

A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Ranga Avinash (38110449)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MAY– 2022



SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Ranga Avinash (Reg No: 38110449)** who carried out the project entitled "**A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES**" under my supervision from June 2021 to November 2021.

INTERNAL GUIDE

Dr Prayla Shyry M.E Ph.D.

HEAD OF THE DEPARTMENT

Dr. S VIGNESHWARI, M.E. PhD.,

Dr. L. Lakshmanan M.E. PhD.,

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I Ranga Avinash (Reg No: 38110449) hereby declare that the Project Report entitled **“A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES”** done by us under the guidance of **Dr Prayla Shyry M.E Ph.D.** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2018-2022.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D., Dean**, School of Computing **Dr. S. Vigneshwari M.E., Ph.D.** and **Dr. L. Lakshmanan M.E., Ph.D.**, Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Dr Prayla Shyry M.E., Ph.D.** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

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 paper proposed to use different machine learning techniques and classification algorithm like KNN, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. It was Experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. Use three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
1	INTRODUCTION	8
	1.1. OVERVIEW	8
	1.2. MACHINE LEARNING	9
	1.3 OBJECTIVE	9
2	LITERATURE SURVEY	10
	2.1 RELATED WORK	10
3	METHODOLOGY	15
	3.1 EXISTING SYSTEM	15
	3.1.1 DISADVANTAGES EXISTING SYSTEM	15
	3.2 PROPOSED SYSTEM	15
	3.2.1 ADVATAGES OF PROPOSED SYSTEM	16
	3.3 ALGORITHMS USED	16
	3.3.1 RANDOM FOREST ALGORITHM	16
	3.3.2 KNN ALGORITHM	18
	3.4 HARDWARE REQUIREMENTS	19
	3.5 SOFTWARE REQUIREMENTS	19
	3.6 DIAGRAMS	19
	3.7 MODULES	28
	3.8 SYSTEM ARCHITECTURE	29
	3.9 LANGUAGE USED	33
4	SYSTEM STUDY	40

	4.1 FEASIBILITY STUDY	40
5	CONCLUSION	48
	5.1 CONCLUSION	48
	REFERENCE	49
	APPENDICES	51
	A. SOURCE CODE	51
	B. SCREENSHOTS	55
	C. PLAGIARISM REPORT	74
	D. JOURNAL PAPER	76

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

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 lacks 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 MACHINE LEARNING

Machine learning could be a subfield of computer science (AI). The goal of machine learning typically is to know the structure information of knowledge of information and match that data into models which will be understood and used by folks. Although machine learning could be a field inside technology, it differs from ancient process approaches.

In ancient computing, algorithms are sets of expressly programmed directions employed by computers to calculate or downside solve. Machine learning algorithms instead give computers to coach on knowledge inputs and use applied math analysis so as to output values that fall inside a particular vary. thanks to this, machine learning facilitates computers in building models from sample knowledge so as to modify decision-making processes supported knowledge inputs.

1.3 OBJECTIVE

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.

CHAPTER 2

LITERATURE SURVEY

2.1 REALTED WORK

[2.1] Statistical features-based real-time detection of drifted Twitter spam

AUTHORS: C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min

Twitter spam has become a critical problem nowadays. Recent works focus on applying machine learning techniques for Twitter spam detection, which make use of the statistical features of tweets. In our labeled tweets data set, however, we observe that the statistical properties of spam tweets vary over time, and thus, the performance of existing machine learning-based classifiers decreases. This issue is referred to as “Twitter Spam Drift”. In order to tackle this problem, we first carry out a deep analysis on the statistical features of one million spam tweets and one million non-spam tweets, and then propose a novel Lfun scheme. The proposed scheme can discover “changed” spam tweets from unlabeled tweets and incorporate them into classifier's training process. A number of experiments are performed to evaluate the proposed scheme. The results show that our proposed Lfun scheme can significantly improve the spam detection accuracy in real-world scenarios.

[2.2] Automatically identifying fake news in popular Twitter threads

AUTHORS: C. Buntain and J. Golbeck

Information quality in social media is an increasingly important issue, but web-scale data hinders experts' ability to assess and correct much of the inaccurate content, or "fake news," present in these platforms. This paper develops a method for automating fake news detection on Twitter by learning to predict accuracy assessments in two credibility-focused Twitter datasets: CREDBANK, a crowdsourced dataset of accuracy assessments for events in Twitter, and PHEME, a dataset of potential rumors in Twitter and journalistic assessments of their accuracies. We apply this method to Twitter content sourced from BuzzFeed's fake news dataset and show models trained against crowdsourced workers outperform models based on journalists' assessment and models trained on a pooled dataset of both crowdsourced workers and journalists. All three datasets, aligned into a uniform format, are also publicly available. A feature analysis then identifies features that are most predictive for crowdsourced and journalistic accuracy assessments, results of which are consistent with prior work. We close with a discussion contrasting accuracy and credibility and why models of non-experts outperform models of journalists for fake news detection in Twitter.

[2.3] A performance evaluation of machine learning-based streaming spam tweets detection

AUTHORS: C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian

The popularity of Twitter attracts more and more spammers. Spammers send unwanted tweets to Twitter users to promote websites or services, which are harmful to normal users. In order to stop spammers, researchers have proposed a

number of mechanisms. The focus of recent works is on the application of machine learning techniques into Twitter spam detection. However, tweets are retrieved in a streaming way, and Twitter provides the Streaming API for developers and researchers to access public tweets in real time. There lacks a performance evaluation of existing machine learning-based streaming spam detection methods. In this paper, we bridged the gap by carrying out a performance evaluation, which was from three different aspects of data, feature, and model. A big ground-truth of over 600 million public tweets was created by using a commercial URL-based security tool. For real-time spam detection, we further extracted 12 lightweight features for tweet representation. Spam detection was then transformed to a binary classification problem in the feature space and can be solved by conventional machine learning algorithms. We evaluated the impact of different factors to the spam detection performance, which included spam to nonspam ratio, feature discretization, training data size, data sampling, time-related data, and machine learning algorithms. The results show the streaming spam tweet detection is still a big challenge and a robust detection technique should take into account the three aspects of data, feature, and model.

[2.4] A model-based approach for identifying spammers in social networks

AUTHORS: F. Fathaliani and M. Bouguessa

In this paper, we view the task of identifying spammers in social networks from a mixture modeling perspective, based on which we devise a principled unsupervised approach to detect spammers. In our approach, we first represent each user of the social network with a feature vector that reflects its behaviour and interactions with

other participants. Next, based on the estimated users feature vectors, we propose a statistical framework that uses the Dirichlet distribution in order to identify spammers. The proposed approach is able to automatically discriminate between spammers and legitimate users, while existing unsupervised approaches require human intervention in order to set informal threshold parameters to detect spammers. Furthermore, our approach is general in the sense that it can be applied to different online social sites. To demonstrate the suitability of the proposed method, we conducted experiments on real data extracted from Instagram and Twitter.

[2.5] Spam detection of Twitter traffic: A framework based on random forests and non-uniform feature sampling

AUTHORS: C. Meda, E. Ragusa, C. Gianoglio, R. Zunino, A. Ottaviano, E. Scillia, and R. Surlinelli

Law Enforcement Agencies cover a crucial role in the analysis of open data and need effective techniques to filter troublesome information. In a real scenario, Law Enforcement Agencies analyze Social Networks, i.e. Twitter, monitoring events and profiling accounts. Unfortunately, between the huge amount of internet users, there are people that use microblogs for harassing other people or spreading malicious contents. Users' classification and spammers' identification is a useful technique for relieve Twitter traffic from uninformative content. This work proposes a framework that exploits a non-uniform feature sampling inside a gray box Machine Learning System, using a variant of the Random Forests Algorithm to identify spammers inside Twitter traffic. Experiments are made on a popular

Twitter dataset and on a new dataset of Twitter users. The new provided Twitter dataset is made up of users labeled as spammers or legitimate users, described by 54 features. Experimental results demonstrate the effectiveness of enriched feature sampling method

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

- Tingminet *al.* provide a survey of new methods and techniques to identify Twitter spam detection. The above survey presents a comparative study of the current approaches.
- On the other hand, S. J. Somanet. *al.* conducted a survey on different behaviors exhibited by spammers on Twitter social network. The study also provides a literature review that recognizes the existence of spammers on Twitter social network.
- Despite all the existing studies, there is still a gap in the existing literature. Therefore, to bridge the gap, we review state-of-the-art in the spammer detection and fake user identification on Twitter

3.1.1 DISADVANTAGES OF EXISTING SYSTEM

- Because of Privacy Issues the Facebook dataset is very limited and a lot of details are not made public.
- having less accuracy
- More complex

3.2 PROPOSED SYSTEM

The proposed framework, the sequence of processes that need to be followed for continues detection of fake job post with active learning from the feedback of the result given by the classification algorithm. This framework can easily be implemented by social networking companies. 1. The detection process starts with

the selection of the post that needs to be tested. 2. After the selection of the post, the suitable attributes (i.e. features) are selected on which the classification algorithm is implemented. 3. The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is feed into the classifier. 4. The classifier determines whether the post is fake or genuine. 5. The classifier may not be 100% accurate in classifying the post so; the feedback of the result is given back to the classifier. 6. This process repeats and as the time proceeds, the no. of training data increases and the classifier becomes more and more accurate in predicting the fake job post.

