

A Major Project Report
on
“FAKE JOB POST PREDICTION USING MLDL”

Submitted to the
JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY
HYDERABAD

In partial fulfilment of the requirement for the award of the degree of
BACHELOR OF TECHNOLOGY

IN
Computer Science and Engineering
BY

R. POOJITHA (19WJ5A0527)

Under the Esteemed Guidance of
Mrs. CH. SUSHMA
Asst. Professor, CSE Dept.



GURU NANAK INSTITUTIONS TECHNICAL CAMPUS (AUTONOMOUS)

School of Engineering and Technology
Ibrahimpattanam, R.R District 501506

2021-2022



GURU NANAK INSTITUTIONS TECHNICAL CAMPUS



Approved by
AICTE - New Delhi



Affiliated to
JNTU - Hyderabad



Accredited by
National Assessment and
Accreditation Council



Accredited by
National Board of
Accreditation

AUTONOMOUS
under Section 2 (f) & 12 (b) of
University Grants Commission Act

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that this project report entitled “**FAKE JOB POST PREDICTION USING MLDL**” by **RAM POOJITHA (19WJ5A0527)** submitted in partial fulfilment the requirements for the degree of **Bachelor of Technology in Computer Science and Engineering** of the **Jawaharlal Nehru Technological University Hyderabad** during the academic year 2018-2022, is a bonafide record of work carried out under our guidance and supervision.

INTERNAL GUIDE

Mrs. CH. Sushma

PROJECT COORDINATOR

Mrs. V. Swathi

HOD CSE2

Mr. V. Devasekhar

EXTERNAL EXAMINER

**RAM Innovative Infotech**

M : +91 9581 012 012
E : raminnovativeinfotech@gmail.com
Flat No.#309, Amrutha Ville,
Opp: Yashoda Hospital, Somajiguda,
Hyderabad-82, Telangana, India
www.raminnovativeinfotech.webs.com

PROJECT COMPLETION CERTIFICATE

This is to certify that the following students of final year B. Tech
Department of Computer Science and Engineering - Guru Nanak
Institutions Technical Campus (GNITC) have completed their training and project
at GNITC successfully.

STUDENT NAME:**Roll No:**

1. RAM POJITHA
2. _____
3. _____
4. _____

19WJ5A0527

The training was conducted on MLDL Technology
for the completion of the project titled "FAKE JOB POST PREDICTION
USING MLDL"
in RAM INNOVATIVE INFOTECH. The project has been
completed in all aspects.


Signature

Acknowledgment

We wish to communicate our true gratitude to **Dr. Rishi Sayal, Associate Director**, GNITC for giving us the helpful condition to bringing through our scholastic calendars and undertaking effortlessly.

We have been genuinely honoured to have a brilliant guide **Mr. V. Devasekhar, Associate Professor** and **HOD CSE2**, GNITC for managing us to investigate the implication of our work and we offer our true thanks towards him for driving us through the consummation of undertaking.

We particularly thank our internal guide **Mrs. CH. Sushma, Asst.Professor**, Department of CSE, for offering consistent help and right recommendations are given in the improvement of the project. At long last, we might want to thank our relatives for their ethical help and support to accomplish objectives.

We particularly thank our major-project coordinator **Mrs. V. Swathi, Asst.Professor**, Department of CSE, for offering consistent help and right recommendations are given in the improvement of the project. At long last, we might want to thank our relatives for their ethical help and support to accomplish objectives.

R. POOJITHA	18WJ1A0527
N. CHANDANA	19WJ5A0527
U. BHANU PRAKASH	18WJ1A05Z2

Fake Job Post Prediction Using MLDL

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	vi
	LIST OF FIGURES	vii
	LIST OF SYMBOLS	viii
1.	CHAPTER 1 : INTRODUCTION	
	1.1 GENERAL	1
	1.2 OBJECTIVE	2
	1.3 SCOPE OF THE PROJECT	2
	1.4 EXISTING SYSTEM	3
	1.3.1 EXISTINGSYSTEM DISADVANTAGES	3
	1.3.2 LITERATURE SURVEY	4
	1.5 PROPOSED SYSTEM	9
	1.5.1 PROPOSED SYSTEM ADVANTAGES	9
2.	CHAPTER 2 :PROJECT DESCRIPTION	
	2.1 GENERAL	10
	2.2 METHODOLOGIES	10
	2.2.1 MODULES NAME	10
	2.2.2 MODULES DESCRIPTION	11
	2.3 TECHNIQUE OR ALGORITHM USED	14
3.	CHAPTER 3 : REQUIREMENTS	
	3.1 GENERAL	19
	3.2 HARDWARE REQUIREMENTS	19
	3.3 SOFTWARE REQUIREMENTS	20
	3.4 FUNCTIONAL REQUIREMENTS	20
	3.5 NON-FUNCTIONAL REQUIREMENTS	21

4.	CHAPTER 4 :SYSTEM DESIGN 4.1 GENERAL 4.2 UML 4.2.1 USE CASE DIAGRAM 4.2.2 CLASS DIAGRAM 4.2.3 OBJECT DIAGRAM 4.2.4 SEQUENCED DIAGRAM 4.2.5 COLLABARATION DIAGRAM 4.2.6 DEPLOYMENT DIAGRAM 4.2.7 ACTIVITY DIAGRAM 4.2.8 STATE DIAGRAM 4.2.9 DATA FLOW DIAGRAM 4.3 SYSTEM ARCHITECTURE	22 23 23 24 25 26 27 28 29 30 31 32
5.	CHAPTER 5 :SOFTWARE SPECIFICATION 5.1 GENERAL	33
6.	CHAPTER 6 :IMPLEMENTATION 6.1 GENERAL	36

7.	CHAPTER 7 :SNAPSHOTS 7.1 GENERAL 7.2 SNAPSHOTS	39 39
8.	CHAPTER 8 :SOFTWARE TESTING 8.1 GENERAL 8.2 DEVELOPING METHODOLOGIES 8.3 TYPES OF TESTING 8.3.1 UNIT TESTING 8.3.2 FUNCTIONAL TEST 8.3.3 SYSTEM TEST 8.3.4 PERFORMANCE TEST 8.3.5 INTEGRATION TESTING 8.3.6 ACCEPTANCE TESTING 8.3.7 BUILD THE TEST PLAN	49 49 49 49 50 50 50 51 51 51
9.	CHAPTER 9 : FUTURE ENHANCEMENT 9.1 FUTURE ENHANCEMENTS	52
10	CHAPTER 10 : 10.1 CONCLUSION 10.2 REFERENCES	53 54

ABSTRACT

In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This project proposed to use different data mining techniques and classification algorithm like KNN, decision tree, random forest classifier and deep neural network to predict a job post if it is real or fraudulent.

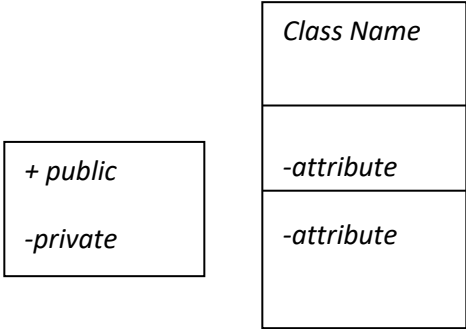
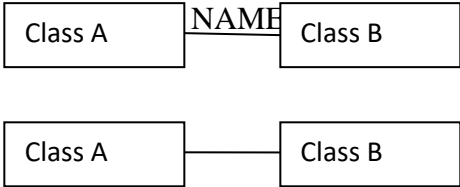
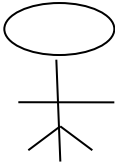
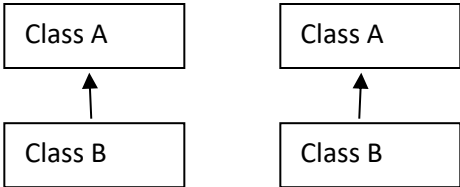
We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task.

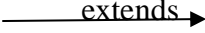


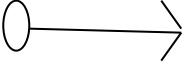
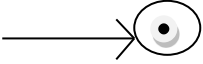
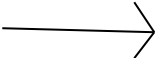
We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

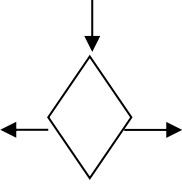
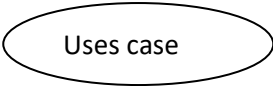
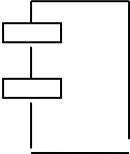
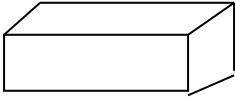
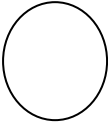
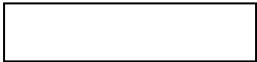
LIST OF FIGURES

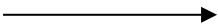
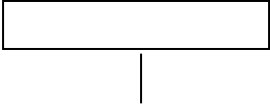
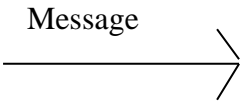
FIGURE NO	NAME OF THE FIGURE	PAGE NO.
4.2.1	Use case Diagram	23
4.2.2	Class Diagram	24
4.2.3	Object Diagram	25
4.2.4	Sequence Diagram	26
4.2.5	Collaboration Diagram	27
4.2.6	Deployment Diagram	28
4.2.7	Activity Diagram	29
4.2.8	State Diagram	30
4.2.9	Data Flow Diagram	31
4.3	System Architecture	32

LIST OF SYMBOLS

S.NO	NOTATION NAME	NOTATION	DESCRIPTION
1.	Class		Represents a collection of similar entities grouped together.
2.	Association		Associations represents static relationships between classes. Roles represents the way the two classes see each other.
3.	Actor		It aggregates several classes into a single classes.
4.	Aggregation		Interaction between the system and external environment

5.	Relation (uses)	uses	Used for additional process communication.
6.	Relation (extends)		Extends relationship is used when one use case is similar to another use case but does a bit more.
7.	Communication		Communication between various use cases.
8.	State		State of the processes.
9.	Initial State		Initial state of the object
10.	Final state		Final state of the object
11.	Control flow		Represents various control flow between the states.

12.	Decision box		Represents decision making process from a constraint
13.	Use case		Interact ion between the system and external environment.
14.	Component		Represents physical modules which are a collection of components.
15.	Node		Represents physical modules which are a collection of components.
16.	Data Process/State		A circle in DFD represents a state or process which has been triggered due to some event or action.
17.	External entity		Represents external entities such as keyboard,sensors,etc.

18.	Transition		Represents communication that occurs between processes.
19.	Object Lifeline		Represents the vertical dimensions that the object communications.
20.	Message		Represents the message exchanged.

CHAPTER 1

INTRODUCTION

1.1 GENERAL

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers.

So, people lack in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information.

Thus the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's belief and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

1.2 OBJECTIVE

We have used different data mining techniques to predict if a job post is fake or not. We have trained EMSCAD data in the classifiers after a pre-processing step. The trained classifier act as an online fake job post detector. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real-life fake job posts.

In this project we have experimented both machine learning algorithms (SVM, KNN, Naive Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99% accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

1.3 SCOPE OF THE PROJECT

This project proposed to Random Forest Classifier to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. The trained classifier shows approximately 98% classification accuracy to predict a fraudulent job post.

1.4 Existing System

- Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers
- So, to reduce the fake job posts, here they used Text Processing to differentiate true and fake posts & Dataset Training
- Zhang proposed an automatic fake detector model to distinguish between true and fake news (including articles, creators, subjects) using text processing. They had used a custom dataset of news or articles posted by PolitiFact website twitter account.
- This dataset was used to train the proposed GDU diffusive unit model. Receiving input from multiple sources simultaneously, this trained model performed well as an automatic fake detector model.

