PANIMALAR
ENGINEERING COLLEGE
An Autonomous Institution, Affiliated to Anna University, Chennai
A Christian Minority Institution
(JAISAKTHI EDUCATIONAL TRUST)
Approved by All India Council for Technical Education

Department of Computer Science and Engineering

REAL & FAKE JOB CLASSIFICATION USING NLTK TECHNIQUE

Team Members Name / Register Number

Kavushick P – 211419104133
Manigandan L – 211419104159
Nareshkumar C – 211419104177

Guide Name & Designation
DR. S. HARIHARAN, M.E., M.Tech., Ph.D.
Professor

Coordinator Name & Designation
Dr. G. SENTHIL KUMAR, M.C.A., M.Phil., M.B.A., M.E., Ph.D.,
Professor

Introduction

Employment scams are on the rise. According to CNBC, the number of employment scams doubled in recent years. The current market situation has led to high unemployment. Economic stress and the coronavirus’s impact have significantly reduced job availability and job loss for many individuals. A case like this presents an appropriate opportunity for scammers. Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. Most scammers do this to get personal information from the person they are scamming. Personal information can contain addresses, bank account details, social security numbers, etc. I am a university student, and I have received several such scam emails. The scammers provide users with a very lucrative job opportunity and later ask for money in return. Or they require investment from the job seeker with the promise of a job. This is a dangerous problem that can be addressed through Machine Learning techniques and Natural Language Processing (NLP). This data contains features that define a job posting. These job postings are categorized as either real or fake. Fake job postings are a tiny fraction of this dataset. That is as excepted. We do not expect a lot of phony job postings.

Objective of the Project

- To avoid fraudulent Job postings on the internet, we target to minimize the number of such frauds through the Machine Learning approach to predict the chances of a job being fake so that the candidate can stay alert and make informed decisions if required.
- The model will use NLP to analyze the sentiments and pattern in the job posting and TF-IDF vectorizer for feature extraction.
- In this model, we are going to use Synthetic Minority Oversampling Technique (SMOTE) to balance the data and for classification, we used Random Forest to predict output with high accuracy, even for the large dataset it runs efficiently, and it enhances the accuracy of the model and prevents the overfitting issue.
- The final model will take in any relevant job posting data and produce a result determining whether the job is real or fake.

Literature Survey

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

Author: Sultana Umme Habiba

Year : 2021

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. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron 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 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.

Literature Survey

Title: Fake Job Recruitment Detection Using Machine Learning Approach.

Author: Samir Bandyopadhyay, Shawni Dutta

Year: 2020

To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.

Literature Survey

Title: Identifying Real and Fake Job Posting-Machine Learning Approach

Author: Devi.A P 1 , Sandhiya.S2 , Gayathri.R 5

Year: 2021

The process of searching jobs is one of the most problematic issue freshers face, this process is used by various scamsters to lure freshers into scams and profit from the students. In order to avoid this, this paper proposes a system with deep learning and flask for front-end, that can identify fake jobs. While browsing for jobs online we saw that many scamsters demanded money for booking slots to interviews that did not exist and also extort money from students with promise of giving them jobs in return, this served as motivation for this proposal.The objectives that are to be considered are: Prediction of real or fake job. The proposed system is basically an ANN classification model based on Multinomial Naive Bayes algorithm to determine fake job posting or real one.. This therefore makes searching of jobs much more efficient and also allows the users to be worry free when they search for jobs online.

Problem Statement

- The task is to build a text classification model using the Natural Language Toolkit (NLTK) that can accurately classify job postings into different job categories based on the job title and job description.

- Given a dataset of job postings with their corresponding job titles and descriptions, the goal is to train a machine learning model that can classify new job postings into one of the predefined job categories, such as Engineering, Sales, Finance, Marketing, etc.

- The model should take into account various text processing techniques such as tokenization, stemming, and stop-word removal, and use appropriate machine learning algorithms to achieve high accuracy in classification.