3.2.1 ADVANTAGES OF PROPOSED SYSTEM

- The social networking sites are making our social lives better but nevertheless there are a lot of issues with using these social networking sites.
- The issues are privacy, online bullying, potential for misuse, trolling, etc. These are done mostly by using fake job post.
- In this project, we came up with a framework through which we can detect a fake job post using machine learning algorithms so that the social life of people become secured.

3.3 ALGORITHMS USED

3.3.1 RANDOM FOREST ALGORITHM

Random forest algorithm can use both for classification and the regression kind of problems. In this you are going to learn, how the **random forest algorithm** works in machine learning for the classification task.

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and

Regression problems in ML. It is based on the concept of **ensemble learning**, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The below diagram explains the working of the Random Forest algorithm:

ALGORITHM USED

1. Randomly select "k" features from total "m" features.
 1. Where $k \ll m$
2. Among the "k" features, calculate the node "d" using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1 to 3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for "n" number of times to create "n" number of trees.

The beginning of random forest algorithm starts with randomly selecting “k” features out of total “m” features. In the image, you can observe that we are randomly taking features and observations.

3.3.2 KNN ALGORITHMS

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

3.4 HARDWARE REQUIREMENTS

System : Pentium i3 Processor

Hard Disk : 500 GB.

Monitor : 15'' LED

Input Devices : Keyboard, Mouse

Ram : 2 GB

3.5 SOFTWARE REQUIREMENTS

Operating system : Windows 10

Coding Language : Python

3.6 DIAGRAM

DATA FLOW DIAGRAM

1. 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.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.
3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that

depicts information flow and the transformations that are applied as data moves from input to output.

4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

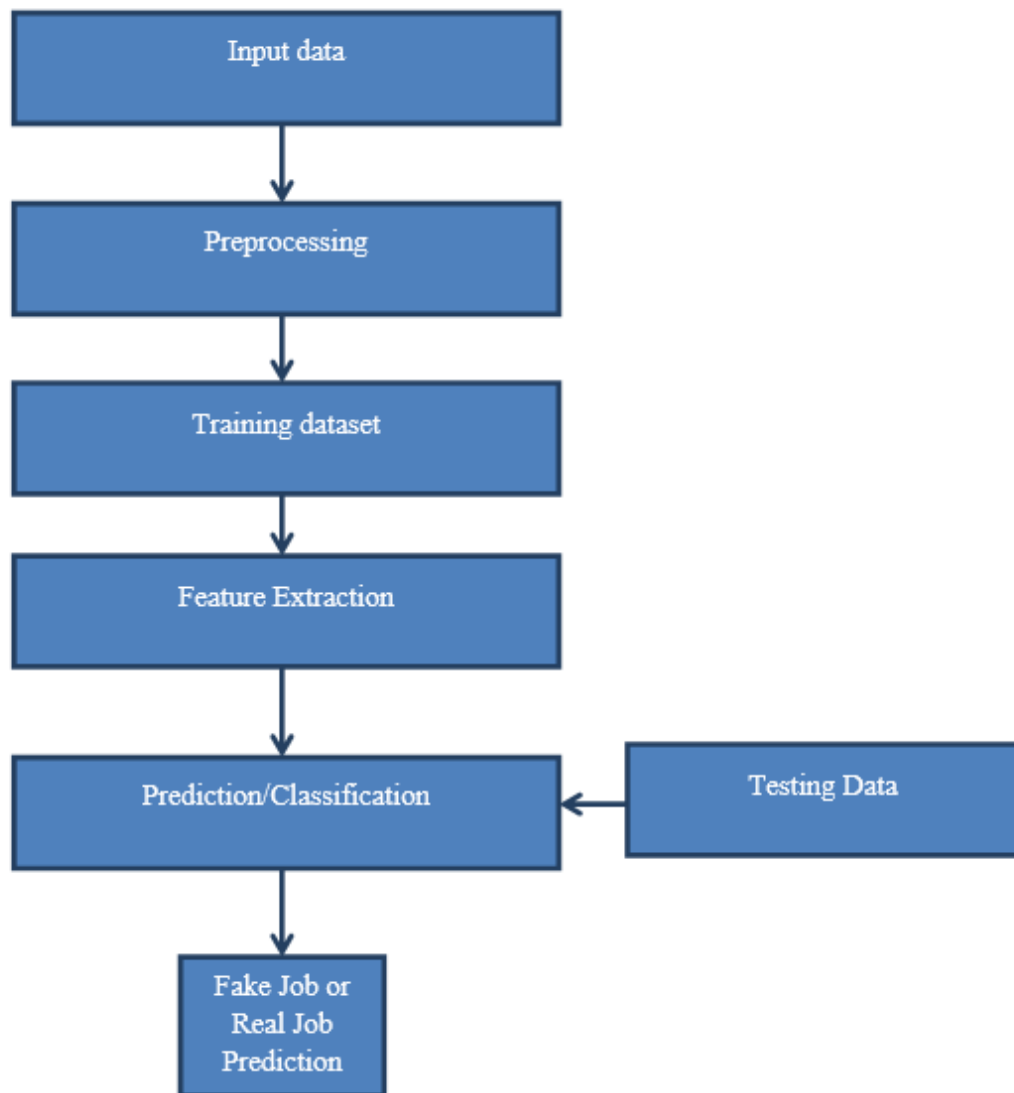


FIG 3.1

UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.

2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.
4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