1.4.1 Existing System Disadvantages

- Less Classification Accuracy
- Less Accuracy
- Less Precision rate

1.4.2 LITERATURE SURVEY:

TITLE: Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset

AUTHOR :S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu

YEAR : 2017

DESCRIPTION

The critical process of hiring has relatively recently been ported to the cloud. Specifically, the automated systems responsible for completing the recruitment of new employees in an online fashion, aim to make the hiring process more immediate, accurate and cost-efficient. However, the online exposure of such traditional business procedures has introduced new points of failure that may lead to privacy loss for applicants and harm the reputation of organizations. So far, the most common case of Online Recruitment Frauds (ORF), is employment scam. Unlike relevant online fraud problems, the tackling of ORF has not yet received the proper attention, remaining largely unexplored until now. Responding to this need, the work at hand defines and describes the characteristics of this severe and timely novel cyber security research topic. At the same time, it contributes and evaluates the first to our knowledge publicly available dataset of 17,880 annotated job ads, retrieved from the use of a real-life system.

TITLE :An Intelligent Model for Online Recruitment Fraud Detection
AUTHOR :B. Alghamdi, F. Alharby
YEAR : 2019

DESCRIPTION

This study research attempts to prohibit privacy and loss of money for individuals and organization by creating a reliable model which can detect the fraud exposure in the online recruitment environments. This research presents a major contribution represented in a reliable detection model using ensemble approach based on Random Forest classifier to detect Online Recruitment Fraud (ORF). The detection of Online Recruitment Fraud is characterized by other types of electronic fraud detection by its modern and the scarcity of studies on this concept. The researcher proposed the detection model to achieve the objectives of this study. For feature selection, support vector machine method is used and for classification and detection, ensemble classifier using Random Forest is employed. A freely available dataset called Employment Scam Aegean Dataset (EMSCAD) is used to apply the model. Pre-processing step had been applied before the selection and classification adoptions. The results showed an obtained accuracy of 97.41%. Further, the findings presented the main features and important factors in detection purpose include having a company profile feature, having a company logo feature and an industry feature.

TITLE : Job Prediction: From Deep Neural Network Models to Applications
AUTHOR : Tin Van Huynh¹, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen¹, and Anh Gia-Tuan Nguyen
YEAR : 2020

DESCRIPTION

Determining the job is suitable for a student or a person looking for work based on their job descriptions such as knowledge and skills that are difficult, as well as how employers must find ways to choose the candidates that match the job they require. In this paper, we focus on studying the job prediction using different deep neural network models including TextCNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN with various pre-trained word embeddings on the IT job dataset. In addition, we proposed a simple and effective ensemble model combining different deep neural network models. Our experimental results illustrated that our proposed ensemble model achieved the highest result with an F1-score of 72.71%. Moreover, we analyze these experimental results to have insights about this problem to find better solutions in the future.

TITLE : FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network

AUTHOR : Jiawei Zhang, Bowen Dong, Philip S. Yu

YEAR : 2020

DESCRIPTION

In recent years, due to the booming development of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by this online fake news easily, which has brought about tremendous effects on the offline society already. An important goal in improving the trustworthiness of information in online social networks is to identify the fake news timely. This paper aims at investigating the principles, methodologies and algorithms for detecting fake news articles, creators and subjects from online social networks and evaluating the corresponding performance. This paper addresses the challenges introduced by the unknown characteristics of fake news and diverse connections among news articles, creators and subjects. This paper introduces a novel automatic fake news credibility inference model, namely FAKEDETECTOR. Based on a set of explicit and latent features extracted from the textual information, FAKEDETECTOR builds a deep diffusive network model to learn the representations of news articles, creators and subjects simultaneously. Extensive experiments have been done on a real-world fake news dataset to compare FAKEDETECTOR with several state-of-the-art models, and the experimental results have demonstrated the effectiveness of the proposed model.

TITLE : Automatic Detection of Cyber Recruitment by Violent Extremists
AUTHOR : Scanlon, J.R. and Gerber, M.S
YEAR : 2014

DESCRIPTION

Growing use of the Internet as a major means of communication has led to the formation of cyber-communities, which have become increasingly appealing to terrorist groups due to the unregulated nature of Internet communication. Online communities enable violent extremists to increase recruitment by allowing them to build personal relationships with a worldwide audience capable of accessing uncensored content. This article presents methods for identifying the recruitment activities of violent groups within extremist social media websites. Specifically, these methods apply known techniques within supervised learning and natural language processing to the untested task of automatically identifying forum posts intended to recruit new violent extremist members. We used data from the western jihadist website Ansar AlJihad Network, which was compiled by the University of Arizona's Dark Web Project. Multiple judges manually annotated a sample of these data, marking 192 randomly sampled posts as recruiting (YES) or non-recruiting (NO). We observed significant agreement between the judges' labels; Cohen's $\kappa=(0.5,0.9)$ at $p=0.01$. We tested the feasibility of using naive Bayes models, logistic regression, classification trees, boosting, and support vector machines (SVM) to classify the forum posts. Evaluation with receiver operating characteristic (ROC) curves shows that our SVM classifier achieves an 89% area under the curve (AUC), a significant improvement over the 63% AUC performance achieved by our simplest naive Bayes model (Tukey's test at $p=0.05$). To our knowledge, this is the first result reported on this task, and our analysis indicates that automatic detection of online terrorist recruitment is a feasible task. We also identify a number of important areas of future work including classifying non-English posts and measuring how recruitment posts and current events change membership numbers over time.

1.5 Proposed System

- The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the jobseekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle is employed that provides information regarding a job that may or may not be suspicious.
- This dataset contains 17,880 number of job posts. This dataset is used in the proposed methods for testing the overall performance of the approach. For better understanding of the target as a baseline, a multistep procedure is followed for obtaining a balanced dataset. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset.
- Random Forest Classifier is applied for classifying job post as fake. The performance measure metrics such as Accuracy, Recall, Precision, and Confusion matrix are used for evaluating the prediction for proposed classifier.

1.5.1 Proposed System Advantages

- This approach reduces the number of trainable attribute effectively with less processing time.
- We have achieved approximately 98% classification accuracy (highest) for Random Forest classifier.
- We have analyzed performance analysis parameters also to check if the model works well at both false positive and false negative samples.
- Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, Random Forest Classifier has proved the best prediction results.

CHAPTER 2

PROJECT DESCRIPTION

2.1 GENERAL:

In this chapter, various supervised machine learning approaches are used. This section provides a general description of these approaches.

2.2 METHODOLOGIES

2.2.1 MODULES NAME:

- **Data Collection**
- **Dataset**
- **Data Preparation**
- **Model Selection**
- **Analyze and Prediction**
- **Accuracy on test set**
- **Saving the Trained Model**

2.2.2 MODULE DESCRIPTION

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Data set link: <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>

Dataset:

The dataset consists of 17880 individual data. There are 18 columns in the dataset, which are described below.

1. job_id - unique vacancy identifier
2. title - headline
3. location - the geographical location of the job advertisement
4. department - corporate department (for example, sales)
5. salary range - indicative salary range (eg 50,000-60,000)
6. company profile - a short description of the company
7. description - detailed description of the job advertisement
8. requirements - the requirements for the vacancy are listed
9. benefits - the proposed benefits are listed;
10. telecommuting - true for remote posts
11. has_company_logo- true if the company logo is present;
12. has_questions - true if test questions are present

13. employment type - type of employment;
14. required experience - necessary experience
15. required education - necessary education
16. industry - industry
17. function - function to be performed
18. fraudulent - indicates whether the job is fraudulent

Data Preparation:

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)

Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data

Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis

Split into training and evaluation sets

Model Selection:

We used Random Forest Classifier algorithm, We got a accuracy of 94.7% on test set so we implemented this algorithm.

Analyze and Prediction:

In the actual dataset, we chose only 2 features:

- 1 Description - detailed description of the job advertisement
- 2 Fraudulent - indicates whether the job is fraudulent

Accuracy on test set:

We got an accuracy of 95.02% on test set.

Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle.

Make sure you have pickle installed in your environment.

Next, let's import the module and dump the model into .pkl file

2.3 TECHNIQUE USED OR ALGORITHM USED

Proposed Algorithm

➤ The Random Forests Algorithm

Let's understand the algorithm in layman's terms. Suppose you want to go on a trip and you would like to travel to a place which you will enjoy.

So what do you do to find a place that you will like? You can search online, read reviews on travel blogs and portals, or you can also ask your friends.

Let's suppose you have decided to ask your friends, and talked with them about their past travel experience to various places. You will get some recommendations from every friend. Now you have to make a list of those recommended places. Then, you ask them to vote (or select one best place for the trip) from the list of recommended places you made. The place with the highest number of votes will be your final choice for the trip.

In the above decision process, there are two parts. First, asking your friends about their individual travel experience and getting one recommendation out of multiple places they have visited. This part is like using the decision tree algorithm. Here, each friend makes a selection of the places he or she has visited so far.

The second part, after collecting all the recommendations, is the voting procedure for selecting the best place in the list of recommendations. This whole process of getting recommendations from friends and voting on them to find the best place is known as the random forests algorithm.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the

tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

How does the algorithm work?

It works in four steps:

Select random samples from a given dataset.

Construct a decision tree for each sample and get a prediction result from each decision tree.

Perform a vote for each predicted result.

Select the prediction result with the most votes as the final prediction.

Advantages:

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.
- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

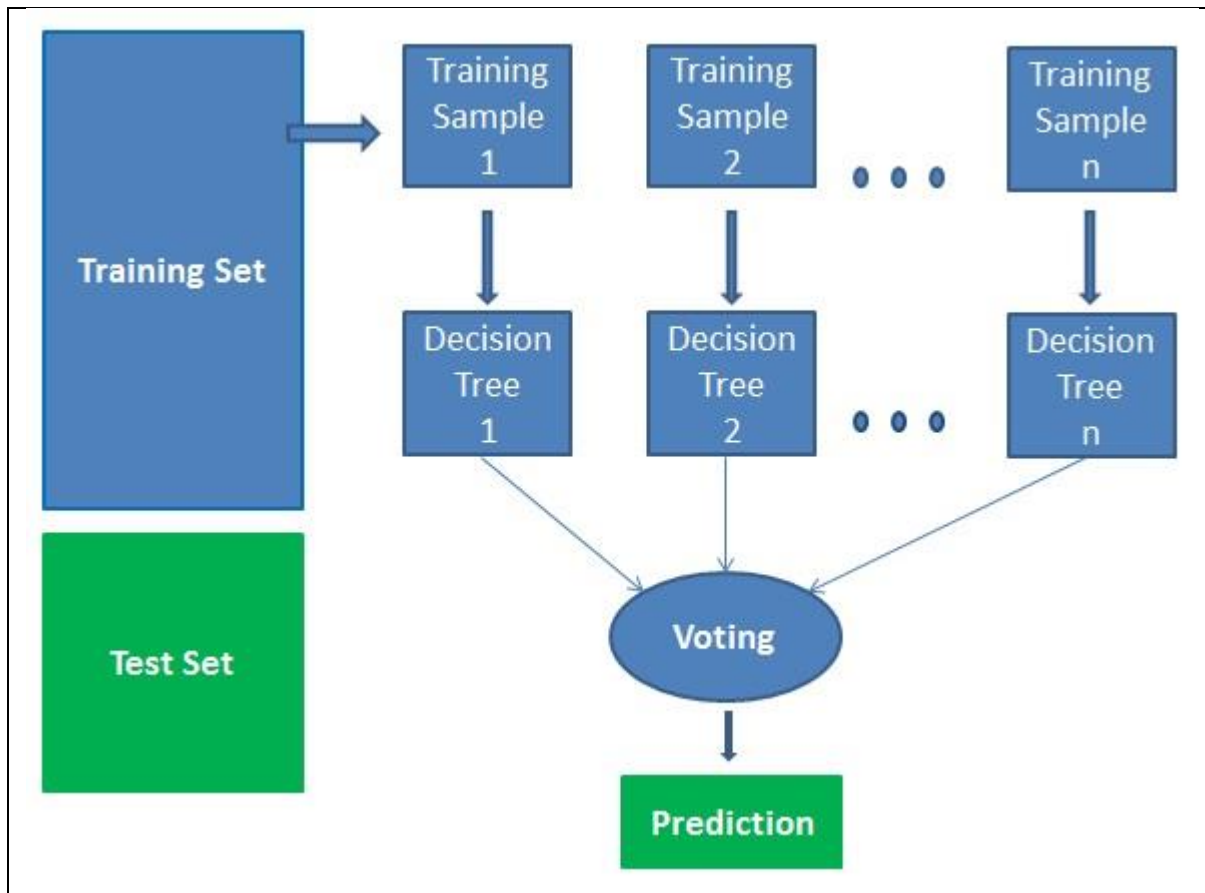


Fig: Random Forest Classifier

Disadvantages:

- Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

Finding important features

Random forests also offers a good feature selection indicator. Scikit-learn provides an extra variable with the model, which shows the relative importance or contribution of each feature in the prediction. It automatically computes the relevance score of each feature in the training phase. Then it scales the relevance down so that the sum of all scores is 1.