- The success of the model will be evaluated based on metrics such as accuracy, precision, recall, and F1 score, using appropriate evaluation techniques such as cross-validation or hold-out validation. The ultimate goal is to build a text classification model that can be used in real-world applications to automate the process of job classification.

Proposed System

- The proposed model is to build a machine learning model that is capable of classifying whether the job is fake or not.
- The fake jobs are considered to be widespread and controlling them is very difficult as the world is developing toward digital everyone now has access to internet and they can post whatever they want. So there is a greater chance for the people to get misguided.
- The machine learning is generally build to tackle these type of complicated task like it takes more amount of time to analyse these type of data manually.
- The machine learning can be used to classify whether the job is fake or not by using the previous data and make them to understand the pattern and improve the accuracy of the model by adjusting parameters and use that model as the classification model.
- Different algorithms can be compared and the best model can be used for classification purpose.

Software / Hardware used

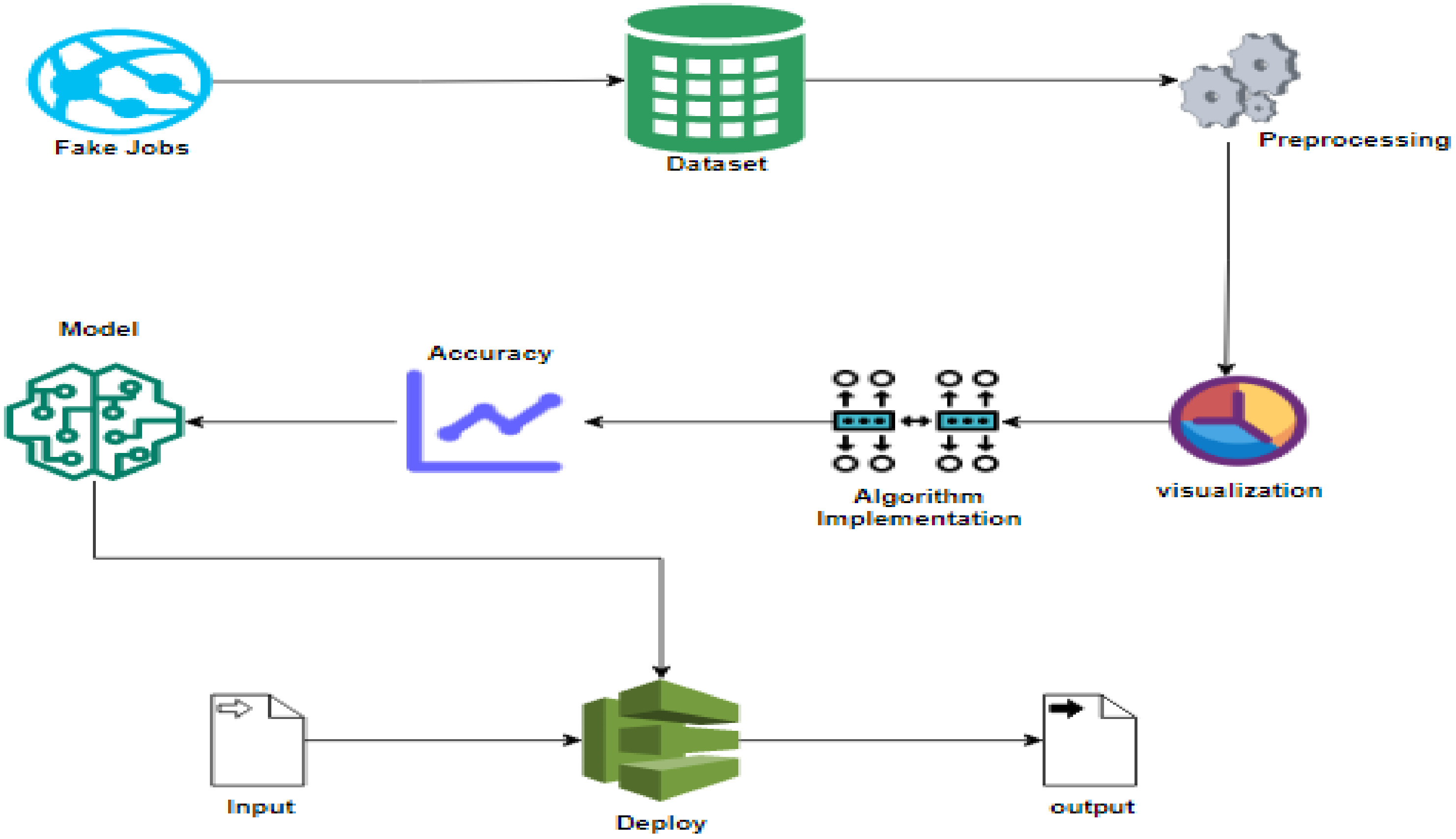
Software Requirements:

- Operating System : Windows 10 or later
- Tool : Anaconda with Jupyter Notebook

Hardware Requirements:

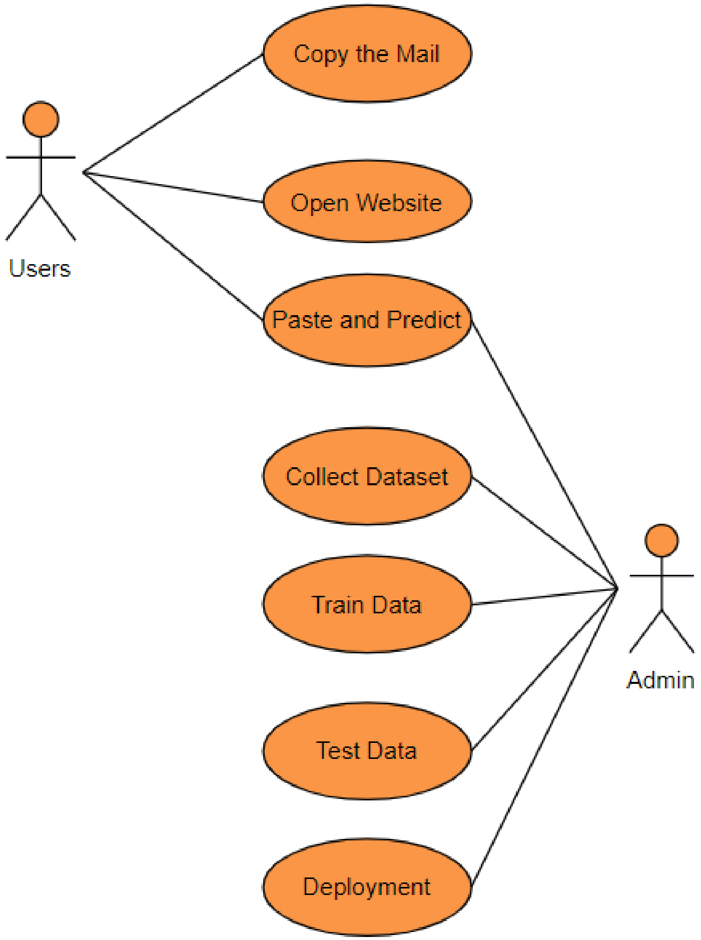
- Processor : Intel i3
- Hard disk : minimum 10 GB
- RAM : minimum 4 GB