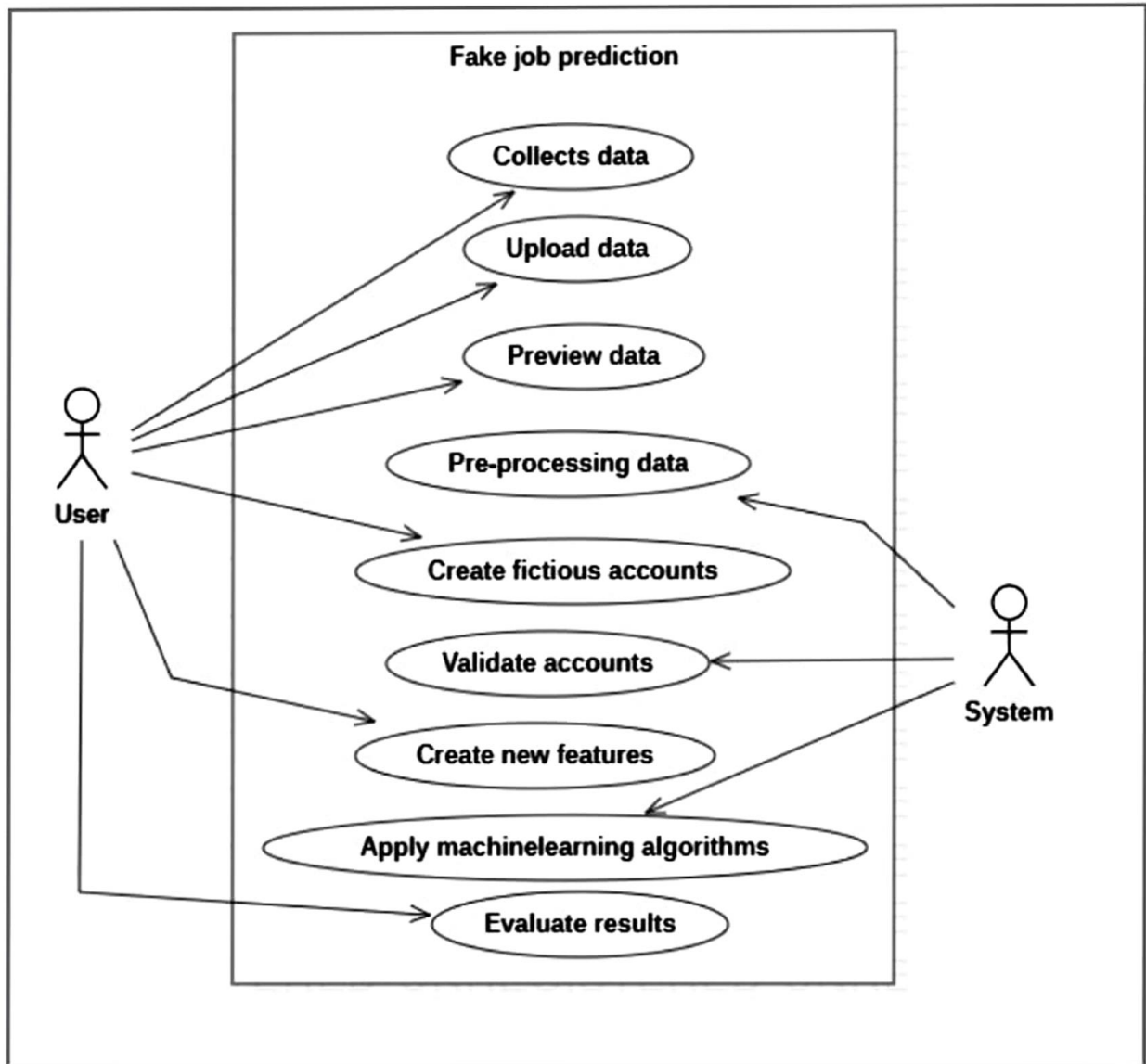


Fig.3.2

ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

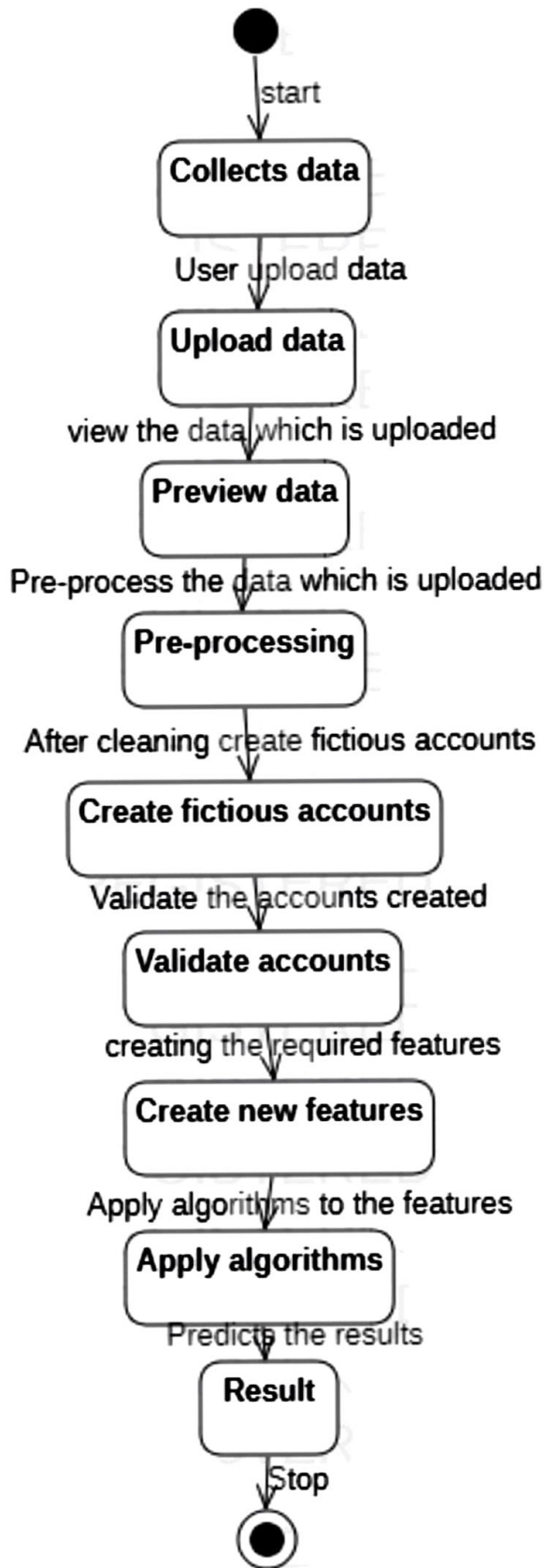


Fig 3.3

SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

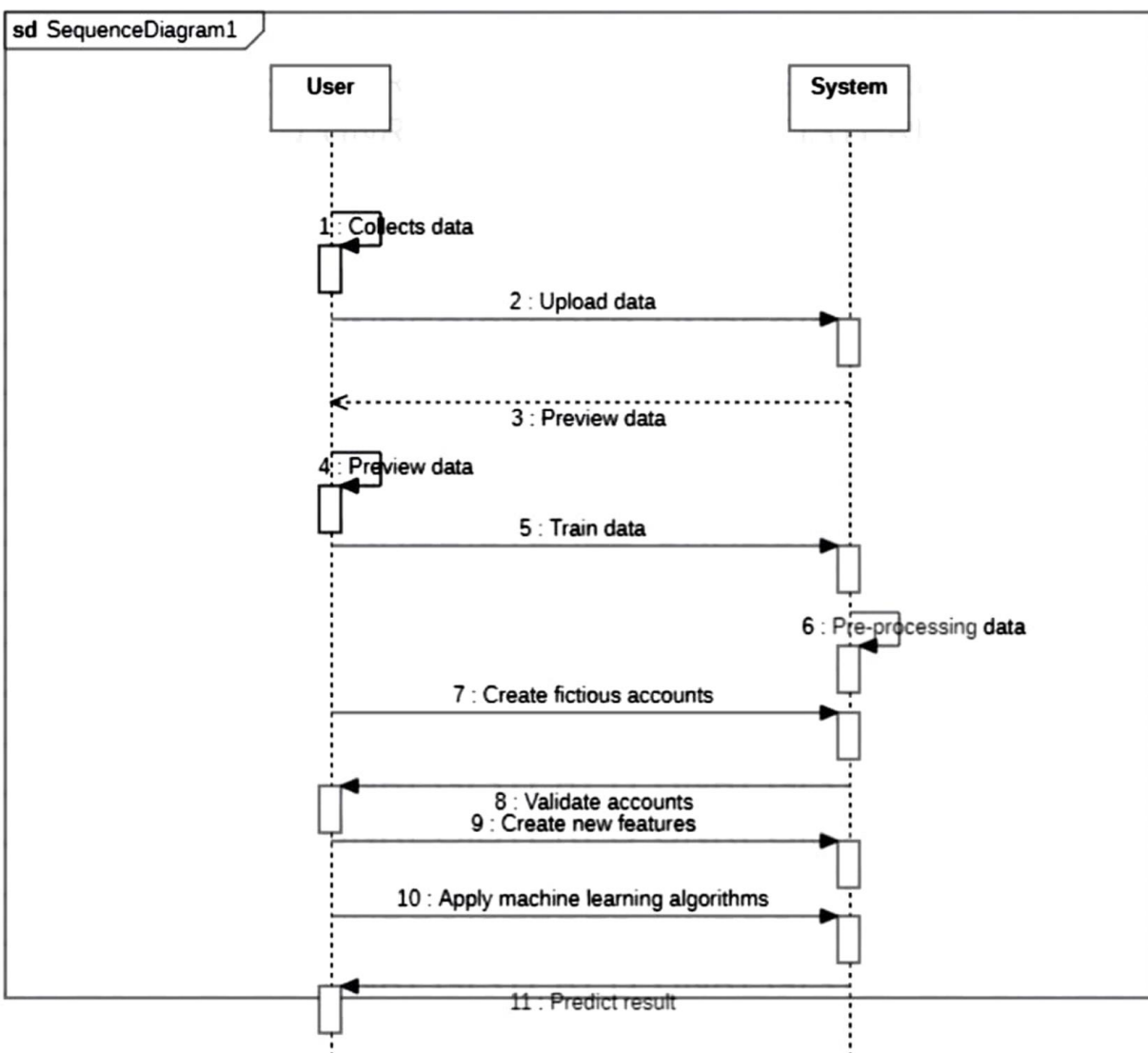


FIG 3.4

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

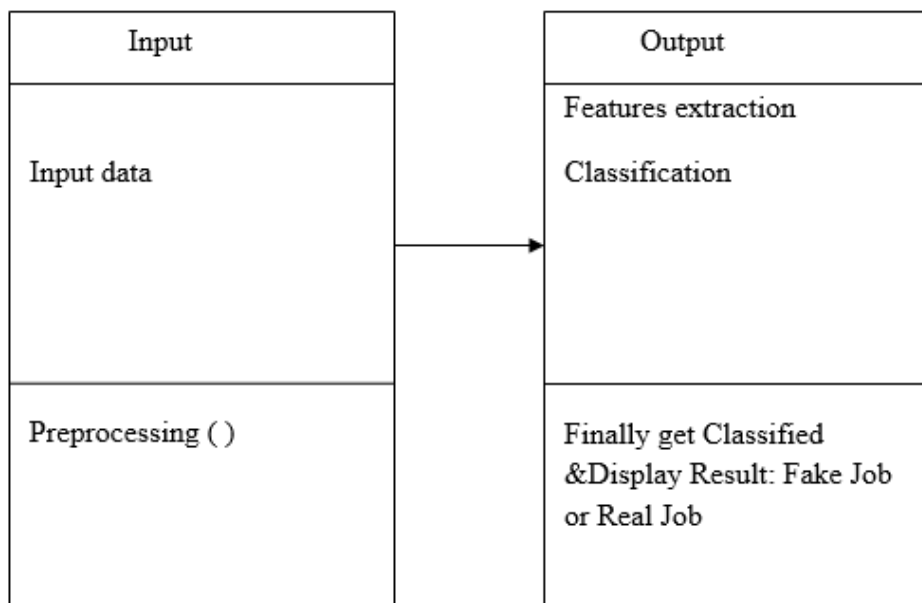


FIG 3.5

3.7 MODULES

- Data Collection
- Pre-Processing
- Train and Test
- Machine Learning Technique
- Detection of Fake Job post

MODULE DESCRIPTION

3.7.1 Data Collection

The dataset was collected from online (www.Kaggle.com) The dataset contains 2180 different samples of job posts. It contains different parameters of job post like Title, Location, department, salary, description, required education etc. Dimensions of dataset is 2180 x 17.

