A Project report on

FAKE ACCOUNT DETECTION USING MACHINE LEARNING AND DATA SCIENCE

A Dissertation submitted to JNTU Hyderabad in partial fulfillment of the academic requirements for the award of the degree.

Bachelor of Technology

in

Computer Science and Engineering

Submitted by

MOIN ASHIQ
19H51A05P0

M. MOUNIKA
19H51A05N8

N. HARIKA RATNA
19H51A05P4

Under the esteemed guidance of

Mrs. M. KAMALA
Assistant Professor

Department of Computer Science and Engineering

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institution under UGC & JNTUH, Approved by AICTE, Permanently Affiliated to JNTUH, Accredited by NBA.)

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD - 501401.

2019- 2023

CMR COLLEGE OF ENGINEERING & TECHNOLOGY

KANDLAKOYA, MEDCHAL ROAD, HYDERABAD – 501401

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Major Project Phase - 1 report entitled **"Fake Account Detection using Machine Learning and Data Science"being** submitted by Moin Ashiq (19H51A05P0), Madham Mounika(19H51A05N8), N. Harika Rathna (19H51A05P4) in partial fulfillment for the award of **Bachelor of Technology in Computer Science and Engineering** is a record of Bonafide work carried out his/her under my guidance and supervision.

The results embody in this project report have not been submitted to any other University or Institute for the award of any Degree.

Mrs. M. Kamala
Asst.Professor
Dept. Of CSE

Dr. Siva Skandha Sanagala
Associate Professor and HOD
Dept. of CSE

Acknowledgement

With great pleasure I want to take this opportunity to express my heartfelt gratitude to all the people who helped in making this project work a grand success.

We are grateful to **Mrs. M. Kamala** Assistant Professor, Dept. of Computer Science and Engineering for his valuable technical suggestions and guidance during the execution of this project work.

We would like to thank **Dr. S.SivaSkandhaSanagala,** Head of the Department of Computer Science and Engineering, CMR College of Engineering and Technology, who is the major driving forces to complete my project work successfully.

We are very grateful to **Dr. Vijaya Kumar Koppula**, Dean-Academic, CMR College of Engineering and Technology, for his constant support and motivation in carrying out the project work successfully.

We are highly indebted to **Dr. V A Narayana,** Principal, CMR College of Engineering and Technology, for giving permission to carry out this project in a successful and fruitful way.

We would like to thank the Teaching & Non- teaching staff of Department of Computer Science and Engineering for their co-operation.

We express our sincere thanks to **Dr. Ch. Gopal Reddy**, Secretary, CMR Group of Institutions, for his continuous care.

Finally, we extend thanks to our parents who stood behind us at different stages of this project. We sincerely acknowledge and thank all those who gave support directly and indirectly in completion of this project work.

MOIN ASHIQ (19H51A05P0)
MOUNIKA (19H51A05N8)
N. HARIKARATNA(19H51A05P4)

TABLE OF CONTENTS

List of Figures

List of Tables

ABSTRACT

Nowadays, online social media is dominating the world in several ways. Day by day the number of users using social media is increasing drastically. The main advantage of online social media is that we can connect to people easily and communicate with them in a better way. This provided a new way of a potential attack, such as fake identity, false information, etc. A recent survey suggests that the number of accounts present in the social media is much greater than the users using it. This suggests that fake accounts have been increased in the recent years. Online social media providers face difficulty in identifying these fake accounts. The need for identifying these fake accounts is that social media is flooded with false information, advertisements, etc.

Traditional methods cannot distinguish between real and fake accounts efficiently. Improvement in fake account creation made the previous works outdated. The new models created used different approaches such as automatic posts or comments, spreading false information or spam with advertisements to identify fake accounts. Due to the increase in the creation of the fake accounts different algorithms with different attributes are use. Previously use algorithms like naïve bayes, support vector machine, random forest has become inefficient in finding the fake accounts. In this research, we came up with an innovative method to identify fake accounts. We used gradient boosting algorithm with decision tree containing three attributes. Those attributes are spam commenting, artificial activity and engagement rate. We combined Machine learning and Data Science to accurately predict fake accounts.