This score will help you choose the most important features and drop the least important ones for model building.

Random forest uses gini importance or mean decrease in impurity (MDI) to calculate the importance of each feature. Gini importance is also known as the total decrease in node impurity. This is how much the model fit or accuracy decreases when you drop a variable. The larger the decrease, the more significant the variable is. Here, the mean decrease is a significant parameter for variable selection. The Gini index can describe the overall explanatory power of the variables.

Random Forests vs Decision Trees

- Random forests is a set of multiple decision trees.
- Deep decision trees may suffer from overfitting, but random forests prevents overfitting by creating trees on random subsets.
- Decision trees are computationally faster.
- Random forests is difficult to interpret, while a decision tree is easily interpretable and can be converted to rules.

Analyze and Prediction:

In the actual dataset, we chose only 8 features

1. telecommuting - true for remote posts
2. has_company_logo- true if the company logo is present;
3. has questions - true if test questions are present
4. employment type - type of employment;

5. required experience - necessary experience
 6. required education - necessary education
 7. industry - industry
 8. function - function to be performed
- result : indicates whether the job is fraudulent

CHAPTER 3

REQUIREMENTS ENGINEERING

3.1 GENERAL

These are the requirements for doing the project. Without using these tools and software's we can't do the project. So we have two requirements to do the project. They are

1. Hardware Requirements.
2. Software Requirements.

3.2 HARDWARE REQUIREMENTS

The hardware requirements may serve as the basis for a contract for the implementation of the system and should therefore be a complete and consistent specification of the whole system. They are used by software engineers as the starting point for the system design. It should what the system do and not how it should be implemented.

- PROCESSOR : DUAL CORE 2 DUOS.
- RAM : 4GB DD RAM
- HARD DISK : 250 GB

3.3 SOFTWARE REQUIREMENTS

The software requirements document is the specification of the system. It should include both a definition and a specification of requirements. It is a set of what the system should do rather than how it should do it. The software requirements provide a basis for creating the software requirements specification. It is useful in estimating cost, planning team activities, performing tasks and tracking the teams and tracking the team's progress throughout the development activity.

- OPERATING SYSTEM : WINDOWS 7/8/10
- PLATFORM : SPYDER3
- PROGRAMMING LANGUAGE : PYTHON, HTML
- FRONT END : SPYDER3

3.4 FUNCTIONAL REQUIREMENTS

A functional requirement defines a function of a software-system or its component. A function is described as a set of inputs, the behavior, the presented result will help us in identifying the behaviour of employees who can be attired over the next time. Experimental results reveal that the logistic regression approach can reach up to 86% accuracy over other machine learning approaches.

3.5 NON-FUNCTIONAL REQUIREMENTS

The major non-functional Requirements of the system are as follows

➤ **Usability**

The system is designed with completely automated process hence there is no or less user intervention.

➤ **Reliability**

The system is more reliable because of the qualities that are inherited from the chosen platform java. The code built by using java is more reliable.

➤ **Performance**

This system is developing in the high level languages and using the advanced front-end and back-end technologies it will give response to the end user on client system with in very less time.

➤ **Supportability**

The system is designed to be the cross platform supportable. The system is supported on a wide range of hardware and any software platform, which is having JVM, built into the system.

➤ **Implementation**

The system is implemented in web environment using struts framework. The apache tomcat is used as the web server and windows xp professional is used as the platform. Interface the user interface is based on Struts provides HTML Tag.

CHAPTER 4

DESIGN ENGINEERING

4.1 GENERAL

Design Engineering deals with the various UML [Unified Modelling language] diagrams for the implementation of project. Design is a meaningful engineering representation of a thing that is to be built. Software design is a process through which the requirements are translated into representation of the software. Design is the place where quality is rendered in software engineering. Design is the means to accurately translate customer requirements into finished product.

4.2 UML Diagrams

4.2.1 USECASE DIAGRAM:

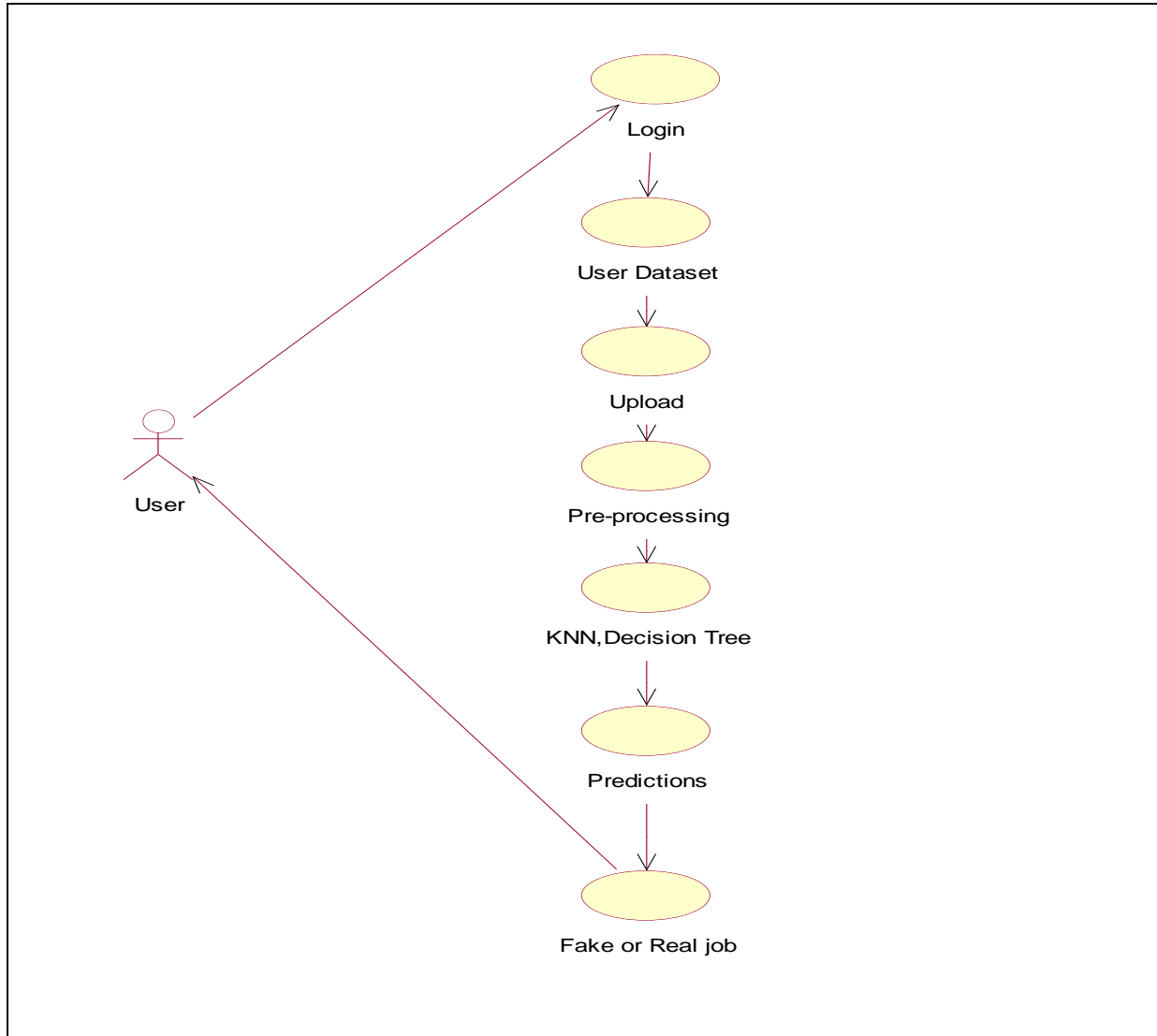


Fig 4.2.1: Usecase Diagram for fake job post prediction using MLDL

EXPLANATION

A use case diagram in the Unified Modeling Language (UML) user has a login. After login it has a user dataset. It can upload a data. It have a preprocessing a data. It will apply a algorithm and it has a predictions of values and then it display output.

4.2.2 CLASS DIAGRAM:

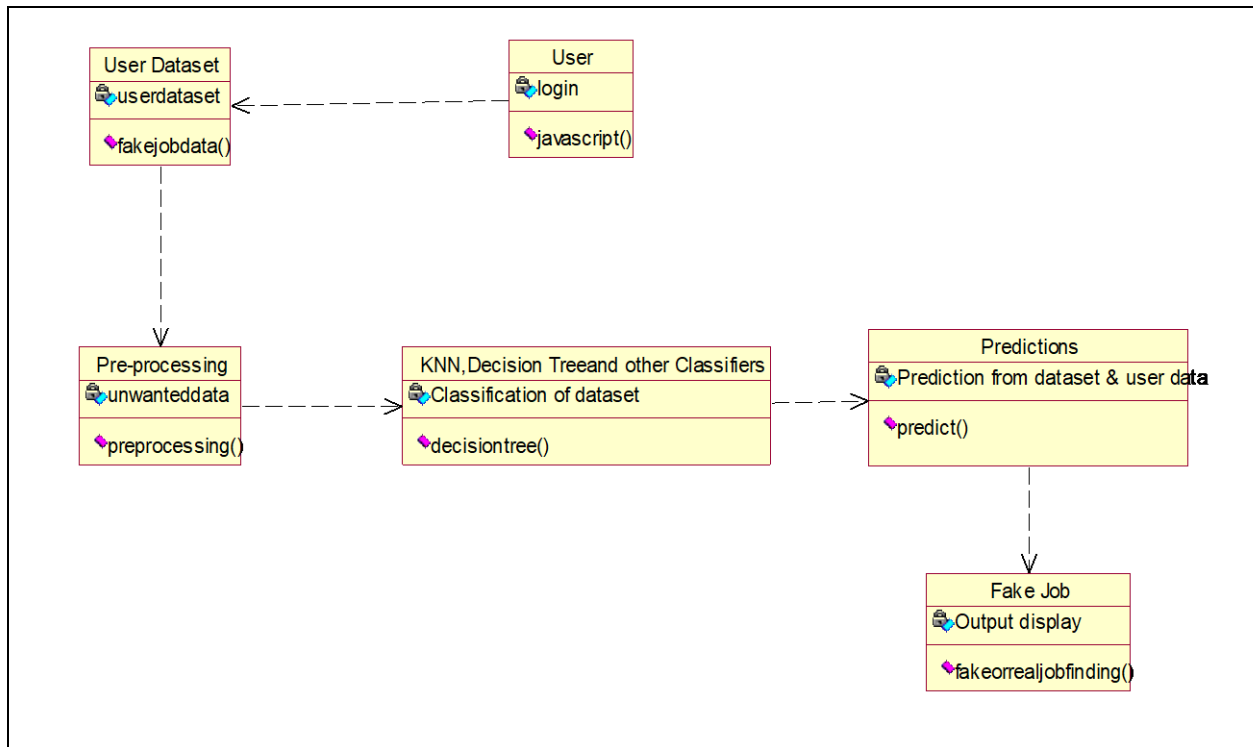


Fig 4.2.2: Class Diagram for fake job post prediction using MLDL

EXPLANATION

In software engineering, a class diagram in the Unified Modelling Language (UML) User has a login with a user id and password. User dataset has a fake job data. It has a pre-processing can remove a unwanted data. It has apply a algorithm KNN, decision tree and other classifier. It has a predictions of data. It detect a fake job.