3.7.2 Pre-Processing

We convert data in to scalar format and then create new features which are passed to algorithm and features are saved in x and labels in y.

3.7.3 Train and Test

we will split the dataset into training dataset and test dataset. We will use 70% of our data to train and the rest 30% to test. To do this, we will create a split parameter which will divide the data frame in a 70-30 ratio.

3.7.4 Machine Learning Technique

After splitting the dataset into training and test dataset, we will instantiate Random Forest Classifier and KNN classifier fit the train data by using ‘fit’ function. Then we will store as model.

3.7.5 Fake Job Post Prediction

In this step details are fed as input in the form of csv of various profiles and prediction is performed. New input csv file with different job profile details is given as input and prediction is performed and details are stored in new csv with prediction results.

3.8 SYSTEM ARCHITECTURE

Describing the overall features of the software is concerned with defining the requirements and establishing the high level of the system. During architectural design, the various web pages and their interconnections are identified and designed. The major software components are identified and decomposed into processing modules and conceptual data structures and the interconnections among the modules are identified. The following modules are identified in the proposed system.

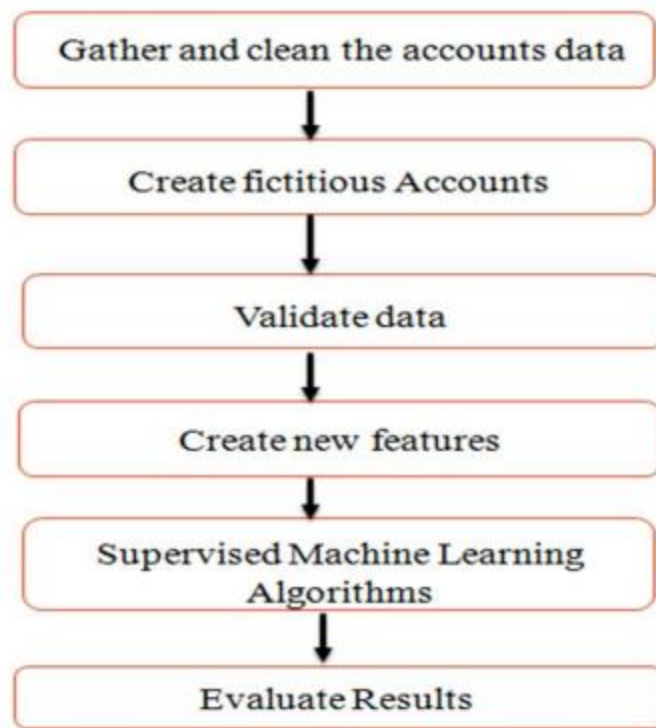


FIG 3.6

The above architecture describes the work structure of the system.

Proposed system is equipped with various Machine Learning tasks and the architecture followed is as shown below. The proposed system collects the dataset which are preprocessed by providing a framework of algorithms using which we can detect fake job post in Facebook by comparing the accuracy of three machine learning algorithms and the algorithm with very high efficiency is found for the given dataset. The different ways in which an algorithm can model a problem is based on its interaction with the experience or environment for the model preparation process that helps in choosing the most appropriate algorithm for the given input data in order to get the best result.

3.8.1 Problem Statement:

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.

INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.
2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.
3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

CHAPTER 4

SYSTEM STUDY

4.1 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ◆ ECONOMICAL FEASIBILITY
- ◆ TECHNICAL FEASIBILITY
- ◆ SOCIAL FEASIBILITY

ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the

research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

SYSTEM TESTING

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.

TYPES OF TESTS

UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs 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.

INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

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.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

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.

WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is used to test areas that cannot be reached from a black box level.

BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box

.you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

6.1 UNIT TESTING:

Unit testing is usually conducted as part of a combined code and unit test phase of the software lifecycle, although it is not uncommon for coding and unit testing to be conducted as two distinct phases.

Test strategy and approach

Field testing will be performed manually and functional tests will be written in detail.

Test objectives

- All field entries must work properly.
- Pages must be activated from the identified link.
- The entry screen, messages and responses must not be delayed.

Features to be tested

- Verify that the entries are of the correct format
- No duplicate entries should be allowed
- All links should take the user to the correct page.

6.2 INTEGRATION TESTING

Software 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.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

6.3 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.

Test Results: All the test cases mentioned above passed successfully. No defects encountered.

CHAPTER-5

CONCLUSION

5.1 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 field 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.

REFERENCES

- [1] C. Chen, S. Wen, J. Zhang, Y. Xiang, J. Oliver, A. Alelaiwi, and M. M. Hassan, “Investigating the deceptive information in Twitter spam,” *Future Gener. Comput. Syst.*, vol. 72, pp. 319–326, Jul. 2017.
- [2] I. David, O. S. Siordia, and D. Moctezuma, “Features combination for the detection of malicious Twitter accounts,” in *Proc. IEEE Int. Autumn Meeting Power, Electron. Comput. (ROPEC)*, Nov. 2016, pp. 1–6.
- [3] M. Babcock, R. A. V. Cox, and S. Kumar, “Diffusion of pro- and anti-false information tweets: The black panther movie case,” *Comput. Math. Org. Theory*, vol. 25, no. 1, pp. 72–84, Mar. 2019.
- [4] S. Keretna, A. Hossny, and D. Creighton, “Recognising user identity in Twitter social networks via text mining,” in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 3079–3082.
- [5] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, “A machine learning approach for Twitter spammers detection,” in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2014, pp. 1–6.
- [6] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, “Real-time Twitter content polluter detection based on direct features,” in *Proc. 2nd Int. Conf. Inf. Sci. Secur. (ICISS)*, Dec. 2015, pp. 1–4.
- [7] H. Shen and X. Liu, “Detecting spammers on Twitter based on content and social interaction,” in *Proc. Int. Conf. Netw. Inf. Syst. Comput.*, pp. 413–417, Jan. 2015.
- [8] G. Jain, M. Sharma, and B. Agarwal, “Spam detection in social media using convolutional and long short term memory neural network,” *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, Jan. 2019.
- [9] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, “A topic-based hidden Markov model for real-time spam tweets filtering,” *Procedia Comput. Sci.*, vol. 112, pp. 833–843, Jan. 2017.

- [10] F. Pierri and S. Ceri, “False news on social media: A data-driven survey,” 2019, arXiv:1902.07539. [Online]. Available: <https://arxiv.org/abs/1902.07539>
- [11] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, “AAFA: Associative affinity factor analysis for bot detection and stance classification in Twitter,” in Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI), Aug. 2017, pp. 356–365.
- [12] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, “Segregating spammers and unsolicited bloggers from genuine experts on Twitter,” IEEE Trans. Dependable Secure Comput., vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.