CHAPTER 1
INTRODUCTION

CHAPTER 1

INTRODUCTION

1.1 Problem Statement:

Social media is an essential part of everyone's life in today's modern world. The main aim of social media is to stay in contact with friends and share news, among other things. The number of people who use social media is rapidly growing. Instagram is a global social media platform that has recently grown in popularity. Instagram has over 1 billion active users, making it one of the most popular social media platforms. People with a large number of followers have been dubbed Social Media Influencers since the introduction of Instagram to the social media scene. These social media influencers have now become a popular place for businesses to promote their goods and services.

Why Fake account detection is important?

- ➢ Gives more security for our profiles.
- ➢ Reduces creation of false news feeds & accounts.
- ➢ Decreasing of abuse, payment fraud and identity of theft.

Fig 1.1 Fake Accounts

The widespread use of social media has turned out to be both a benefit and a liability for society. The use of social media for online fraud and the dissemination of false information is rapidly growing. On social media, fake accounts are the most popular source of false information. Businesses that spend a lot of money on social media influencers need to know if the following they've gotten is organic or not.

1.2. Research Objective

In today's Modern society, social media plays a vital role in everyone's life. The general purpose of social media is to keep in touch with friends, sharing news, etc. The number of users in social media is increasing exponentially. Instagram has recently gained immense popularity among social media users. With more than 1 billion active users, Instagram has become one of the most used social media sites. After the emergence of Instagram to the social media scenario, people with a good number of followers have been called Social Media Influencers. These social media influencers have now become a go-to place for the business organization to advertise their products and services.

1.3 Project Scope and Objectives:

Project Scope:

In the present generation, the social life of everyone has become associated with the online social networks. Adding new friends and keeping in contact with them and their updates has become easier. The online social networks have impact on the science, education, grassroots organizing, employment, business, etc. Researchers have been studying these online social networks to see the impact they make on the people. Teachers can reach the students easily through this making a friendly environment for the students to study, teachers nowadays are getting themselves familiar to these sites bringing online classroom pages, giving homework, making discussions, etc. which improves education a lot. The employers can use these social networking sites to employ the people who are talented and interested in the work, their background check can be done easily.

Fig 1.2 Graph Showing Increase number of fake accounts

Objectives:

1. To reduce creation of fake accounts.

2. Save money, time, & details.

3.Give more security for their profiles

2. No abuse content is created

CHAPTER 2
BACKGROUND WORK

CHAPTER 2

BACKGROUND WORK

2.1. FAKE ACCOUNT DETECTION USING SVM:

2.1.1. Introduction

SVM (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 construct their solution as a weighted sum of SVs, which are only a subset of the training input .

2.1.2. Merits, Demerits and Challenges

Merits:

i. New features and parameters are added.

ii. Perform better than existing ones.

Demerits:

i. Not accurate.

ii. Limited data to train.

iii. High variance problems on increasing dataset.

iv. Only few parameters are used.

Challenges:

i. Use more parameters.

ii. Make all problems satisfied.