4.2.3 OBJECT DIAGRAM:

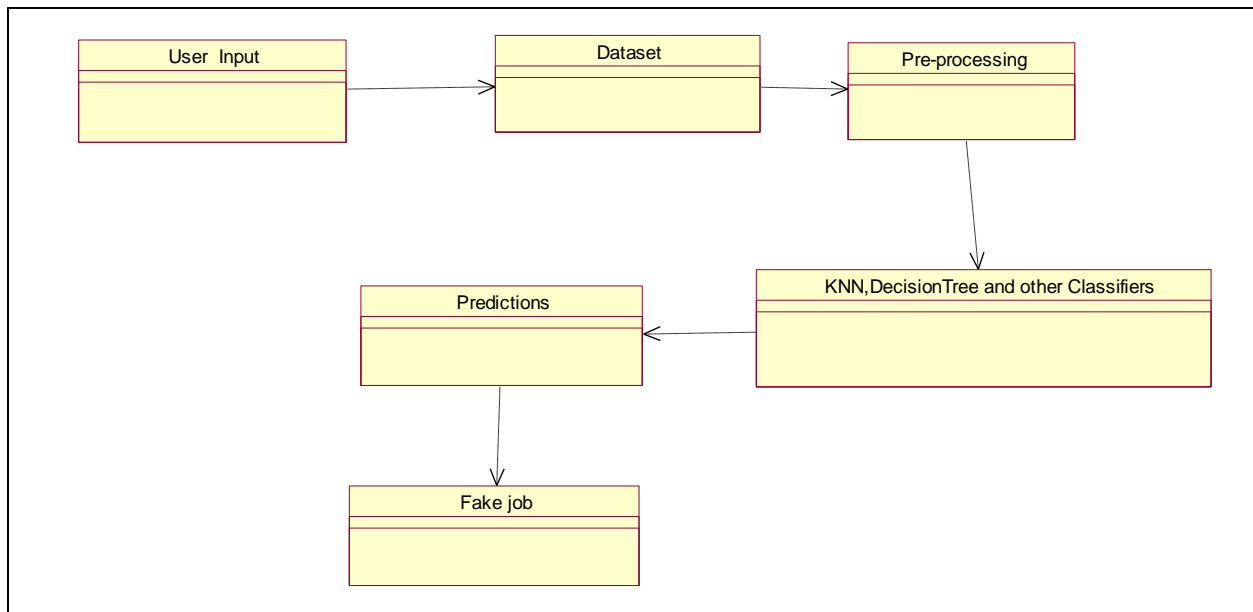


Fig 4.2.3: Object Diagram for fake job post prediction using MLDL

EXPLANATION:

In the above diagram tells about the flow of objects between the classes. It also a User input data to link with a dataset. It also have a preprocessing data to the algorithm. It have a predictions of the values. It also recognize a fake job and it display a output.

4.2.4 SEQUENCE DIAGRAM:

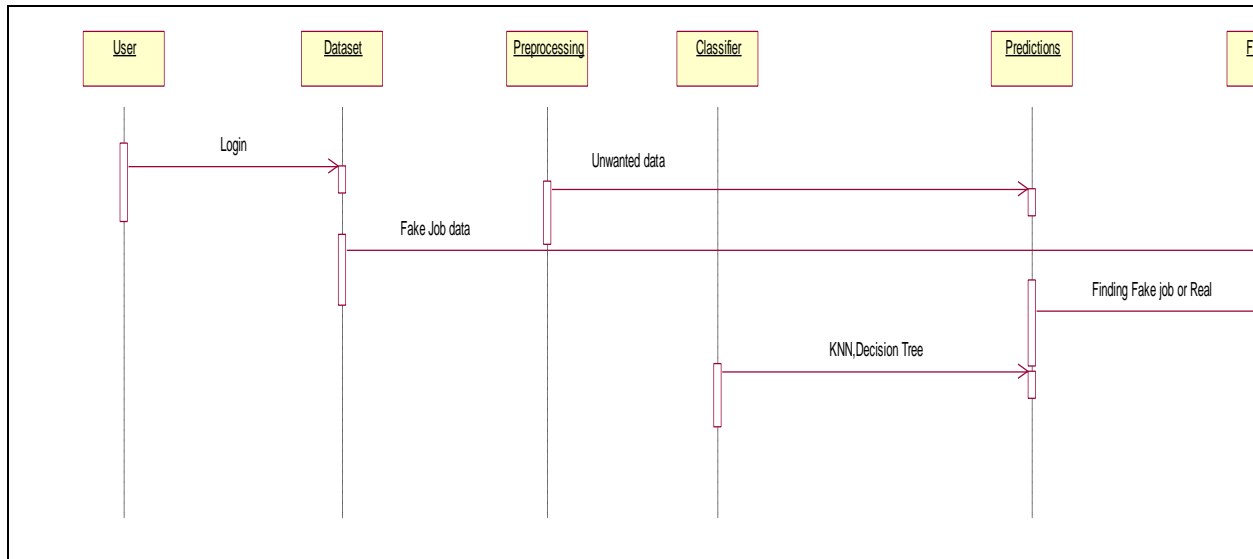


Fig 4.2.4: Sequence Diagram for fake job post prediction using MLDL

EXPLANATION

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It also be a user has link with dataset it sends a message to dataset. From the dataset it has a recognized a fake job. It also have a pre-processing has removes unnecessary data. Classifier has a algorithm performed in KNN, decision trees. It has a predictions of a fake job.

4.2.5 COLLABRATION DIAGRAM:

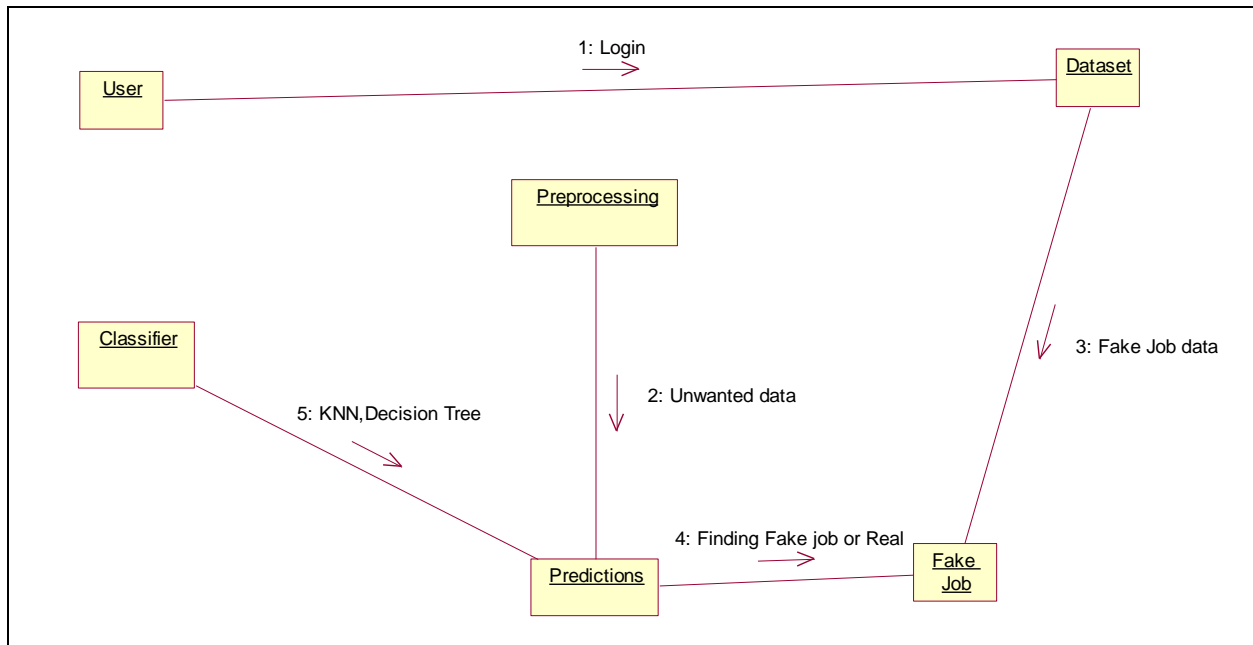


Fig 4.2.5: Collaboration Diagram for fake job post prediction using MLDL

EXPLANATION

A collaboration diagram, also called a communication diagram, shows the interactions between objects. In this diagram, the User has a login with a data. The Dataset has it detects a fake job data. The Classifier it has a apply algorithms in a KNN ,decision trees and links to the predictions. Preprocessing it also have a removes a unwanted data from predictions to predict values.

4.2.6 DEPLOYMENT DIAGRAM:

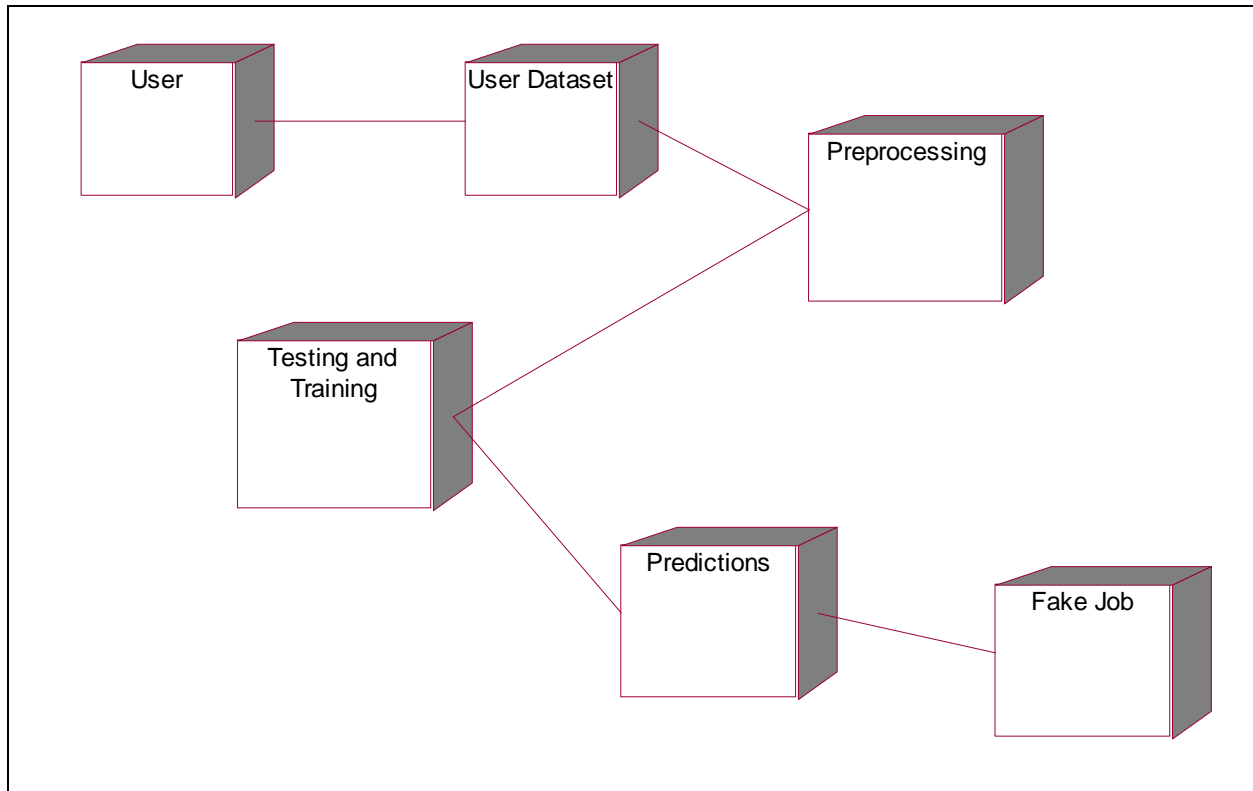


Fig 4.2.6 : Deployment Diagram for fake job post prediction using MLDL

EXPLANATION

Deployment diagrams is a kind of structure diagram used in the user has a link with a user dataset it was also have a pre-processing a data and it has a training and testing data from the dataset. It has a predictions to predict a values and it has a detect a fake job data.