APPENDICES

A. Source 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(debug=True)
```

B. SCREEN SHOTS



Fig 5.1

- It is the home page of our website.

The image shows a web application interface for 'FAKE JOB POST PREDICTION'. At the top, there is a navigation bar with the title 'FAKE JOB POST' on the left and links for 'HOME', 'LOGIN', 'UPLOAD', and 'PREVIEW' on the right. Below the navigation bar is a large banner image of a modern glass skyscraper. Overlaid on this banner is the text 'FAKE JOB POST PREDICTION' in a bold, sans-serif font, with 'FAKE JOB POST' in white and 'PREDICTION' in red. Below the banner, the word 'Login' is displayed in a red, sans-serif font. Underneath 'Login' are two input fields: the first is labeled 'Username' and the second is labeled 'Password'. Both labels are in a small, black, sans-serif font. Below the password field is a red button with the word 'Login' in white, sans-serif font.

FAKE JOB POST

HOME LOGIN UPLOAD PREVIEW

FAKE JOB POST PREDICTION

Login

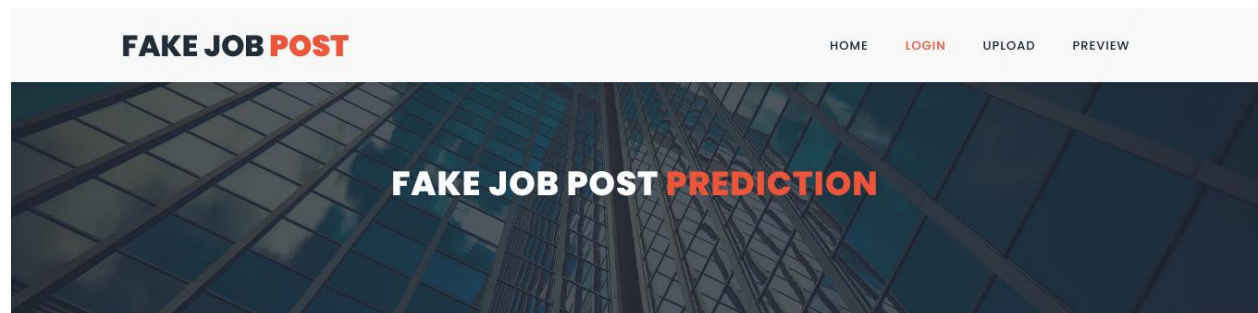
Username

Password

Login

Fig 5.2

- This is the static login page for the user. It was created as static login so that every one can access it. Username and password is common for any user.



Login

Username

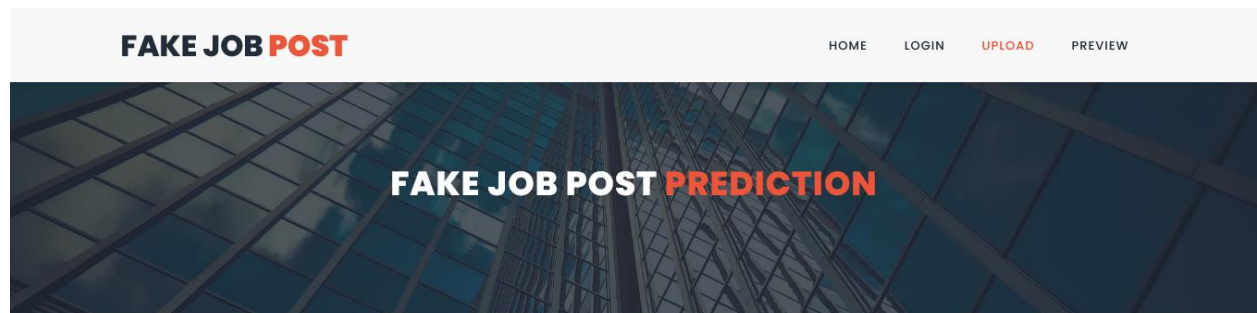
admin

Password

•••••

Login

Fig 5.3



Upload

Browse... upload.csv

Upload

Fig 5.4

- This is the upload section of our website. Dataset should be uploaded in the page so that it can train the data.

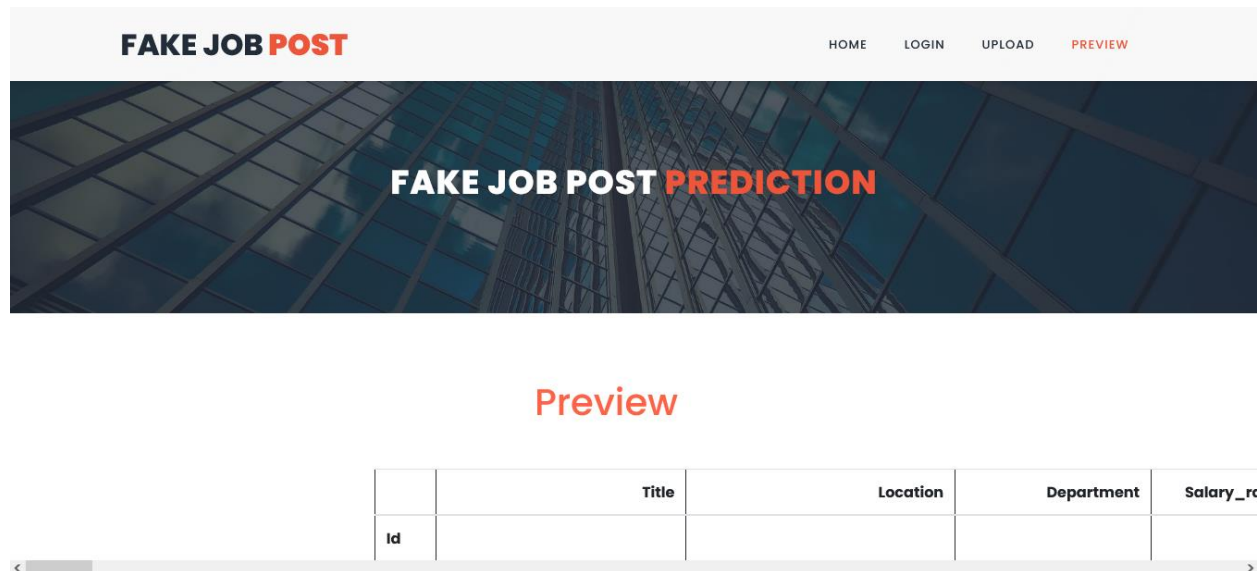


Fig 5.5

- This is the preview section of dataset.it displays the details of dataset like the parameters of dataset
- In this preview section all the sample data is displayed .

FAKE JOB POST					
			HOME	LOGIN	UPLOAD
					PREVIEW
Preview					
	Title	Location	Department	Salary_range	
Id					
1	Marketing Intern	US, NY, New York	Marketing	-	We're Food52, and we've created a about food and cooking. We attract Daily, TechCrunch, and on the Toda
2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	-	90 Seconds, the worlds Cloud Videc http://90#URL_fbe6559afac620a3c teams in 5 countries, 90 Seconds pi Singapore.http://90#URL_fbe6559c
3	Commissioning Machinery Assistant (CMA)	US, IA, Wever	-	-	Valor Services provides Workforce S employing the unique skills, experie

Fig 5.6

FAKE JOB POST		HOME		LOGIN	UPLOAD	PREVIEW	POSTING
ID	Job Title	Location					
2179	Business Technology Analyst	US , Jersey City	-	-			

Click to Train | Test

Fig 5.7

- At the end of preview page we can train the data using KNN algorithm.

FAKE JOB POST

[HOME](#)[LOGIN](#)[UPLOAD](#)[FRAUDULENT JOB POST](#)[TEXT PROCESSING](#)

FAKE JOB POST PREDICTION

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

prediction is:

Fig 5.8

- This is the main part of the project we can predict the job post is legit or fake.By choosing different parameters of the job post like employment type,required education,required experience and function.

FAKE JOB POST

HOME

LOGIN

UPLOAD

FRAUDULENT JOB POST

TEXT PROCESSING

FAKE JOB POST PREDICTION

Fake Job Post Prediction

Enter The Details

Telecommuting:

NO

Has_company_logo:

YES

Has_questions:

NO

Employment_type:

Other

Required_experience:

Internship

Required_education:

Bachelor's Degree

Function:

Marketing

submit

prediction is:

Fig 5.9

FAKE JOB POST

HOMELOGINUPLOADFRAUDULENT JOB POSTTEXT PROCESSING

FAKE JOB POST PREDICTION

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

prediction is:

Legit Job Post

Fig 5.10

- For corresponding parameters of the job post the result is Legit Job

FAKE JOB POST

HOMELOGINUPLOADFRAUDULENT JOB POSTTEXT PROCESSING

FAKE JOB POST PREDICTION

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

prediction is:

Fake Job Post

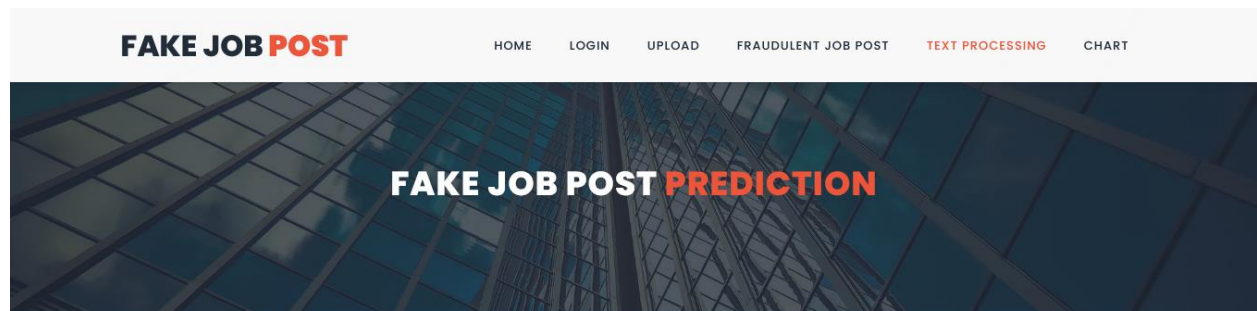
Fig 5.11

- For the corresponding parameters the given job post is a fake post

The screenshot shows the 'Fake Job Post Prediction' web application. At the top, a navigation bar contains the logo 'FAKE JOB POST' and links for 'HOME', 'LOGIN', 'UPLOAD', 'FRAUDULENT JOB POST', 'TEXT PROCESSING' (highlighted in red), and 'CHART'. Below the navigation bar is a large banner image of a modern glass skyscraper at night, with the text 'FAKE JOB POST PREDICTION' overlaid in white and red. The main content area features the title 'Fake Job Post Text Prediction' in red. Below the title is a text input field with the placeholder 'Enter The Details!'. Underneath the input field is a red button labeled 'Predict'. At the bottom of the form, the text 'prediction is:' is displayed.

Fig 5.12

- This is the Text prediction page of our website.
- In this page we can predict the post whether a legit or fake using text of job (description)



JOB TITLE: Itemization Review ManagerLOCATION: Fort Worth, TX
DEPARTMENT: Itemization ReviewREPORTS TO: VP Operations
GENERAL DESCRIPTION:Responsible for the overall aspects of Itemization Review operations: Personnel Hiring, Quality Control of Process, Workflow, monitoring the tracking of and accountability of staff regarding production standards and department expectations.DUTIES AND RESPONSIBILITIES:Oversee company's

Predict

prediction is:

Fig 5.13

FAKE JOB POST

HOME

LOGIN

UPLOAD

FRAUDULENT JOB POST

TEXT PROCESSING

CHART

FAKE JOB POST PREDICTION

Fake Job Post
Text Prediction

Enter The Details!

Predict

prediction is:

Legit Job Post

Fig 5.14

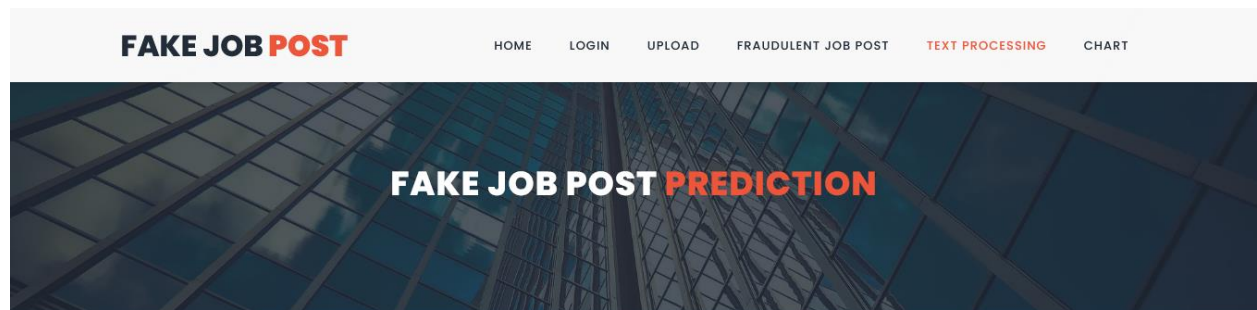


Fig 5.15

FAKE JOB POST

HOME

LOGIN

UPLOAD

FRAUDULENT JOB POST

TEXT PROCESSING

CHART

FAKE JOB POST PREDICTION

Fake Job Post
Text Prediction

Enter The Details!

Predict

prediction is:

Fake Job Post

Fig 5.16

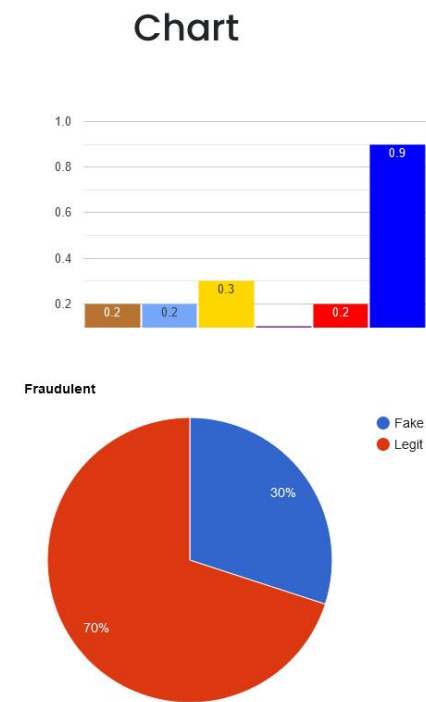
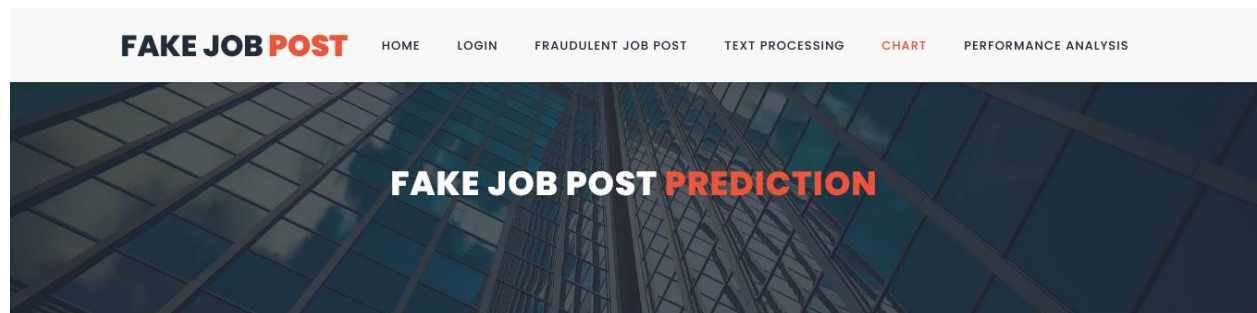


Fig 17

- It is the analysis of data using chart representation.

A. PLAGIARISM REPORT

A COMPARATIVE STUDY ON FAKE JO...

ORIGINALITY REPORT

9%

SIMILARITY INDEX

9%

INTERNET SOURCES

6%

PUBLICATIONS

5%

STUDENT PAPERS

PRIMARY SOURCES

1

"Table of Contents", IEEE Transactions on Computational Social Systems, 2015

Publication

1%

2

www.jpinfotech.org

Internet Source

1%

3

Shivangi Gheewala, Rakesh Patel. "Machine Learning Based Twitter Spam Account Detection: A Review", 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018

Publication

1%

4

researchr.org

Internet Source

1%

5

www.ijrte.org

Internet Source

1%

6

www.pantechelearning.com

Internet Source

1%

7

www.ijert.org

Internet Source

1%

8	www.iosrjournals.org Internet Source	1 %
9	Submitted to Victoria University Student Paper	1 %
10	bionyt.s807.sureserver.com Internet Source	<1 %