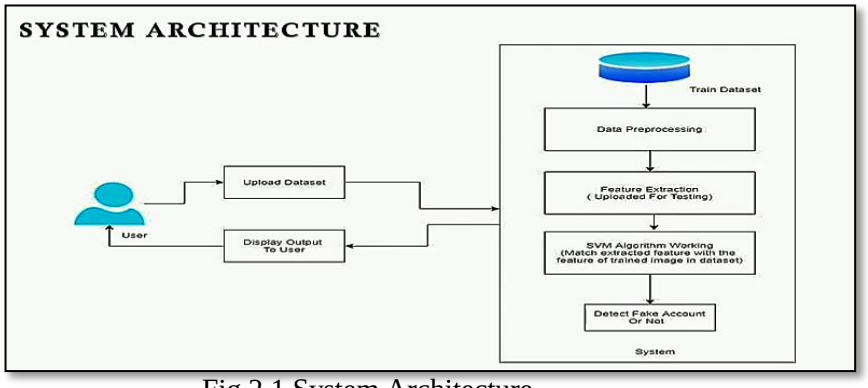


Fig 2.1 System Architecture

2.1.3. Implementation of FAKE ACCOUNT DETECTION USING SVM

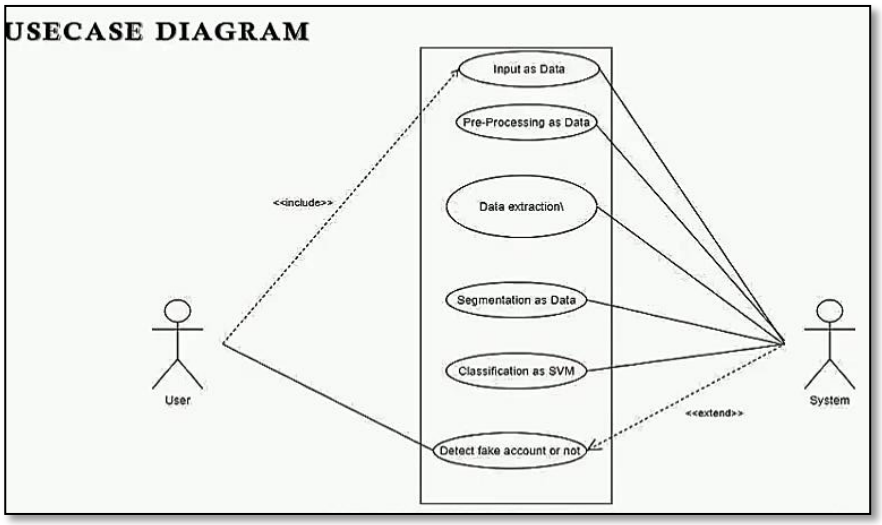


Fig 2.2 UML Diagrams

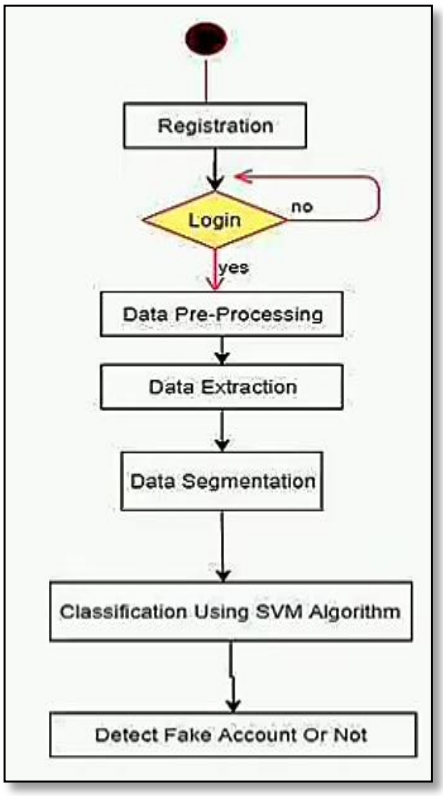


Fig 2.3 Activity diagram

Results:

Fig 2.4 Fake Account Detection Using SVM

2.2. FAKE ACCOUNT DETECTION USING DNN:

2.2.1. Introduction

The research study by Simranjit Kaur et al is based on implementing a k-mean clustering algorithm on vector set to increase efficiency. To detect spam emails using neural networks the two phases namely training and testing are needed to be done. The process of detecting spam and phishing emails using feed forward neural network. The paper has 11 features have been implemented as a binary value 0 or 1 with value 1 indicating this feature appeared in the tested email and value 0 indicating non-appearance case.

2.2.2. Merits, Demerits and Challenges

Merits:

i. Making grouping increase accuracy.

ii. Detects fast.

iii. Implementation done by using binary values.

Demerits:

i. No spam detection method is used.

ii. Not robust.

iii. It only maps to small number of character classes.

iv. Not a language-insensitive pattern matching features.

Challenges:

i. Need more fastest programming.

ii. All the character classes should be involved.