4.2.7 ACTIVITY DIAGRAM:

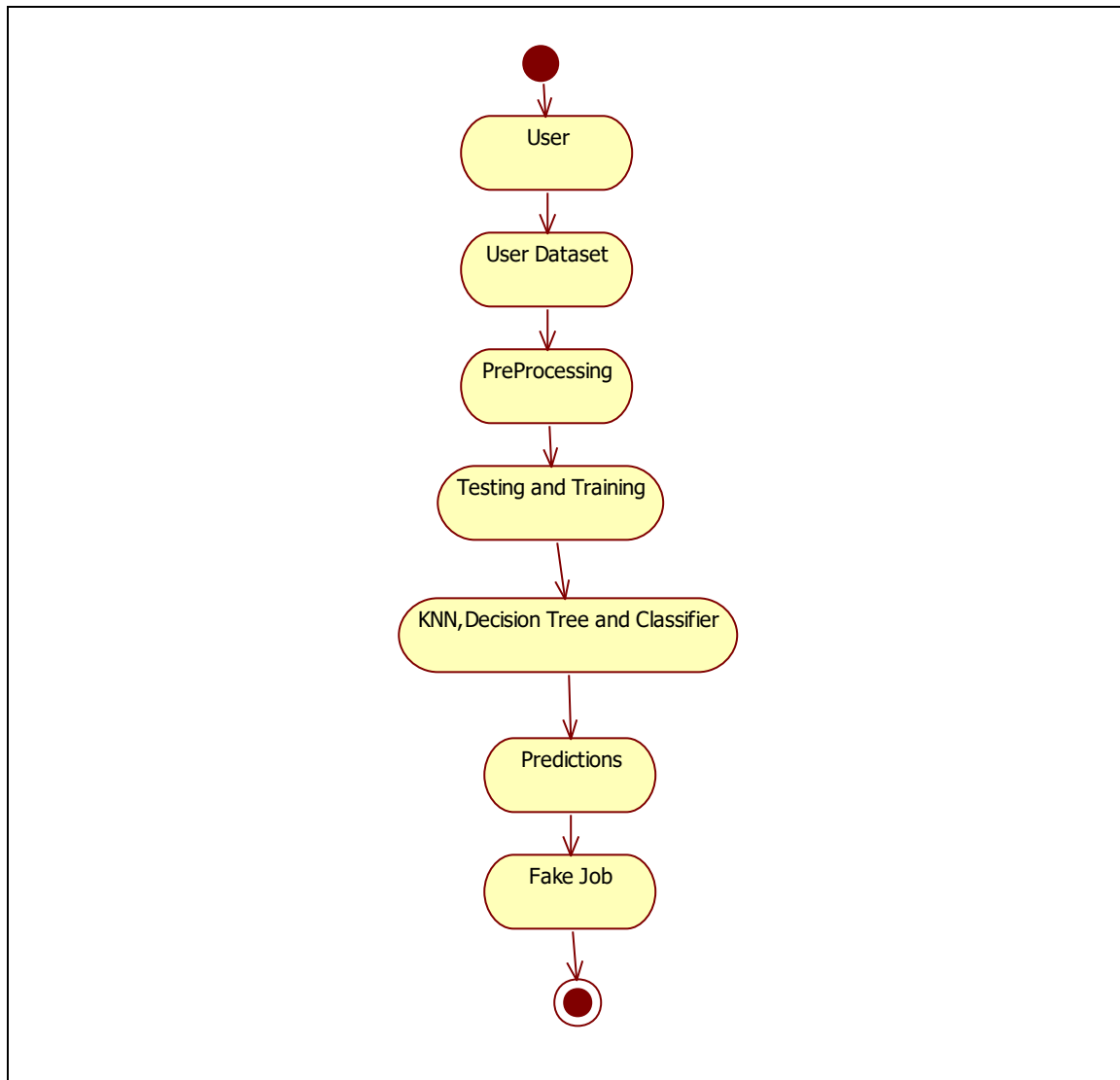


Fig 4.2.7: Deployment Diagram for fake job post prediction using MLDL

EXPLANATION

Activity diagrams are graphical shows a step by step manner operations. It has a user has to login. It was a user dataset. It also have a pre-processing a data. It has a testing and training a data to the dataset. It apply a algorithms in KNN ,decision trees and other. it has also have a predictions to detect a fake job.

4.2.8 STATE DIAGRAM:

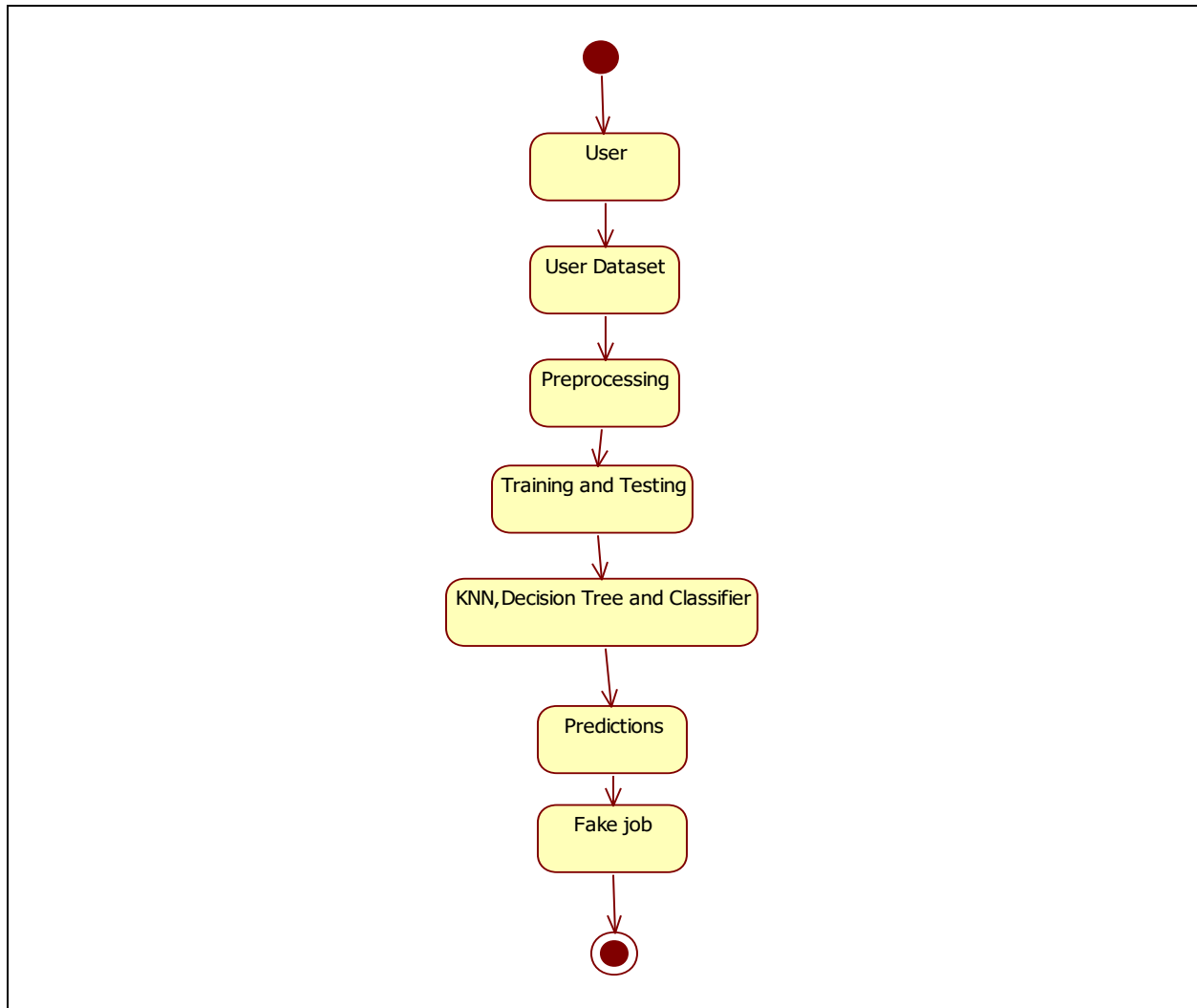


Fig 4.2.8: State Diagram for fake job post prediction using MLDL

EXPLANATION:

State diagram are a loosely defined diagram to show workflows of stepwise activities and actions, with support for choice, iteration and concurrency. State diagrams require that the system it describes a It has a user has to login. It was a user dataset. It also have a pre-processing a data. It has a testing and training a data to the dataset. It apply a algorithms in knn,decision trees and other. it has also have a predictions to detect a fake job.

4.2.9 DATA FLOW DIAGRAM:

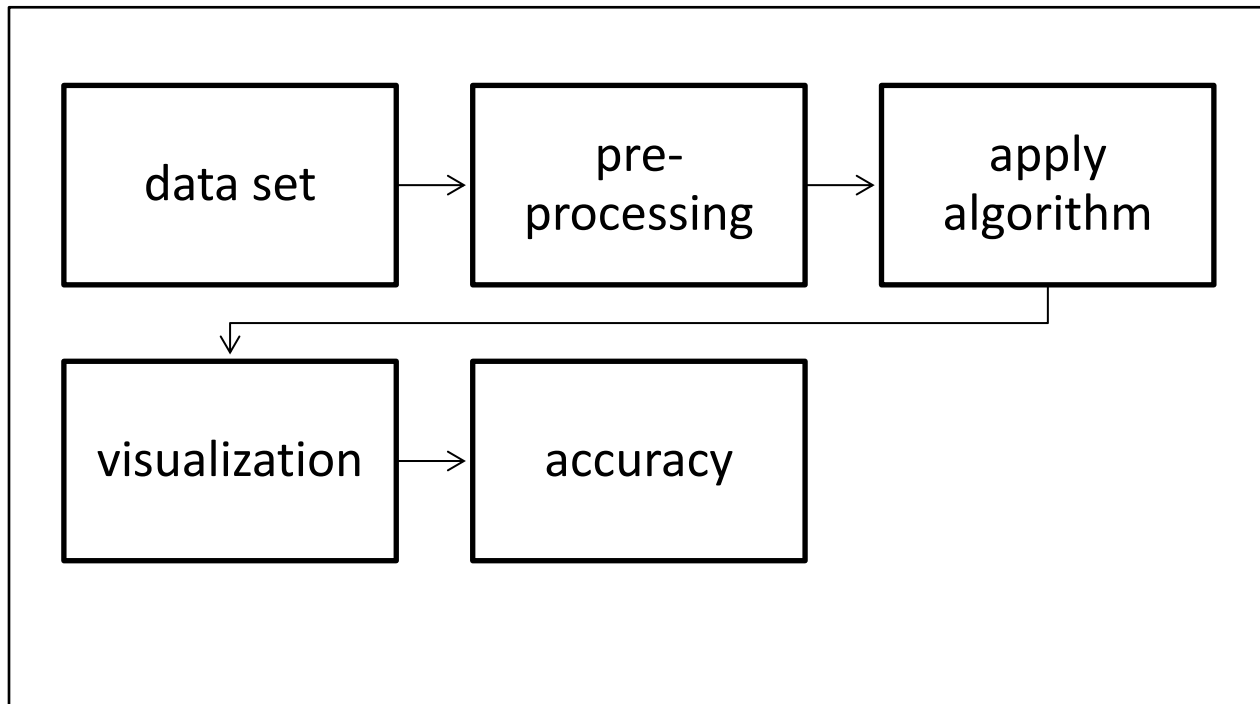


Fig 4.2.9: Data Flow diagram for fake job post prediction using MLDL

EXPLANATION

- The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
- The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components.
- DFD shows how the information moves through the system and how it is modified by a series of transformations. DFD is also known as bubble chart. DFD may be partitioned into levels that represent increasing information flow and functional detail.