11	jpinfotech.org Internet Source	<1 %
12	cps-vo.org Internet Source	<1 %
13	ijettjournal.org Internet Source	<1 %
14	computer.ieeeprojects.org Internet Source	<1 %
15	Faiza Masood, Ghana Ammad, Ahmad Almogren, Assad Abbas, Hasan Ali Khattak, Ikram Ud Din, Mohsen Guizani, Mansour Zuair. "Spammer Detection and Fake User Identification on Social Networks", IEEE Access, 2019 Publication	<1 %

A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES

Student name: A.Ranga Chakradhar-38110450

Ranga Avinash-38110449

Guide name: Dr.S. PRAYLA SHYRY M.E. Ph.d.

ABSTRACT:

Lately, with the progression of current innovation and advertising, the advancement of new positions has turned into a test in this day and age. Along these lines, crafted by distorting predictions will be of worry to all. Likewise with numerous order assignments, making bogus predictions requires a great deal. The development requires the utilization of an assortment of strategies for AI and sequencing calculations, for example, KNN, memory sequencing, swarm collectors, and profound organizations, to anticipate whether the innovation is valid or bogus. We tried the Employment Scam Aegean Dataset (EMSCAD), which contains 18,000 occasions. We utilized three complete segments for this profound pipeline. The classifications shrouded in the preparation represent around 98% of the classifications (DNN) for foreseeing extortion.

Keywords: Fake Job, Online Recruitment, Machine Learning, Ensemble Approach.

INTRODUCTION

Today, modern and mechanical improvement opens up extraordinary freedoms for work searchers to make new and various positions. Alongside work posting, position searchers can pick dependent on schedule, degree, experience, qualification, and then some. The Internet and web-based media have added to the enlistment interaction. Social media assumes a major part in the accomplishment of enlistment, as it relies upon publicizing. Publicizing via online media and internet has set out new and new open doors for sharing data. All things being equal, the quick expansion in the quantity of open positions expands the level of deceitful promotions that badger work searchers. Accordingly, individuals are more averse to publicize for new positions to keep up with their wellbeing and security in the field of individual, expert

and expert data. Subsequently, the genuine motivation behind publicizing work through web-based media and internet business is a troublesome issue to accomplish their validity and believability. Innovation is intended to make life more straightforward and better for people around us, yet it doesn't make instability in proficient life. Assuming that work declarations can be effectively separated and bogus declarations can be anticipated, it will be a major advance in enlisting new representatives. Spreading fake positions can be a wellspring of dissatisfaction and disappointment for those searching for a task. The system attempts to anticipate employment opportunities by opening another window to the difficulties confronting representatives.

LITERATURE SURVEY

Fake Profile Identification in Online Social Networks

Sk.Shama, K.Siva Nandini, P.Bhavya Anjali, K. Devi Manaswi/ 2019

Innovation is blasting today. Cell phones are turning out to be increasingly shrewd. Innovation is tied in with observing new companions, observing companions is the Internet of Things in everybody's day to day existence, and it's not difficult to track down their own advantages. In any case, the expansion of long-range interpersonal communication locales leads to a great deal of issues, like deception of their personalities and web-based learning. Island clients have pointless information during swimming, which has been sent by fake clients. Studies have shown that 20-40 percent of online media profiles like Facebook are phony. Along these lines, understanding phony profiles on a public site prompts an answer utilizing a class. It doesn't contain inside and out data about the Facebook app.

Detecting Fake User Accounts on Different Social Media Networks

Sachin Ingle, Satish Borade , Sagar Awasare/ 2019

This will find fake user from multiple Social Networking platform by matching their friends network, matching profile details and their writing styles. In this way, we are matching different user accounts on multiple social network platforms and finding matched account. In this way we can detect fake user. Along with this we are improving efficiency of previous works in this area. To work on these disadvantages, we are proposing a system that can help to detect fake user from multiple social networking platforms and detect those users who have two accounts on different SMN's site with different name means same user creating

multiple accounts but different name for hiding identity.

Implications of Various Fake Profile Detection Techniques in Social Networks

Dr. Sanjeev Dhawan, Ekta/ 2018

The motivation behind this page is to cause to notice Facebook and distinguish fake qualifications. Facebook is broadly utilized on informal organizations, where clients can share messages, recordings, and recordings, and clients can add more companions to their profiles. Yet, it is hard to know whether a renewed individual is genuine. It very well may be an awful business. Distinguish culprits or profiles of different specialized cheats wanted. In this article, we have attempted to investigate various methods of correlation dependent on various projects for drawing various aspects. This application is just for programming dependent on client data or client wallets. It does exclude different choices, for example, the Facebook application.

Statistical features-based real-time detection of drifted Twitter spam

C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min / 2017

The proposed program can distinguish "change" tweets in unwritten tweets and carry out them in preparing. A few examinations have been directed to assess the proposed program. The outcomes show that our Lfun program can extraordinarily further develop spammers who know the reality of the world. There are likewise hindrances to our Lfun program. The upside of "old" spam is that it kills the impacts of "spam drift" so we can post genuine spam tweets before very long.

Automatically identifying fake news in popular Twitter threads

C. Buntain and J. Golbeck / 2018

In request to discover the bogus data on Twitter, we discussed the valid and solid motivations behind why non-proficient models are more than fake models. The fundamental exercises are restricted to different exercises among CREDBANK and PHEME, which can change the exchange of force. Assuming the essential designations of our models are altogether unique, contrasts in construction and capacity might be because of more restricted issues than plan abilities.