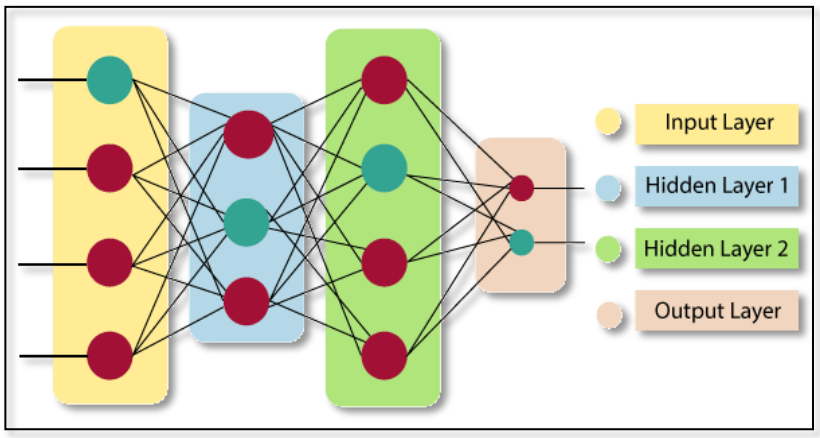
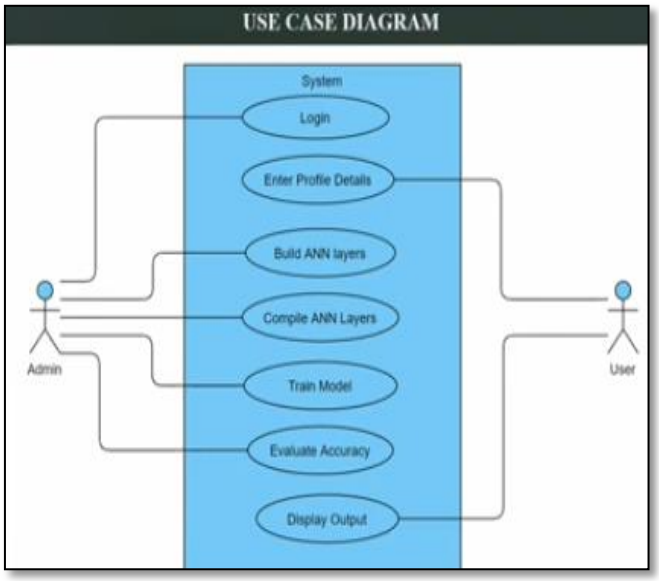


Fig 2.5 Challenges Layers

2.2.3. Implementation of Fake account detection using DNN:

UML Diagrams



Results:

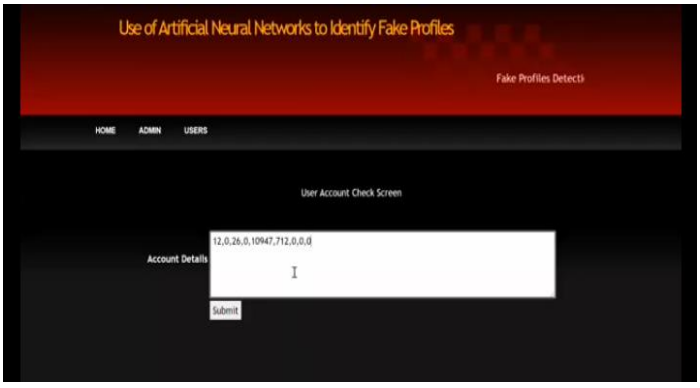


Fig 2.6 Fake account detection using DNN

2.3. FAKE ACCOUNT DETECTION USING RANDOM FOREST

2.3.1. Introduction

Random forest is a supervised learning algorithm that is used for both classifications as well as regression. Similarly, the 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. It is an ensemble method that is better than a single decision tree because it reducesthe over-fitting by averaging the result.

2.3.2. Merits, Demerits and Challenges

Merits:

i. More accurate.

ii. Creates many variations of trees.

iii. Give best outcomes than others.

Demerits:

iv. Not accurate than ours.

v. Uses only few features like comments & comportments.

Challenges:

vi. Use more features like complaints, etc.

vii. Can automatically remove the fake one.