System Architecture



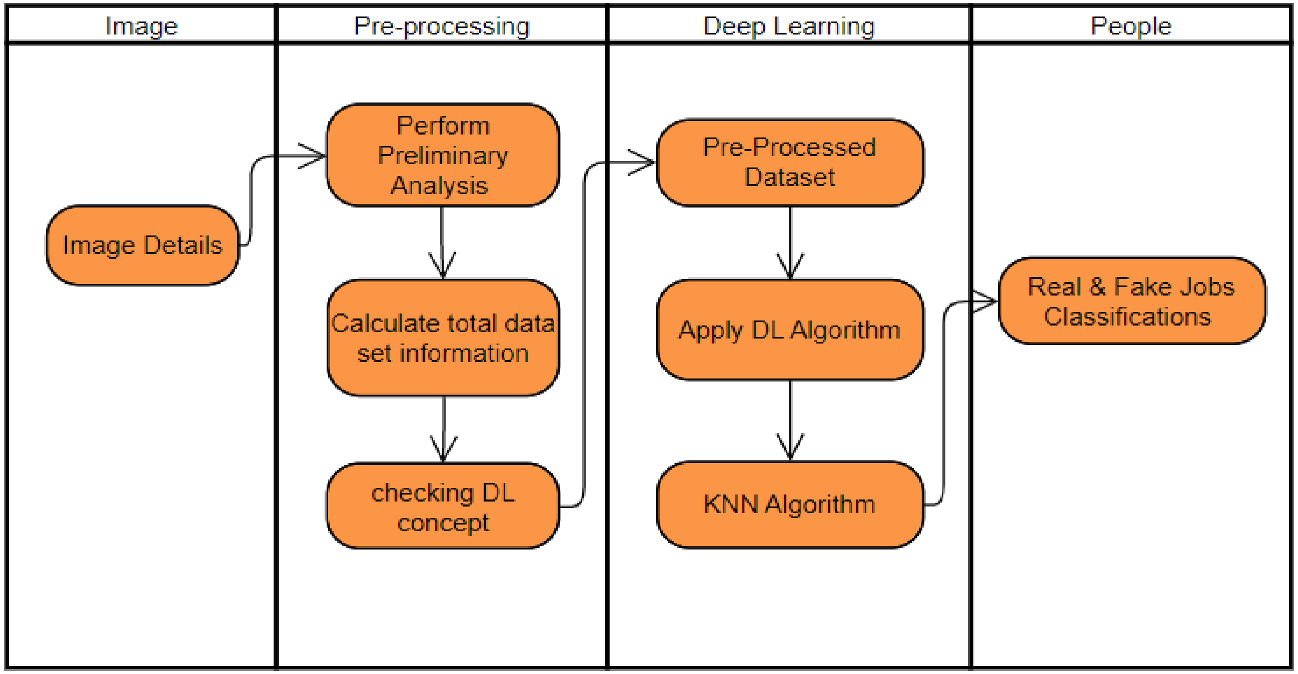
System Design

- **Use Case Diagram**



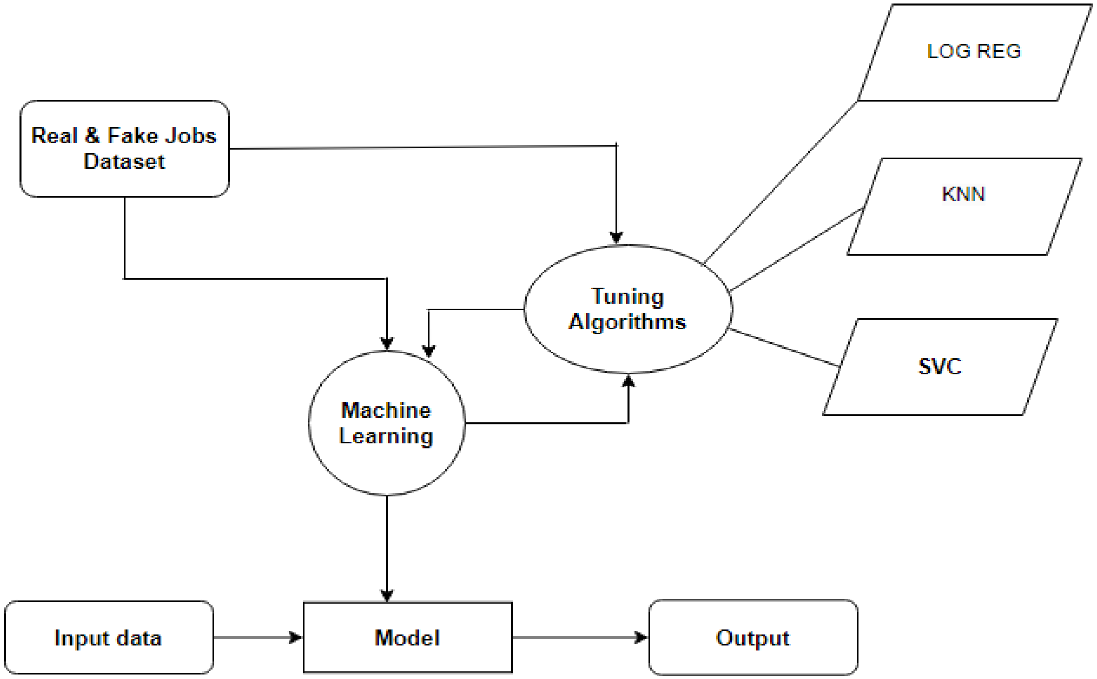
System Design

- **Activity Diagram**



System Design

- **Data Flow Diagram**



Module Description

Data Pre-processing:

- Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset.
- If the data volume is large enough to be representative of the population, you may not need the validation techniques.