A performance evaluation of machine learning-based streaming spam tweets detection

C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian / 2015

The huge spam issue on Twitter has effectively grabbed the eye of analysts. A few scientists have concentrated on the idea of spam, trailed by various Twitter spam identification exercises. Accordingly, we examine the means engaged with setting it up in two phases: 1) portrayal, and 2) spam discovery on Twitter. On this page we give an essential outline of ML calculations for looking for spam tweets. To do this examination, we initially gathered a sum of 600 million tweets. We currently create around 6.5 million messages utilizing Trend Micro's Web Reputation System

A model-based approach for identifying spammers in social networks

F. Fathaliani and M. Bouguessa / 2017

In this article, we will take a gander at how to recognize a spam sender on a public site blended in with highlights dependent on the making of an uncontrolled strategy for distinguishing spammers. In our methodology, we initially address every informal community client with a vector that mirrors their conduct and associations with different members. The exploration introduced in this article shows that our uncontrolled techniques work better (now and again better) than numerous different strategies. Notwithstanding, our methodology permits us to investigate the advantages of genuine without neglected data and study hall access.

Spam detection of Twitter traffic: A framework based on random forests and non-uniform feature sampling

C. Meda, E. Ragusa, C. Gianoglio, R. Zunino, A. Ottaviano, E. Scillia, and R. Surlinelli / 2016

This work offers a wide scope of choices to browse before the Black Box Machine Learning Machine, utilizing an arbitrary woods memory calculation choice to recognize the spam sender in the rush hour gridlock load. The exploration depends on the most famous Twitter information bundles and the new Twitter client information bundles. Luckily, there are numerous Internet clients who use microblogging to disturb and spread terrible things. Distinguishing clients and recognizing spammers is a valuable method for decreasing traffic to Twitter's unknown substance.

EXISTING SYSTEM

Tingmin et al. Examine new ways and strategies to distinguish Twitter spam. The above research shows a relative investigation

of innovation. Then again, S. J. Somanet. al. Twitter has explored different practices of spammers via online media. The concentrate additionally incorporated a survey of the books that recognize spam on Twitter. Notwithstanding the exploration, there is as yet a hole in the current writing. So we are trying the furthest down the line innovation to distinguish spammers and recognize fake clients on Twitter to clear the hole.

DISADVANTAGES EXISTING SYSTEM

Because of Privacy Issues the Facebook dataset is very limited and a lot of details are not made public.

Having less accuracy

More complex

PROPOSED SYSTEM

At the arranged level, various moves should be made to follow the phony profile and gain from the criticism given by the calculation. This level can be effectively carried out by a long-range informal communication organization. 1. The intellectual cycle starts with choosing a test profile. 2. In the wake of choosing the profile, select the properties (i.e. properties) that are suitable for the space where the calculation is carried out. 3. Capacities prohibited via preparing. Plan customary trainings as you give new data and feed. 4. The rundown demonstrates whether it is bogus or valid. 5. Since the classifications can't be 100% all together; orders input thoughts. 6. This cycle is rehashed, and over the long run, no. How much data preparing is expanding and the order of phony names is turning out to be more sensible.

ADVANTAGES OF PROPOSED SYSTEM

Web-based media doesn't work on our lives, yet there are numerous issues with utilizing web-based media. Issues incorporate individual life, online badgering, misuse, and harassing. This is generally done utilizing fake profiles. In this venture, we utilized AI to recognize fake personalities and build up ways of securing individuals' lives.

ARCHITECTURE DIAGRAM



DESCRIPTION

The given system has an assortment of AI apparatuses, and the office was executed as depicted previously. The given framework gathers the information given before the calculation level, utilizing which we can decide the phony profile on Twitter by contrasting it and the reality of the most fit calculation on the three AI machines and the given informational index. Calculations There are various ways of recognizing issues dependent on the exhibition, experience, or climate in which the inspecting system chooses the fitting calculation dependent on

the data gave to accomplish the best outcomes.

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

System : Pentium Dual Core.

Hard Disk : 120 GB.

Monitor : 15'' LED

Input Devices : Keyboard, Mouse

Ram : 4 GB.

SOFTWARE REQUIREMENTS:

Operating system: Windows 7/10.

Coding Language : Python

MODULES:

Data Collection

Pre-Processing

Train and Test

Machine Learning Technique

Detection of Fake Profiles

Data Collection

We will be using a Python Library called Tweepy to connect to the Twitter API and collect the data. We download tweets containing certain key words, to incorporate the words or hash tags that contain relevant keyword related to fake users. Some of the most important fields are:

text, which contains the text included in the tweet.

created_at, which is a timestamp of when the tweet was created.

user, which contains information about the user that created the tweet, like the username and user id.

Pre-Processing

We scalarize the data and experiment through the calculation and the capacities put away in the x and the mark in the y.

Train and Test

We show the substance of the metadata anticipated from extra data about the client's tweets, while the substance is planned to see the nature of the client's instant message and the message utilized in the article.

Machine Learning Technique