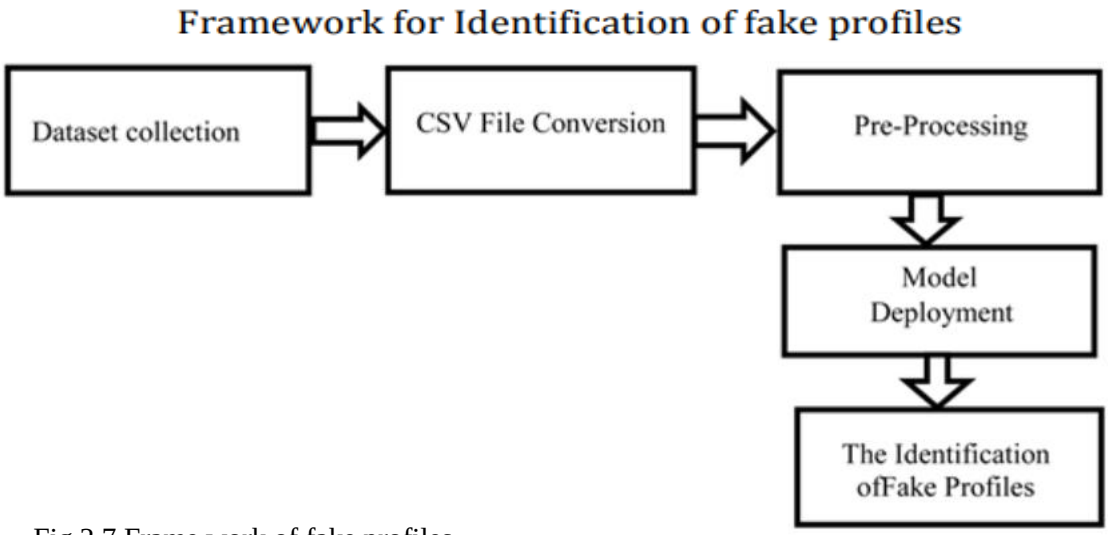
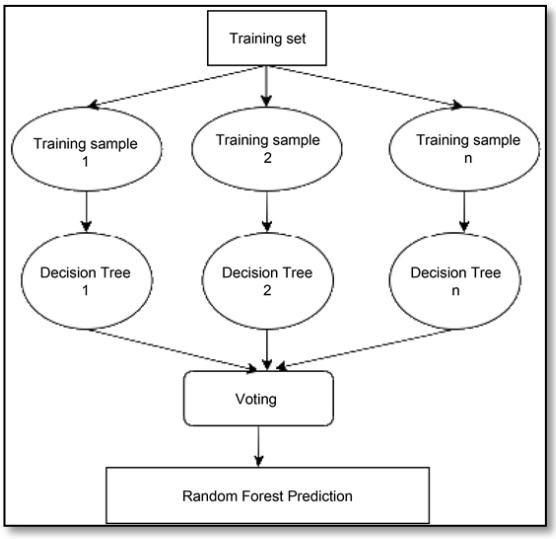


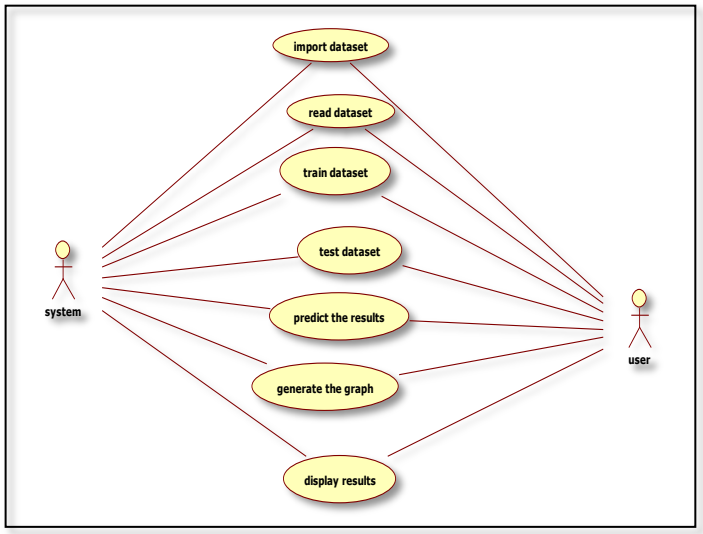
Fig 2.7 Frame work of fake profiles

2.3.3. Implementation of Fake account detection using Random Forest

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

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





USE CASE DIAGRAM

Results:

Fig 2.8 Fake account detection using Random Forest

CHAPTER 3
RESULTS AND DISCUSSION

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Comparison of Existing Solutions:

We also conducted experiments using SGD + Momentum weight updates and found out that it takes too long to cover the entire data set. We ran our model up to 20 epochs after which it began to over fit. Thus identifying the profile is real or fake. We used sparse vector representation of tweets for training the classifier. We identify that the presence of bigrams features significantly improved the accuracy. The overall accuracy across all machine learning models was very high with the highest being 94.43% using Deep Neural Networks and 94% using Random Forest method and finally 90.01% using Support Vector Machine algorithm. These results are just below what one would expect from getting the prediction right by chance.