- However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset.
- To finding the missing value, duplicate value and description of data type whether it is float variable or integer
- . The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

Module Description

Data Disualization:

- Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data
- . Data visualization provides an important suite of tools for gaining a qualitative understanding.
- This can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more.
- With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance.
- Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Module Description

KNN

- 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 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.

Module Description

Logistic Regression

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

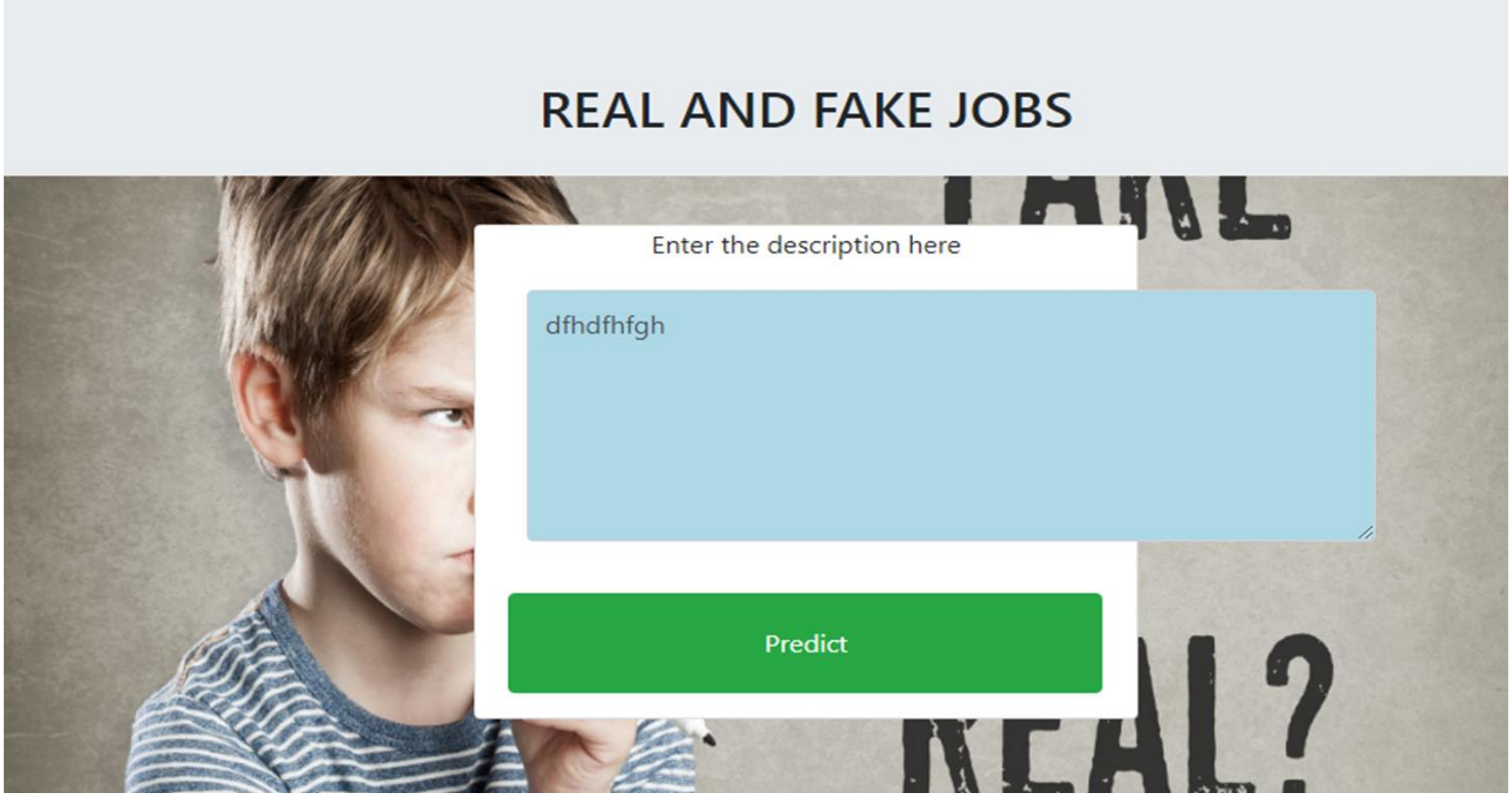
- Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