Support Vector Machine (SVM): Support-vector machines (SVMs), and vector networks are controlled learning strategies and preparing calculations that investigate data utilized in examination and search techniques. In light of the preparation information (supervised learning), the calculation that delivers the best hyperplane records new models. Nerve organization: A muscle network is an organization of neurons, or circles, or in current speech, an organization of living things comprised of fake jobs. For fake jobs, a neural network (NN) is a gathering of regular and fake nerve cells used to process computerized information dependent on a mix of information. Typical memory: Normal calculation memory is calculation control. As the name recommends, this calculation makes backwoods with many trees. By and large, the more trees there are in the

backwoods, the more wonderful the timberland looks. Thusly, an enormous number of trees fill in the woodland, which gives genuine outcomes.

Detection of Fake Profiles

Knowing a Fake Profile, a system that gathers a bunch of data created prior to giving the level of the calculation we use can recognize a phony profile on Facebook by contrasting the reality of three machine calculations, calculations, and best practices on a given informational collection. Calculations are various methods of distinguishing an issue dependent on the presentation, experience, or climate in which the examining system chooses the suitable calculation for the data gave to accomplish the best outcomes.

ALGORITHMS USED

K-Nearest Neighbour (KNN)
Random Forest

K-Nearest Neighbour (KNN) Algorithm:

KNN is utilized to look and sort close to the k connection from the data gave. We use knn to choose the best and best motion pictures from every one of the films chose for this venture. Uneven Damage: One-Sided Damage (SVD) is one of the arithmetical strategies utilized as a technique for lessening AI. SVD is a framework augmentation strategy that decreases how much data set on resources by diminishing the time from N-guide to K-model ($K < N$). In the application framework, SVD is utilized as a separating accomplice. It utilizes a network design where each line addressing the client and every section addresses a similar article. The parts of this network are a rundown of clients. Factorization of this network is done by

breaking down a similar worth. It observes framework objects from the high level (utilizing highlight point) network objects. One-esteem disintegration is the method for breaking a grid into three more: $A = USVT$ where an is the network of a $m \times n$ device, U is the symmetrical $a \times r$ to one side of the lattice, demonstrating the connection between the client and the secret article, and S is a $r \times r$. The corner-to-corner framework that decides the force of each secret item and V is the askew $a \times n$ on the right half of the grid, addressing the connection between the sort and the secret article.

K, N, K, N , calculation AI calculation programming can be exceptionally basic things. Likewise see that it is near the worth of the item (proceeding with K, N) coming about because of the given (stage K, n) stage.

STEP 1: Start

STEP 2: Income: $D =$

STEP 3: Another instance of planning $x = (x_1 \dots x_n)$

STEP 4: Count $(x_i, c_i) d(x_i, x)$ for each case.

STEP 5: Reduce $d(x_i, x)$ over the base, $(I = 1 \dots N)$

STEP 6: x : For instance, select K close to D_kx

STEP 7: Scores x, D_kx General classifications

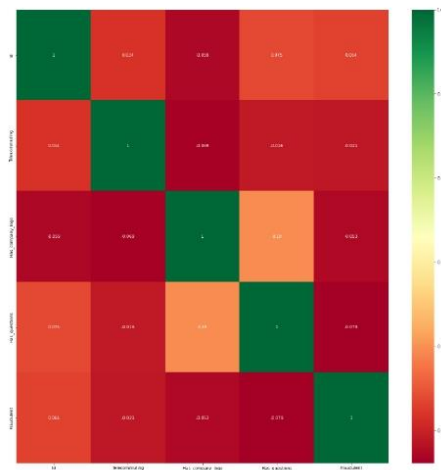
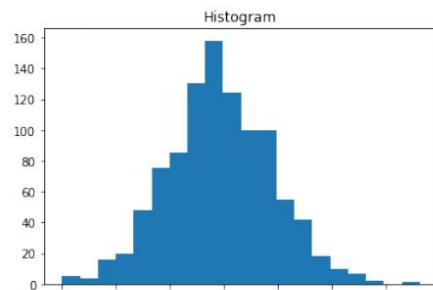
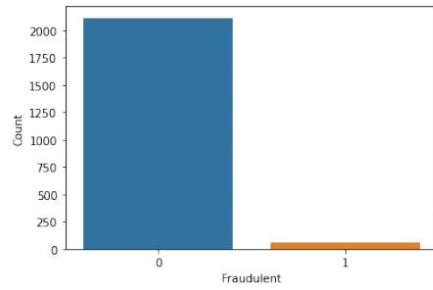
Stage 8: Completion

RANDOM FOREST

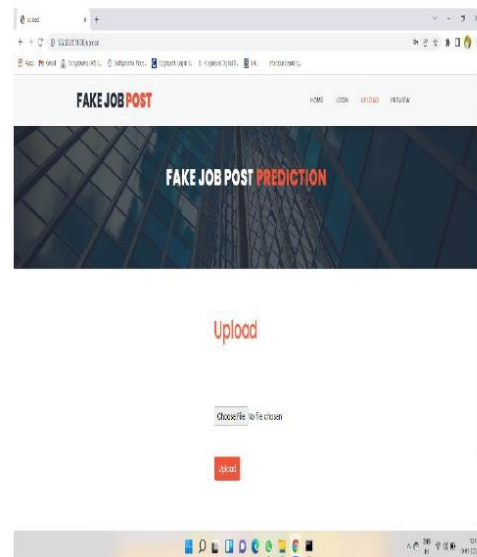
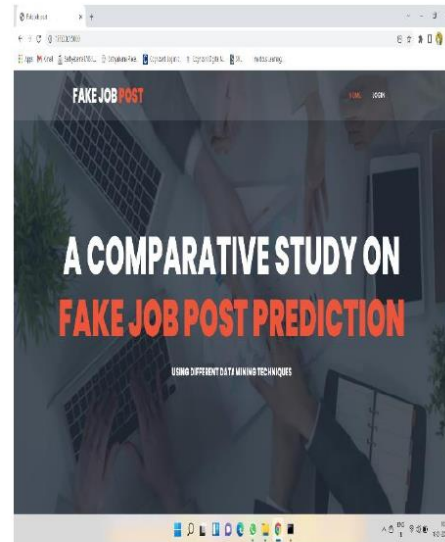
Random Forest Classifier is applied for clasifying job post as fake. The performance measure metrics such as Accuracy, Recall,

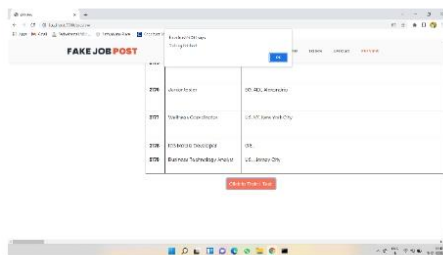
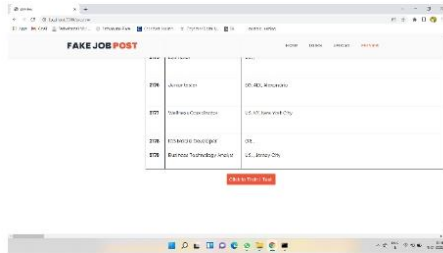
Precision, and Confusion matrix are used for evaluating the prediction for proposed classifier.

JOB POST DATASETS



OUTPUT SCREENSHOTS





CONCLUSION

Understanding position misrepresentation will just direct work searchers to get something authentic for the organization. To battle ladies' extortion, various AI calculations should be tended to in this article. Review techniques are utilized to give different proportions of execution to identify working environment misrepresentation. Research shows that ranger service isn't a distributed device. The normal technique is 98.27% more exact than the current strategy.

REFERENCES

- [1] C. Chen, S. Wen, J. Zhang, Y. Xiang, J. Oliver, A. Alelaiwi, and M. M. Hassan, "Investigating the deceptive information in Twitter spam," *Future Gener. Comput. Syst.*, vol. 72, pp. 319–326, Jul. 2017.
- [2] I. David, O. S. Siordia, and D. Moctezuma, "Features combination for the detection of malicious Twitter accounts," in *Proc. IEEE Int. Autumn Meeting Power, Electron. Comput. (ROPEC)*, Nov. 2016, pp. 1–6.
- [3] M. Babcock, R. A. V. Cox, and S. Kumar, "Diffusion of pro- and anti-false information tweets: The black panther movie case," *Comput. Math. Org. Theory*, vol. 25, no. 1, pp. 72–84, Mar. 2019.
- [4] S. Keretna, A. Hossny, and D. Creighton, "Recognising user identity in Twitter social networks via text mining," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2013, pp. 3079–3082.
- [5] C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, "A machine learning approach for Twitter spammers detection," in *Proc. Int. Carnahan Conf. Secur. Technol. (ICCST)*, Oct. 2014, pp. 1–6.
- [6] W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, "Real-time Twitter content polluter detection based on direct features," in *Proc. 2nd Int. Conf. Inf. Sci. Secur. (ICISS)*, Dec. 2015, pp. 1–4.
- [7] H. Shen and X. Liu, "Detecting spammers on Twitter based on content and social interaction," in *Proc. Int. Conf. Netw. Inf. Syst. Comput.*, pp. 413–417, Jan. 2015.

- [8] G. Jain, M. Sharma, and B. Agarwal, "Spam detection in social media using convolutional and long short term memory neural network," *Ann. Math. Artif. Intell.*, vol. 85, no. 1, pp. 21–44, Jan. 2019.
- [9] M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, "A topic-based hidden Markov model for real-time spam tweets filtering," *Procedia Comput. Sci.*, vol. 112, pp. 833–843, Jan. 2017.
- [10] F. Pierri and S. Ceri, "False news on social media: A data-driven survey," 2019, arXiv:1902.07539. [Online]. Available: <https://arxiv.org/abs/1902.07539>
- [11] S. Sadiq, Y. Yan, A. Taylor, M.-L. Shyu, S.-C. Chen, and D. Feaster, "AAFA: Associative affinity factor analysis for bot detection and stance classification in Twitter," in *Proc. IEEE Int. Conf. Inf. Reuse Integr. (IRI)*, Aug. 2017, pp. 356–365.
- [12] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, "Segregating spammers and unsolicited bloggers from genuine experts on Twitter," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.