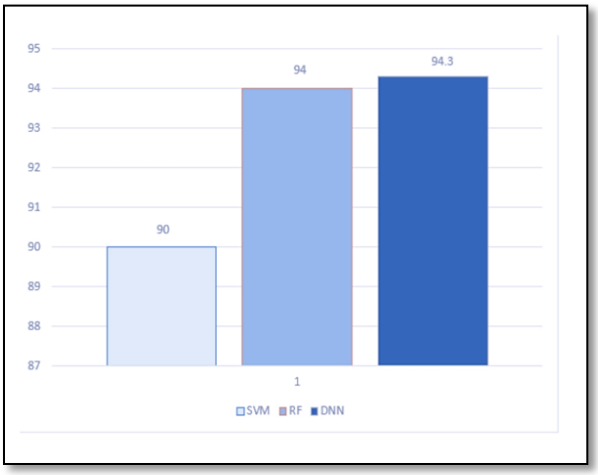
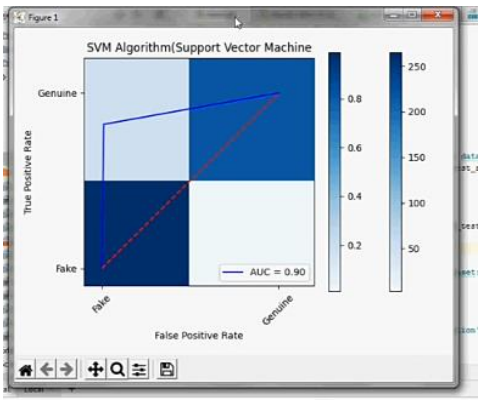
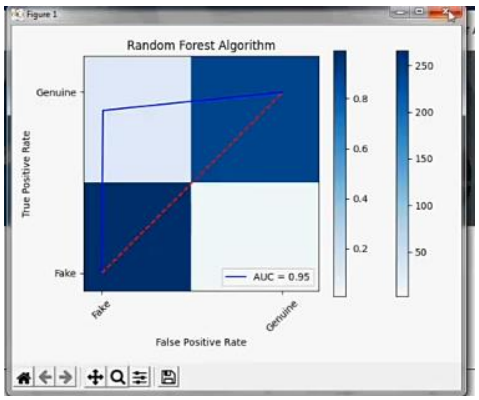