Testing

- Unit Testing

Test Case ID	Description	Test Steps	Expected Result	Actual Result	Pass/ Fail
1	Test the Data analyzsis module to ensure that it can correctly identify real or fake	1. Input a real job offer 2. Input a fake job offer. 1. Run the data analysis module	The data analysis module should correctly identify the real or fake job.	The data analysis moduleshould correctlyidentify the real or fake job	Pass
2	Test the deployment module to ensure that it can identify the real and fake job	1. Get the user input 2. Run the deployment module	The deployment module should identify the real or fake job	The deployment module should identify the real or fake job	Pass

Screen Shots

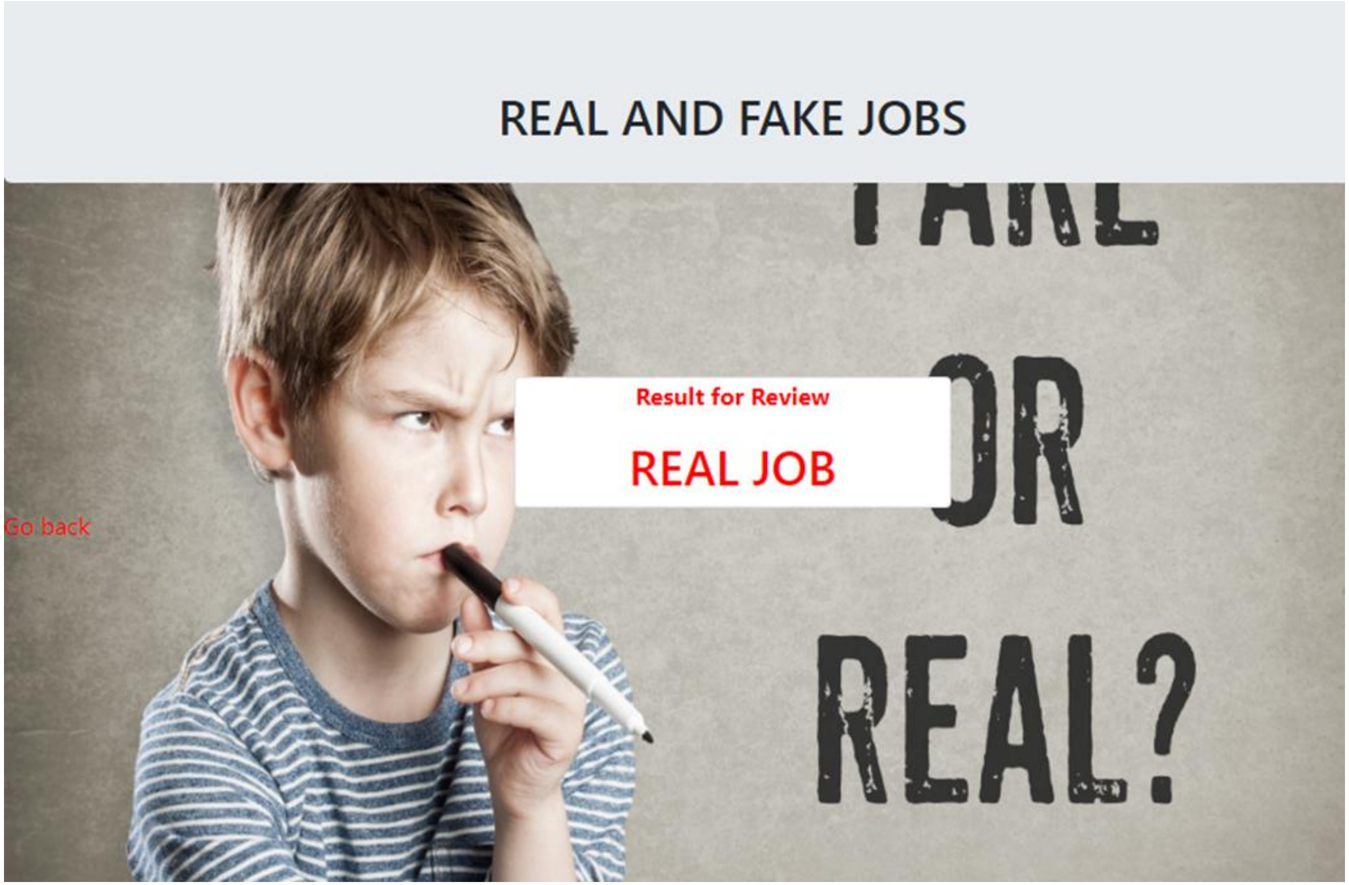


REAL AND FAKE JOBS

Enter the description here

dfhdfhfgh

Screen Shots



REAL AND FAKE JOBS

Result for Review

REAL JOB

Conclusion / Feature Enhancement

Conclusion

- The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation.
- The best accuracy on public test set of higher accuracy score algorithm will be find out.
- The founded one is used in the application which can help to find the Real and fake jobs.

Future work

- Deploying the project in the cloud.
- To optimize the work to implement in the IOT system.
- The Future work includes fraudulent job notification to connect with cloud

References

- Shawni Dutta and Prof. Samir Kumar Bandyopadhyay, “Fake Job Recruitment Detection Using Machine Learning Approach”, International Journal of Engineering Trends and Technology (IJETT) – Volume 68 Issue 4- April 2020.
- Sultana Umme Habiba, Md. Khairul Islam, Farzana Tasnim, “A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques”, 2nd International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), 2021.
- Devi.A P, Sandhiya.S , Gayathri.R, “Identifying Real and Fake Job Posting-Machine Learning Approach”, IARJSET Vol. 8, Issue 8, August 2021.
- Mr. Gulshan, Mr. Mukund, Mr. Ajay, Mr. Pankaj Kumar, Mrs. Aruna M, Dr. Malatesh S, ” Fake Job Post Prediction Using Machine Learning Algorithms”, IJIRT | Volume 9 Issue 3, August 2021.