a binary classification algorithm that finds the maximum separation hyper plane between two classes.
- SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The support vectors are the data points that are closest to the separating hyperplane.



- ➢ The figure illustrates linear classification, with + indicating data points of type 1, and - indicating data points of type 0.
- ➢ The datasets that we have used cannot be classified using linear classifier.
- ➢ The implementation of SVM is done on Matlab.

Technology Stack

Libraries:

- Numpy
- Pandas
- Matlplotlib
- pip
- Scikit-learn
- Sexmachine

Referances

- Yasyn Elyusufi (&) , Zakaria Elyusufi, and M’hamed Ait Kbir (2020) : **Social Networks Fake Profiles Detection Using Machine Learning Algorithms** In LIST Laboratory, Faculty of Sciences and Technologies, Tangier, Morocco.
- 1P. Srinivas Rao, 2Dr. Jayadev Gyani, 3Dr.G.Narsimha (2018) : **Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP** In Research Scholar, JNTUK, Kakinada, India.
- Gayathri A, Radhika S, Mrs. Jayalakshmi S.L(2018).: **Detecting Fake Accounts in Media Application Using Machine Learning.**
- S. P. Maniraj, Harie Krishnan G, Surya T, Pranav R (2019) : **Fake Account Detection using Machine Learning and Data Science** In International Journal of Innovative Technology and Exploring Engineering (IJITEE)

- Ahmed El Azab, Amira M. Idrees, Mahmoud A. Mahmoud, Hesham Hefny (2016): **Fake Account Detection in Twitter Based on Minimum Weighted Feature set** in International Journal of Computer, Electrical, Automation, Control and Information Engineering.
- Ananya Dey1, Hamsashree Reddy2 ,Manjistha Dey3 and Niharika Sinha4 (2019): **Detection of Fake Accounts in Instagram Using Machine Learning** In National Institute of Technology, Tiruchirappalli,India.
- Akshatha T M, Dr. M. N Veena (2020): **Machine Learning Framework for Detecting Spammer and Fake Users on Twitter** In International journal of engineering research & technology (IJERT)
- K. Ojo, A. (2019) : **Improved Model for Detecting Fake Profiles in Online Social Network: A Case Study of Twitter** In University of Ibadan, Nigeria.

- Kristo Radion Purba, David Asirvatham, Raja Kumar Murugesan (2020): **Classification of instagram fake users using supervised machine learning algorithms** in International Journal of Electrical and Computer Engineering (IJECE)
- Sneha Rane, Megha Ainapurkar, Ameya Wadekar (2018): **Detection of Compromised Accounts in Online Social Network** In International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
- S.Sandeep Bhat, M. Vishnu Priya (2020): **Recognition Of Fake Profile In Online Social Networks Using Machine Learning** In International journal of engineering research & technology (IJERT)
- Mohammed Jabardi, Imohammed Jabard (2020) : **Twitter Fake Account Detection and Classification using Ontological Engineering and Semantic Web Rule Language** In Karbala International Journal of Modern Science.

- Yeshwant Singh, Subhasish Banerjee (2018): **Fake (Sybil) Account Detection using Machine Learning** in National Institute of Technology.