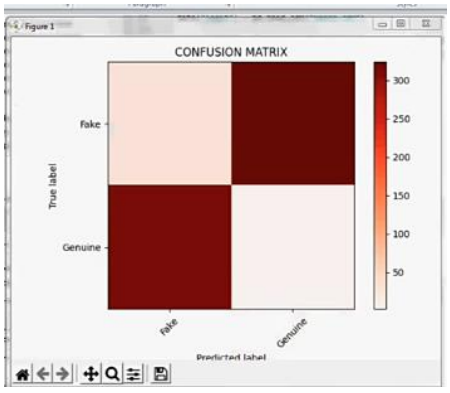
Fig 3.1 Graph of comparison



SVM



Random forest

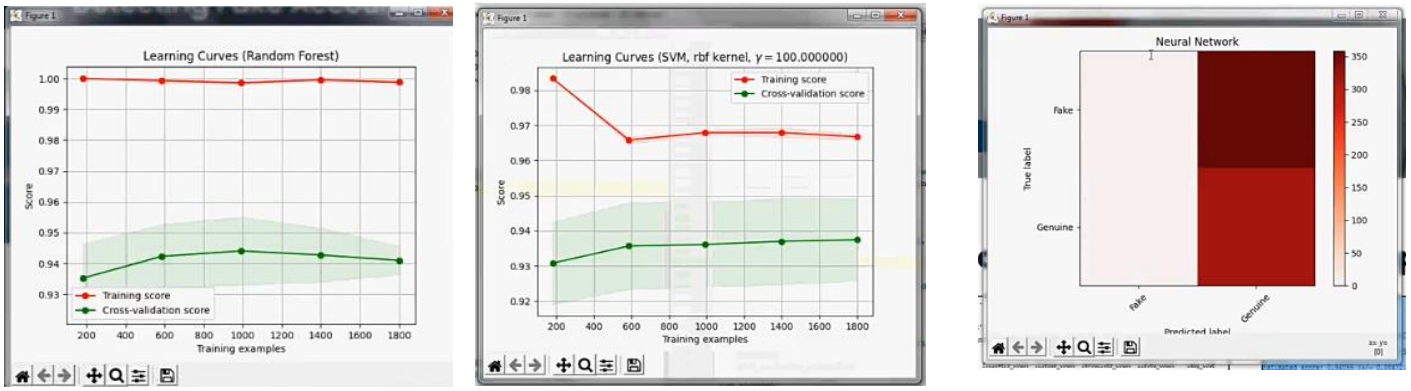


Neural Network

Fig 3.2 Metrics of Algorithm

Classifiers	Feature selection	Xgboost	Adaboost	GBM
Accuracy	Chi2	0.958	0.942	0.952
Precision		0.951	0.911	0.939
Recall		0.898	0.887	0.906

Table 3.1 Boosting classifier performance on features



SVM **Random forest** **Neural Network**

Fig 3.3 Comparison of graph

3.2. Data Collection and Performance metrics

i. Uploading the data.
ii. Dataset pre-processing
iii. Choosing a feature
iv. A method for detecting fraudulent accounts and comparing the results.

A	B	C	D	E	F	G	H	I	J
Raju	5/6/1985	Male	#7882,4th Cross,Rajajinagar	Rajajinagar	raju123@gmail.com	9535866270	Hp Laptop	Laptops are manufactured by HP and distributed all over the world	Rajajinagar
Kannan	9/4/2000	Male	#672,4th Cross,Vijayanagar	Vijayanagar	Kann.123@gmail.com	9535866270	Woment Safety	Women safety is question in India	Malleshwaram
Gokul	8/2/1998	Male	#829,18th Cross,Malleshwaram	Malleshwaram	Gokul.123@gmail.com	9535866270	Covid19	Corona virus is very dnagerous virus	Malleshwaram
Mala	9/6/1994	Female	#893,7th Cross,Yeshwantpur	Yeshwantpur	Mala.123@gmail.com	9535866270	Birds Fever	Bird Fever is common now a days in India	Yeshwantpur
Raju	5/6/1985	Male	#8928,7th Main,Gopi Nagar	Gopinagar	raju123@gmail.com	9535866270	CAA Protest	CAA Protest will be very dangerous scheme In India	Gopinagar
Vishnu	6/7/1998	Male	#627,7th Cross,Wilson Garden	Wilson Garden	Vishnu123@gmail.com	9535866270	Formers Protest	Formers protest will be success In India by Panjab Formers	Ashok Nagar
Gopi	9/6/1994	Male	#893,17th Cross,Yeshwantpur	Yeshwantpur	Gopi.123@gmail.com	9535866270	Dengue	It may spread to human being also	Yeshwantpur
Sasi	12/2/2000	Male	#8928,7th Main,RM Nagar	RM Nagar	Sasi123@gmail.com	9535866270	CAA	CAA Protest will be very dangerous scheme In India	RM Nagar
Kumar	6/7/1998	Male	#627,7th Cross,Wilson Garden	Wilson Garden	Kumar123@gmail.com	9535866270	Acer Desktop	Acer Desktop is manufactured by Acer and seeing throughout world	Wilson Garden
Suja	9/6/1994	Female	#893,7th Cross,Yeshwantpur	Yeshwantpur	Suja.123@gmail.com	9535866270	Sunflower	Sunflower oil is good to health and low cloestral	Rajajinagar
Sumo	5/6/1985	Male	#8928,7th Main,Banasawadi	Banasawadi	Sumo123@gmail.com	9535866270	CAA Protest1	CAA Protest will be very dangerous scheme In India	Banasawadi
Sarashwathi	6/7/1998	Female	#781,7th Cross,Wilson Garden	Wilson Garden	Sarashwathi123@gmail.c	9535866270	Eagle	Eagles are very fast in capturing small birds	Wilson Garden
Gopiraj	9/6/1994	Male	#12,7th Cross,Yeshwantpur	Yeshwantpur	Gopiraj.123@gmail.com	9535866270	H1N1	H1N1 is spreading from Birds	Yeshwantpur
Kamal	5/6/1985	Male	#89,17th Main,Samrajnagar	Samrajnagar	Kamal123@gmail.com	9535866270	CAA Protest2	CAA Protest will be very dangerous scheme In India	Samrajnagar
Vishnu	6/7/1998	Male	#627,7th Cross,Wilson Garden	Rajajinagar	Vishnu123@gmail.com	9535866270	Flue	Flue is a type of fever and spreads from animals	Ashok Nagar