4.3 SYSTEMARCHITECTURE:

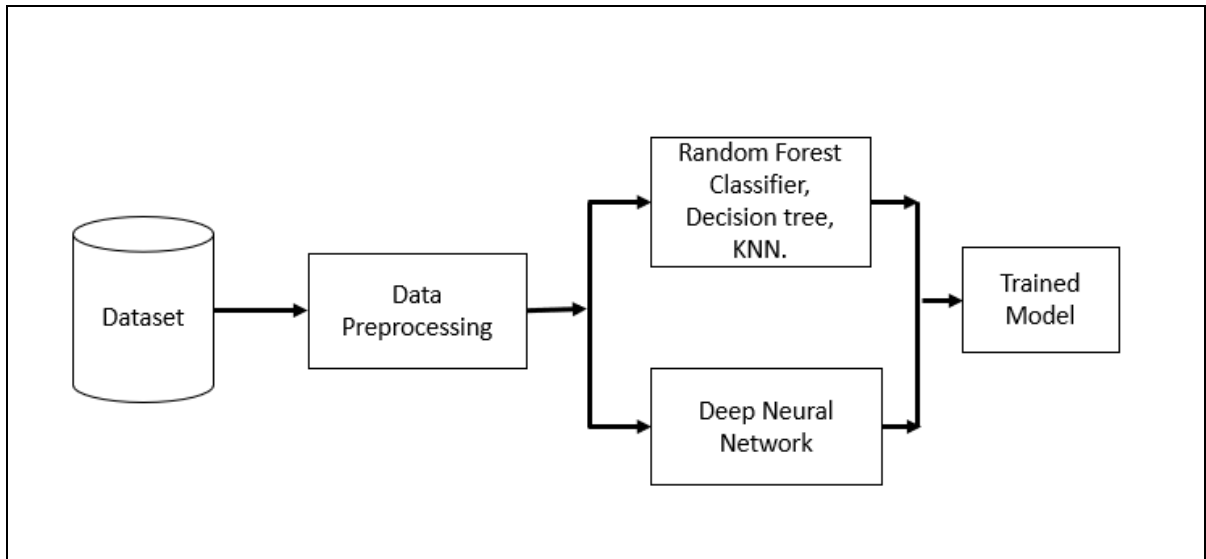


Fig 4.3: System Architecture for fake job post prediction using MLDL

We have used different data mining techniques to predict if a job post is fake or not. We have trained EMSCAD data in the classifiers after a pre-processing step. The trained classifier act as an online fake job post detector. An Artificial Neural Network (ANN) which contains multiple layers between the input and output layer is called Deep Neural Network. DNN works on feed forward algorithm. Data flow is directed from input to output layer . DNN creates a number of virtual neurons initialized with a random numerical value as connection weights. This weight is multiplied with the input and produce an output between 0 and 1.

The training process adjust the weights to classify the output efficiently. Added layers make the model to learn rare patterns which leads the model to overfitting. Dropout layers reduce the number of trainable parameters to make the model generalized. In this paper, we have used a sequential model of dense layers for training the data, relu as activation function and adam as optimizer. During training process, adam calculates individual learning rates on different parameters as this is an adaptive learning Other classifiers K Nearest Neighbor, Random Forest Classifier, Decision Tree, are the classifiers where our work dataset is trained.

CHAPTER 5

DEVELOPMENT TOOLS

5.1 GENERAL

Python

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

History of Python

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk, and Unix shell and other scripting languages.

Python is copyrighted. Like Perl, Python source code is now available under the GNU General Public License (GPL).

Python is now maintained by a core development team at the institute, although Guido van Rossum still holds a vital role in directing its progress.

Importance of Python

- **Python is Interpreted** – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** – You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented** – Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

- **Python is a Beginner's Language** – Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

Features of Python

- **Easy-to-learn** – Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- **Easy-to-read** – Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** – Python's source code is fairly easy-to-maintain.
- **A broad standard library** – Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.
- **Interactive Mode** – Python has support for an interactive mode which allows interactive testing and debugging of snippets of code.
- **Portable** – Python can run on a wide variety of hardware platforms and has the same interface on all platforms.
- **Extendable** – You can add low-level modules to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** – Python provides interfaces to all major commercial databases.
- **GUI Programming** – Python supports GUI applications that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** – Python provides a better structure and support for large programs than shell scripting.

Apart from the above-mentioned features, Python has a big list of good features, few are listed below –

- It supports functional and structured programming methods as well as OOP.
- It can be used as a scripting language or can be compiled to byte-code for building large applications.
- It provides very high-level dynamic data types and supports dynamic type checking.
- IT supports automatic garbage collection.
- It can be easily integrated with C, C++, COM, ActiveX, CORBA, and Java.

Libraries used in python:

- NumPy - mainly useful for its N-dimensional array objects.
- pandas - Python data analysis library, including structures such as data frames.
- matplotlib - 2D plotting library producing publication quality figures.
- scikit-learn - the machine learning algorithms used for data analysis and data mining tasks.

CHAPTER 6

IMPLEMENTATION

6.1 GENERAL

To defend against the proposed attacks and improve the robustness of ML-based NIDSs, we introduce two probable methods in prior work, and then propose a novel defense scheme named adversarial feature reduction.

Adversarial training: This is a promising method widely used to defend against adversarial examples in the image domain by retraining the classifiers with correctly-labeled adversarial examples. However in our traffic-space attack, it can only reduce the attack effectiveness by limiting the generation of adversarial features.

Feature selection: This is an important step in feature engineering to remove redundant/irrelevant dimensions of features used in ML models, which can effectively improve detection performance and robustness.

Adversarial feature reduction: We propose a novel scheme to explain and defend against such traffic-space adversarial attacks. In a nutshell, we proactively simulate the proposed attack and then calculate the degree to which the value of each feature dimension in the mutated traffic is close to the adversarial features compared to original value (see Appendix D for details). The proximity rates of each feature dimension can be viewed as the adversarial robustness scores. Our main claim is the high dimensionality of features gives attackers an opportunity to exploit some vulnerable dimensions to evade detection.

Hence, we propose an intuitive defense scheme by deleting partial feature dimensions with low robustness scores.

CODE:

```
import numpy as np
import pandas as pd
from flask import Flask, request, jsonify, render_template, redirect, flash, send_file
from sklearn.preprocessing import MinMaxScaler
from werkzeug.utils import secure_filename
import pickle
app = Flask(__name__) #Initialize the flask App
model = pickle.load( open('random.pickle', 'rb') )
vecs = pickle.load( open('vectorizers.pickle', 'rb') )
classifiers = pickle.load( open('classifiers.pickle', 'rb') )
@app.route('/')
@app.route('/index')
def index():
    return render_template('index.html')
@app.route('/chart')
def chart():
    return render_template('chart.html')
@app.route('/performance')
def performance():
    return render_template('performance.html')
@app.route('/login')
def login():
    return render_template('login.html')
@app.route('/upload')
def upload():
    return render_template('upload.html')
@app.route('/preview',methods=["POST"])
def preview():
    if request.method == 'POST':
        dataset = request.files['datasetfile']
        df = pd.read_csv(dataset,encoding = 'unicode_escape')
        df.set_index('Id', inplace=True)
        return render_template("preview.html",df_view = df)
    @app.route('/fake_prediction')
def fake_prediction():
    return render_template('fake_prediction.html')
@app.route('/predict',methods=['POST'])
def predict():
    features = [float(x) for x in request.form.values()]
    final_features = [np.array(features)]
    y_pred = model.predict(final_features)
    if y_pred[0] == 1:
        label="Fake Job Post"
    elif y_pred[0] == 0:
```

```

        label="Legit Job Post"
    return render_template('fake_prediction.html', prediction_texts=label)
@app.route('/text_prediction')
def text_prediction():
    return render_template("text_prediction.html")
@app.route('/job')
def job():
    abc = request.args.get('news')
    input_data = [abc.rstrip()]
    # transforming input
    tfidf_test = vecs.transform(input_data)
    # predicting the input
    y_preds = classifiers.predict(tfidf_test)
    if y_preds[0] == 1:
        labels="Fake Job Post"
    elif y_preds[0] == 0:
        labels="Legit Job Post"
    return render_template('text_prediction.html', prediction_text=labels)
if __name__ == "__main__":
    app.run()

```

CHAPTER 7

SNAPSHOTS

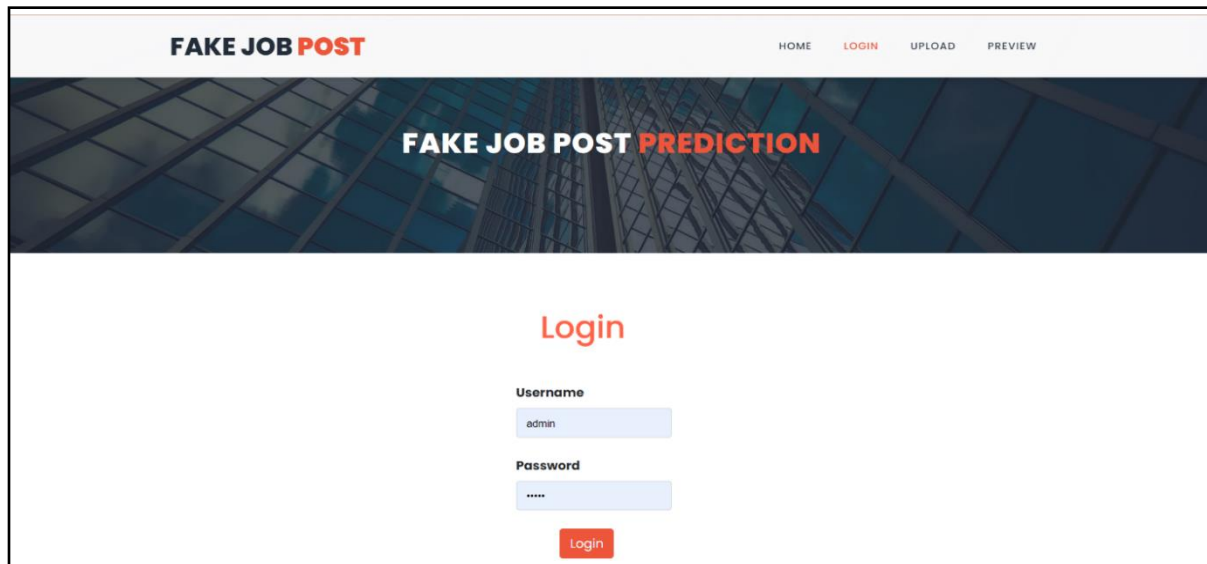
7.1 GENERAL

This project implements like application using python and the Server process is maintained using the SOCKET & SERVERSOCKET and the Design part is played by Cascading Style Sheet.

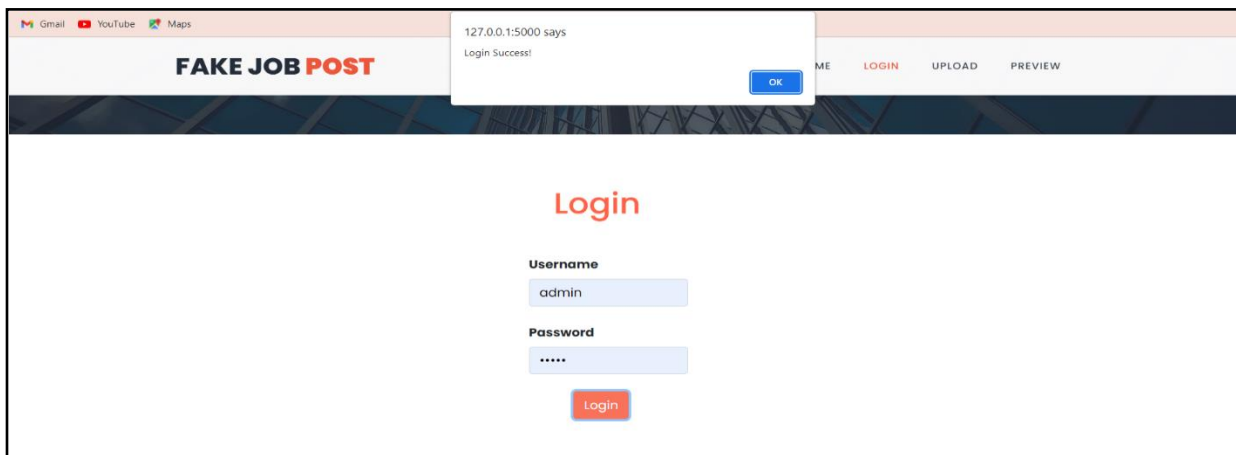
7.2 SNAPSHOTS



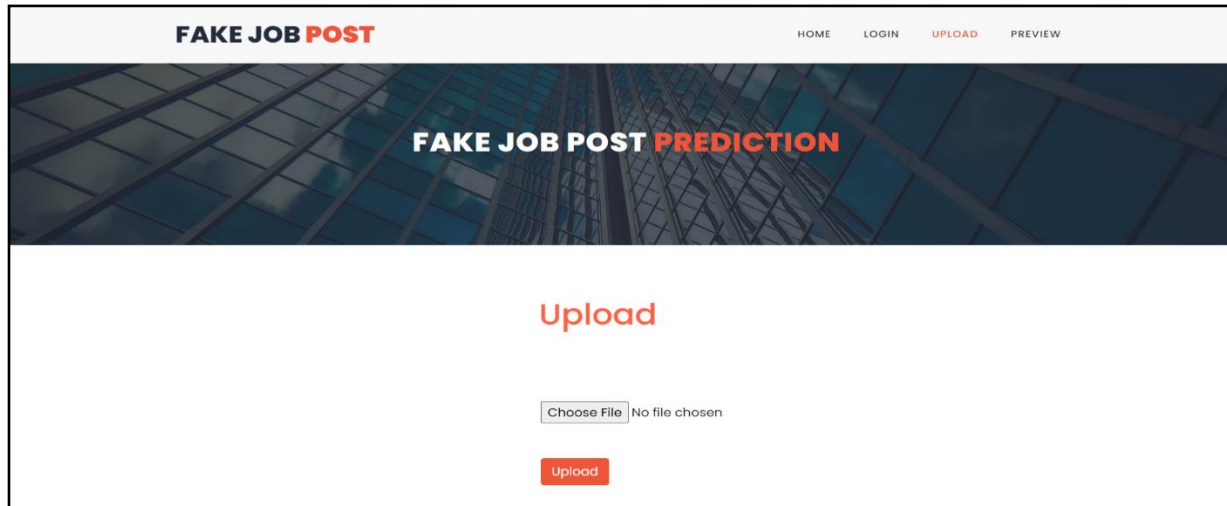
The above snapshot contains interface of the app.



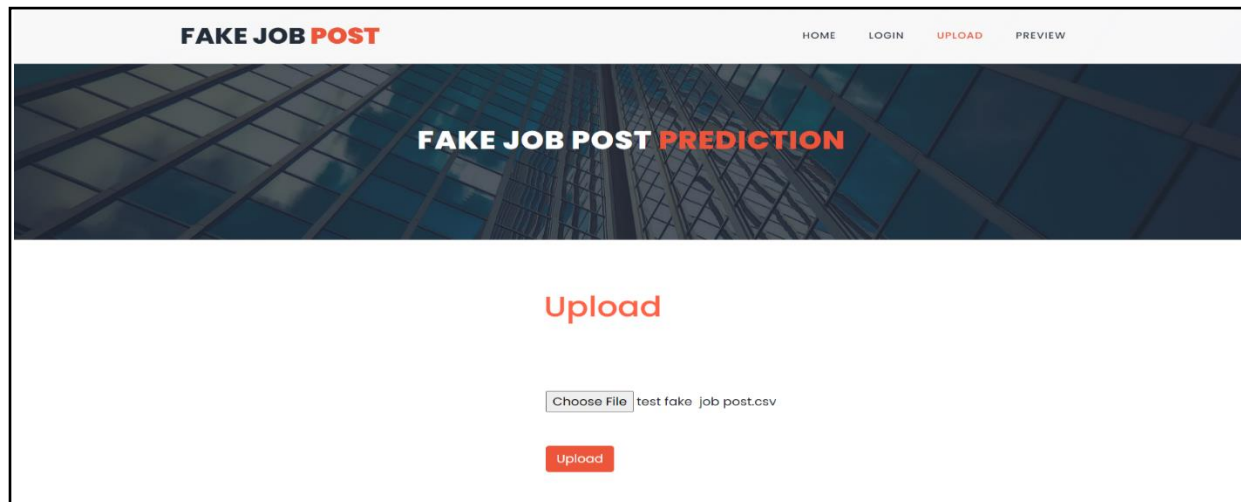
when login button is clicked it displays the login page as shown above.



The login page allows the user to enter the username and password as admin so by entering the details we can login successfully.



After login we will be navigated for uploading of our data file known as upload.csv.



Click on upload button.

FAKE JOB POST
[HOME](#)
[LOGIN](#)
[UPLOAD](#)
[PREVIEW](#)

Preview

	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	function	fraudulent
Id								
1	No	Yes	No	Other	Internship	Bachelor's Degree	Marketing	Legit Job
2	No	Yes	No	Full-time	Not Applicable	Bachelor's Degree	Customer Service	Legit Job Post
3	No	Yes	Yes	Full-time	Mid-Senior level	High School or equivalent	Other	Fake Job Post
4	No	No	No	Not Mentioned employment_type	Mid-Senior level	Bachelor's Degree	Information Technology	Fake Job Post
5	No	Yes	Yes	Full-time	Mid-Senior level	Bachelor's Degree	Health Care Provider	Legit Job Post
6	No	Yes	Yes	Contract	Mid-Senior level	Bachelor's Degree	Information Technology	Legit Job Post
7	Yes	No	No	Not Mentioned employment_type	Mid-Senior level	Bachelor's Degree	Information Technology	Fake Job Post
8	No	No	Yes	Not Mentioned employment_type	Mid-Senior level	Bachelor's Degree	Information Technology	Fake Job Post

[Click to Train | Test](#)

Then it displays the attributes in csv file. The attributes are very useful in predicting whether the job post id legit or fake and gives a detail description about why the job post is fake or legit.

The screenshot shows the FAKE JOB POST web application. At the top, there's a navigation bar with links for HOME, LOGIN, UPLOAD, and PREVIEW. A modal window displays the message "localhost:5000 says Training finished!" with an OK button. Below the modal is a table with 9 columns and 3 rows of data. At the bottom, there's a button labeled "Click to Train | Test".

6	No	Yes	Yes	Contract	Mid-Senior level	Bachelor's Degree	Information Technology	Legit Job Post
7	Yes	No	No	Not Mentioned employment_type	Mid-Senior level	Bachelor's Degree	Information Technology	Fake Job Post
8	No	No	Yes	Not Mentioned employment_type	Mid-Senior level	Bachelor's Degree	Information Technology	Fake Job Post

The dataset contain attributes like telecommuting, has_company_logo, has_questions, employment_type,...etc. We can click on click to train test button to start prediction and we can predict whether the job post is fake or legit.

The screenshot shows the FAKE JOB POST web application with the title "Fake Job Post Prediction". Below the title is a section labeled "Enter The Details" with several input fields for job post attributes. A submit button is located at the bottom.

FAKE JOB POST HOME LOGIN UPLOAD **FRAUDULENT JOB POST** TEXT PROCESSING

Fake Job Post Prediction

Enter The Details

Telecommuting: NO

Has_company_logo: NO

Has_questions: NO

Employment_type: Full-time

Required_experience: Mid-Senior level

Required_education: Master's Degree

Function: Marketing

submit

We can enter the details manually from the dataset and start prediction by clicking submit button. The details contain the attributes which are present in the data file or data set.

FAKE JOB POST HOME LOGIN UPLOAD **FRAUDULENT JOB POST** TEXT PROCESSING

Enter The Details

Telecommuting: NO

Has_company_logo: NO

Has_questions: NO

Employment_type: Full-time

Required_experience: Mid-Senior level

Required_education: Master's Degree

Function: Marketing

submit

prediction is:
Legit Job Post

The prediction is can be legit job post or fake job post according to the our parameters values.

FAKE JOB POST HOME LOGIN **UPLOAD** PREVIEW

FAKE JOB POST PREDICTION

Upload

Choose File test text_prediction.csv

Upload

Next click on upload button on the top of home page. After we will be navigated for uploading of our data file known as test text_prediction.csv. Click on the upload button.

FAKE JOB POST

HOME LOGIN UPLOAD PREVIEW

Preview

	description	fraudulent	Unnamed
Id			
1	Organised - Focused - Vibrant - Awesome!Do you have a passion for customer service? Slick typing skills? Maybe Account Management? ...And think administration is cooler than a polar bear on a jetski? Then we need to hear you!We are the Cloud Video Production Service and operating on a global level. Yeah, it's pretty cool. Serious about delivering a world class product and excellent customer service.Our rapidly expanding business is looking for a talented Project Manager to manage the successful delivery of video projects, manage client communications and drive the production process. Work with some of the coolest brands on the planet and learn from a global team that are representing NZ is a huge way!We are entering the next growth stage of our business and growing quickly internationally. Therefore, the position is bursting with opportunity for the right person entering the business at the right time.90 Seconds, the worlds Cloud Video Production Service - http://90#URL_fbe6559afac620a3cd2c22281f7b8d0eef56a73e3d9a311e2f1ca13d081dd630#90 Seconds is the worlds Cloud Video Production Service enabling brands and agencies to get high	-	Legit Job Post

FAKE JOB POST

HOME LOGIN UPLOAD PREVIEW

	test activities in line with work instructions and Standard Operating Procedures, under direct guidance and supervision. Ensure all calibrated equipment used within the build, test process is within calibration date and rated for the work activities. Maintain workshop plant and equipment including test equipment. Ensure work area is maintained in a safe and tidy manner. Play a pro-active role in housekeeping and continuous improvement initiatives.		
5	We are seeking a seasoned Temporary Accounts Receivable professional with 3-5 years of experience working with high end dept stores. The applicant will support the A/R coordinator with the workload as it applies to Accounts Receivable. The ideal candidate will be able multi-task with duties involving but not limited to the below primary responsibilities:Cash PostingHandle large volume of chargebacks; dispute and recoveryClosing out RMAs and issuing creditsCredit issuance for approved deductionsAssistance with sending statements, invoices and following up with collections of past duesOther duties as assignedSecondary responsibilitiesAssist with Month end duties such as closingAssist with monitoring our factored client aging and Private clients	Legit Job Post	-
6	Optometric practice is seeking a full-time Optical / Sales for our Colorado Springs, Colorado location. To apply for this position, please submit your application via this link: #URL_dc6a4e8df8c88cf7bb611c27adf835b2ea6d40cec837463b39bb6ba9bca8852#?i=MTkz and select the Optical / Sales (Colorado Springs, Colorado) position from the Job Opening drop-down menu.For more information about our company, please visit our web site at . We are an equal opportunity employer.	Fake Job Post	-

Click to Train | Test

FAKE JOB POST

localhost:5000 says
Training finished!

HOME LOGIN UPLOAD PREVIEW

	ideal candidate will be able multi-task with duties involving but not limited to the below primary responsibilities:Cash PostingHandle large volume of chargebacks; dispute and recoveryClosing out RMAs and issuing creditsCredit issuance for approved deductionsAssistance with sending statements, invoices and following up with collections of past duesOther duties as assignedSecondary responsibilitiesAssist with Month end duties such as closingAssist with monitoring our factored client aging and Private clients		
6	Optometric practice is seeking a full-time Optical / Sales for our Colorado Springs, Colorado location. To apply for this position, please submit your application via this link: #URL_dc6a4e8df8c88cf7bb611c27adf835b2ea6d40cec837463b39bb6ba9bca8852#?i=MTkz and select the Optical / Sales (Colorado Springs, Colorado) position from the Job Opening drop-down menu.For more information about our company, please visit our web site at . We are an equal opportunity employer.	Fake Job Post	-

Click to Train | Test

Then it displays the job post articles with some attributes like job description, fraudulent, unnamed,...etc. Click on the train/test button.

FAKE JOB POST

HOMELOGINUPLOADFRAUDULENT JOB POSTTEXT PROCESSINGCHART

Fake Job Post
Text Prediction

Optometric practice is seeking a full-time Optical / Sales for our Colorado Springs, Colorado [location A](#). To apply for this position, please submit your application via this [link A](#). #URL_dc6a4e8df8c88cf7bb61lc27fadf835b2ea6d40cec837463b39bb6ba9bca8852#?i=MTkzA and select the Optical / Sales (Colorado Springs, Colorado) position from the Job Opening drop-down [menu](#). For more information about our company, please visit our web site [at A](#). We are an equal opportunity employer.