Table 3.2 Data set

CHAPTER 4
CONCLUSION

CHAPTER 4
CONCLUSION

4.1 Conclusion

In this Project we 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. Our evaluation on both in-sample and out- of-sample data showed strong performance, and we have used the system in production to find and restrict more than 250,000 accounts. In this work we evaluated our framework on clusters created by simple grouping based on registration date and registration IP address. In future work we expect to run our model on clustering that are created by grouping on other features, such as ISP and other time periods, such as week or month.

Another promising line of research is to use more sophisticated clustering algorithms such as k-means or hierarchical clustering. While these approaches may be fruitful, they present obstacles to operating at scale: k- means may require too many clusters (i.e., too large a value of k) to produce useful results and clustering of data may be too intensive for classifying millions of accounts in Online Social Network.

From a modelling perspective, one important direction for future work is to apply feature sets used in other spam detection models, and hence to realize multi-model ensemble prediction. Another direction is to make the system robust against adversarial attacks, such as a botnet that diversifies all features, or an attacker that learns from failures.

CHAPTER 5
REFERENCES

CHAPTER 5
REFERENCES

[1]. Estee Van Der Walt and Jan Eloff,"Using Machine Learning to Detect Fake Identities:Bots vs Humans"IEEE Trans. Emerg.TopicsComput. Intell., vol. 1, no. 1, pp. 61–71 March 2018.

[2]. Loredana Caruccio,DomenicoDesiato and Giuseppe Polese"Fake Account Identification in Social Networks" IEEE International Conference on Big Data.,, vol. 9, no. 6, pp. 811–824,2018.

[3]. SarahKhaled,Neamat El-Tazi and Hoda M. O. Mokhtar"Detecting Fake Accounts on Social Media" IEEE International Conference on Big Data.., vol.6 pp 101-110 ,2018.

[4]. SuyashSomani and Somya Jain "Resolving Identities on FaceBook and Twitter" Tenth International Conference on Contemporary Computing (IC3), 10- 12 August 2017.

[5]. FrancescoBuccafurri, Gianluca Lax,Denis Migdal, Serena Nicolazzo, Antonino Nocera and Christophe Rosenberger"Contrasting False Identities in Social Networks by Trust Chains and Biometric Reinforcement " International Conference on Cyberworlds vol 5,2017.

[6]. SuprajaGurajala, Joshua S White, Brian Hudson,Brian R Voter and Jeanna N Matthews"Profile characteristics of fake T witter accounts"Big Data &Society,July–December 2016: 1–13.

[7]. Simranjit. Kaur. Tuteja, ‘‘A survey on classification algorithms for email spam filtering,’’ International Journal Eng. Sci., vol. 6, no. 5, pp. 5937–5940, 2016.

[8]. M.A.Devmane and N.K.Rana "Detection and Prevention of Profile Cloning in Online Social Networks"IEEE International Conference on Recent Advances and Innovations in Engineering,May 09-11, 2014.

[9]. SaraKeretna ,Ahmad Hossny and Doug Creighton "Recognising User Identity in T witter Social Networks via Text Mining"IEEE International Conference on Systems, Man, and Cybernetics,2013.

[10]. MohamedT orky, Ali Meligy and Hani Ibrahim"Recognizing Fake Identities In Online Social Networks Based on a Finite Automaton Computer Approach"International Journal of Applications,2016.