Predict

FAKE JOB POST

HOMELOGINUPLOADFRAUDULENT JOB POSTTEXT PROCESSINGCHART

Fake Job Post
Text Prediction

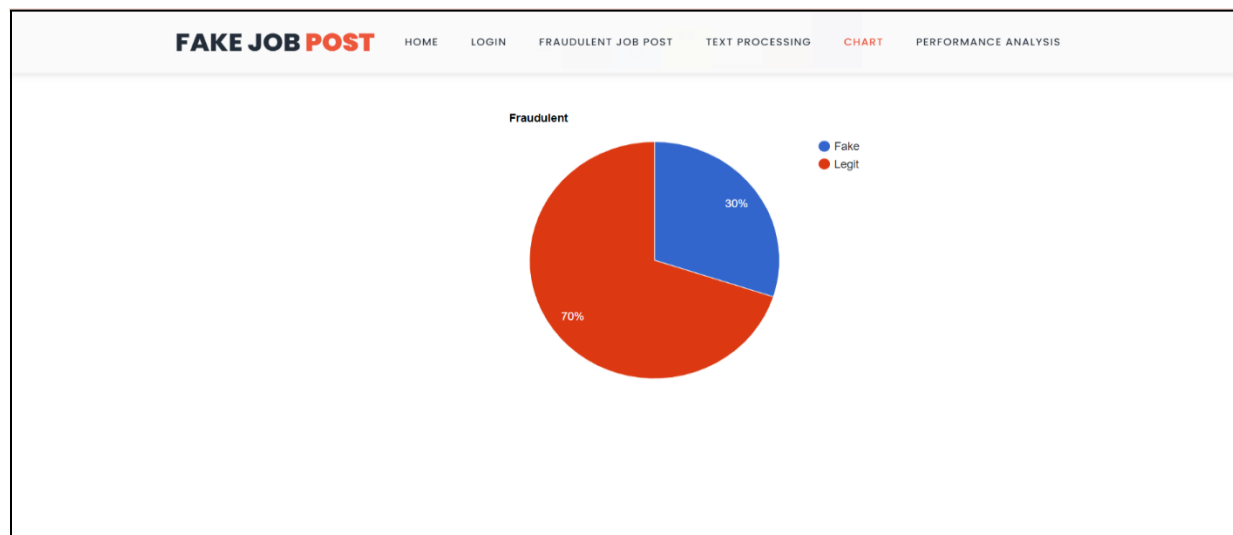
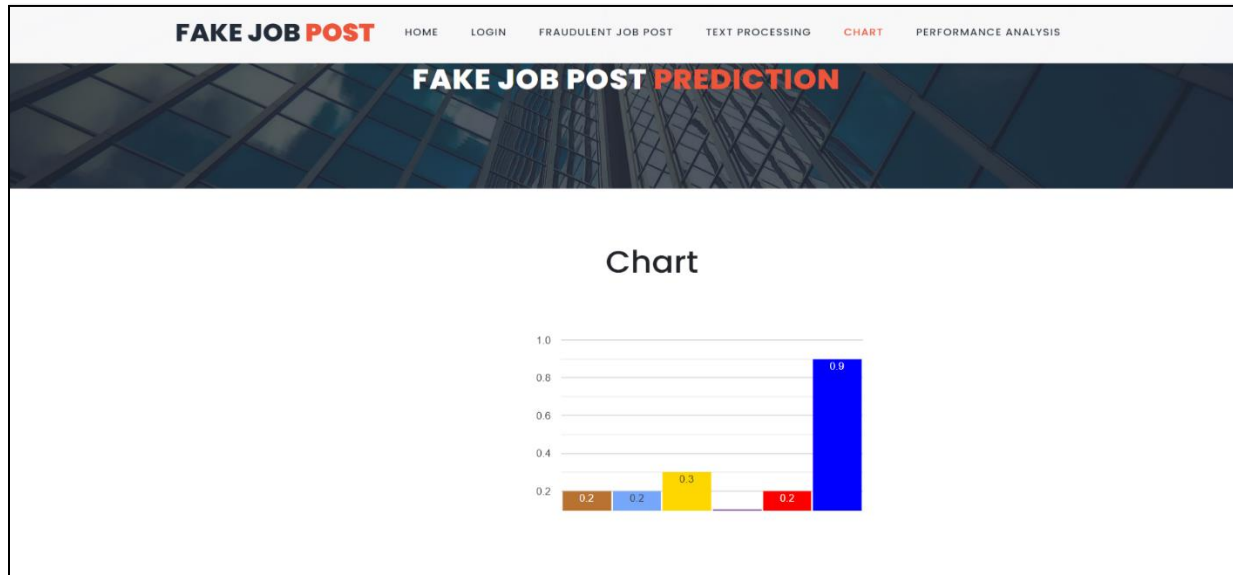
Enter The Details!

Predict

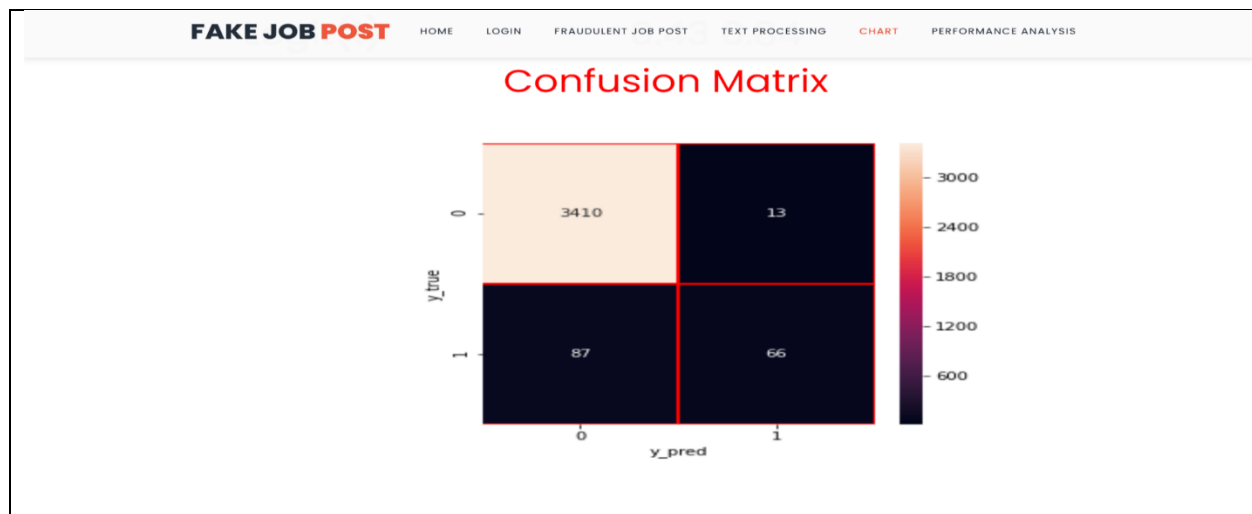
prediction is:

Fake Job Post

Next a text field will be displayed, copy & paste any one of the job post article from the test_prediction.csv data file in the text field and click on the predict button. Then it display whether the job post article is fake job post or legit job post.



After that click on Chart button on the top of the page, then a flow chart & pie chart will be displayed according to the prediction on how much fake or legit the job post is.



FAKE JOB POST HOME LOGIN FRAUDULENT JOB POST TEXT PROCESSING CHART PERFORMANCE ANALYSIS

PERFORMANCE ANALYSIS

Precision and recall

	Recall	Precision
Fake(0)	0.98	1.00
Legit(1)	0.43	0.84

After that click on the Performance Analysis, then confusion matrix & performance analysis will be displayed.

Performance Analysis is a factor which contain precision and recall which helps in accuracy of the result.

CHAPTER 8

SOFTWARE TESTING

8.1 GENERAL

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

8.2 DEVELOPING METHODOLOGIES

The test process is initiated by developing a comprehensive plan to test the general functionality and special features on a variety of platform combinations. Strict quality control procedures are used.

The process verifies that the application meets the requirements specified in the system requirements document and is bug free. The following are the considerations used to develop the framework from developing the testing methodologies.

8.3Types of Tests

8.3.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program input produce valid outputs. All decision branches and internal code flow should be validated.

It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific

business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

8.3.2 Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals. Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.
Invalid Input : identified classes of invalid input must be rejected.
Functions : identified functions must be exercised.
Output : identified classes of application outputs must be exercised.
Systems/Procedures: interfacing systems or procedures must be invoked.

8.3.3 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration-oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

8.3.4 Performance Test

The Performance test ensures that the output be produced within the time limits, and the time taken by the system for compiling, giving response to the users and request being send to the system for to retrieve the results.

8.3.5 Integration Testing

integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects.

The task of the integration test is to check that components or software applications, e.g., components in a software system or – one step up – software applications at the company level – interact without error.

8.3.6 Acceptance Testing

User Acceptance Testing is a critical phase of any project and requires significant participation by the end user. It also ensures that the system meets the functional requirements.

Acceptance testing for Data Synchronization:

- The Acknowledgements will be received by the Sender Node after the Packets are received by the Destination Node
- The Route add operation is done only when there is a Route request in need
- The Status of Nodes information is done automatically in the Cache Updation process

8.2.7 Build the test plan

Any project can be divided into units that can be further performed for detailed processing. Then a testing strategy for each of this unit is carried out. Unit testing helps to identify the possible bugs in the individual component, so the component that has bugs can be identified and can be rectified from errors.

CHAPTER 9

FUTURE ENHANCEMENT

9.1 FUTURE ENHANCEMENT

In this project, we analysed the possible aspects of employment scam, an unexplored up to now research field that calls for further investigation, and we introduced EMSCAD, a publicly available dataset containing both real-life legitimate and fraudulent job ads. As shown, ORF is a relative new field of variable severity that can escalate quickly to extensive scam.

We also experimented with the EMSCAD dataset. Preliminary, yet, detailed results show that text mining in conjunction with metadata can provide a preliminary foundation for job scam detection algorithms. We strongly believe that the provided dataset can be used as a part of an automated anti-scam solution by ATS to train classifiers or gain deeper knowledge to the characteristics of the problem.

It is also anticipated to trigger and fuel further research efforts to this very interesting, yet still in its infancy area. In future works, we intend to expand EMSCAD and enrich the ruleset by focusing on user behavior, company and network data as well as user-content-IP collision patterns. Moreover, we would like to employ graph modeling and explore connections between fraudulent job ads, companies, and users. Ultimately, our goal is to propose an applicable employment fraud detection tool for commercial purposes.

CHAPTER 10

CONCLUSION & REFERENCES

10.1 CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, KNN, Naive Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99% accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

10.2 REFERENCES

- [1] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, *Future Internet* 2017, 9, 6; doi:10.3390/fi9010006.
- [2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, *Journal of Information Security*, 2019, Vol 10, pp. 155176, <https://doi.org/10.4236/iis.2019.103009>.
- [3] Tin Van Huynh¹, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen¹, and Anh Gia-Tuan Nguyen, “Job Prediction: From Deep Neural Network Models to Applications”, *RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020.
- [4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKE DETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, *IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.
- [5] Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by Violent Extremists”, *Security Informatics*, 3, 5, 2014, <https://doi.org/10.1186/s13388-014-0005-5>
- [6] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv Prepr. arXiv1408.5882*, 2014.
- [7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model,” *arXiv Prepr. arXiv1911.03644*, 2019.
- [8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, “Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification,” *Neurocomputing*, vol. 174, pp. 806814, 2016.
- [9] C. Li, G. Zhan, and Z. Li, “News Text Classification Based on Improved BiLSTM-CNN,” in *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 2018, pp. 890-893.
- [10] K. R. Remya and J. S. Ramya, “Using weighted majority voting classifier combination for relation classification in biomedical texts,” *International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, 2014, pp. 1205-1209.

- [11] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security & Its Applications, 8, 55-72. <https://doi.org/10.5121/imsa.2016.8405>.
- [12] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. "Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.
- [13] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.
- [14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.
- [15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. In Proceedings of the 21st international conference on World Wide Web, Lyon, France, 16-20 April 2012; ACM: New York, NY, USA, 2012; pp. 201-210.
- [16] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174. <https://doi.org/10.1016/j.eij.2014.07.002>.