

**School of Computer Science and Engineering**

**J Component report**

**Programme : B. Tech CSE (Core)**

**Course Title : Machine Learning**

**Course Code : CSE4020**

**Slot : D1**

**Title: Online Job Recruitment Scam Detection**

**Team Members:**

**Suravaram Rishabh Reddy | 20BCE1545**

**Kiran BSVS | 20BCE1582**

**Faculty: Dr. Bhargavi.R**   **Sign:**

**Date:**

**DECLARATION**

**I HEREBY DECLARE THAT THE REPORT TITLED “ONLINE JOB RECRUITMENT SCAM DETECTION” SUBMITTED BY ME TO VIT CHENNAI IS A RECORD OF BONA-FIDE WROK UNDERTAKEN BY ME UNDER THE SUPERVISION OF DR. BHARGAVI.R, SCHOOL OF COMPUTER SCIENCE AND ENINEERING, VELLORE INSTITUTE OF TECHNOLOGY, CHENNAI**

Suravaram Rishabh Reddy

Signature of the Candidate

Kiran BSVS

Signature of the Candidate

Suravaram Rishabh Reddy

Reg. No. 20BCE1597

Kiran BSVS

Reg. No. 20BCE1582

**CERTIFICATE**

**CERTIFIED THAT THIS PROJECT REPORT TITLED ONLINE JOB RECRUITMENT SCAM DETECTION IS A BONAFIDE WORK OF SURAVARAM RISHABH REDDY (20BCE1545), KIRAN BSVS (20BCE1582) CARRIED OUT THE PROJECT WORK UNDER MY SUPERVISION AND GUIDANCE FOR CSE4020 – MACHINE LEARNING.**

Dr. Bhargavi.R

**SCOPE, VIT Chennai**

**ACKNOWLEDGEMENT**

**WE WISH TO EXPRESS OUR SINCERE THANKS AND DEEP SENSE OF GRATITUDE TO OUR PROJECT GUIDE, DR. BHARGAVI.R , SCHOOL OF COMPUTER SCIENCE AND ENGINEERING FOR HIS CONSISTENT ENCOURAGEMENT AND VALUABLE GUIDANCE OFFERED TO US THROUGHOUT THE COURSE OF THE PROJECT WORK.**

**WE ALSO TAKE THIS OPPORTUNITY TO THANK ALL THE FACULTY OF THE SCHOOL FOR THEIR SUPPORT AND THEIR WISDOM IMPARTED TO US THROUGHOUT THE COURSES.**

**WE THANK OUR PARENTS, FAMILY, AND FRIENDS FOR BEARING WITH US THROUGHOUT THE COURSE OF OUR PROJECT AND FOR THE OPPORTUNITY THEY PROVIDED US IN UNDERGOING THIS COURSE IN SUCH A PRESTIGIOUS INSTITUTION.**

**Suravaram Rishabh Reddy**

**20BCE1545**

**Kiran BSVS**

**20BCE1582**

**ABSTRACT**

The project aims to create a machine learning model that can detect fraudulent online job recruitment strategies by using Machine Learning Algorithms. The goal of the research is to develop a precise and efficient model that can instantly identify probable scams. In order to do this, the project will apply the CRISP-DM methodology, which includes several steps like exploratory data analysis, data sampling, data pre-processing, text processing, feature selection, encoding categorical variables, classification using ML algorithms, hyper-parameter tuning, and evaluation.

The overall objective of this project is to develop a machine learning model that can recognize online job recruitment fraud with accuracy, hence preventing job seekers from falling for scams.

**CONTENTS**

Declaration. . . . . . . . . . . . . . . . . . . . . . . . . . . ……………... …… ii

Certificate. . . . . . . . . . . . . . . . . . . . . . . . . . . . .. …………...............iii

Acknowledgement. . . . . . . . . . . . . . . . . . . . . . . . . . . ……………...iv

Abstract. . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………...................v

1) Introduction

1.1 Objective and goal of the project . . . . . . . . . . . . . . . . . …….vii

1.2 Problem Statement. …. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vii

1.3 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………...vii

1.4 Challenges . . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………....vii

2) Literature Survey. . . . . . . . . . . . . . . . . . . . . . . . . . . . . …………......…..viii

3) Requirements Specification. . . . . . . . . . . . . . . . . . . . . . . . . . . . . ……….ix

3.1 Hardware Requirements . . . . . . . . . . . . . . . . . . . . . . . . . … ix

3.2 Software Requirements . . . . . . . . . . . . . . . . . . . . . . . . . . . . ix

4) Implementation of System . . . . . . . . . . . . . . . . . . . . . . . .. .. . . . . . . ix

4.1 Data Collection . . . . . . . . . . . . . . . . . . . . . . . . . .… …. ix

4.2 Data Pre-Processing… . . . . . . . . . . . . . . . . . . . . . . . . . . . . .. ix

4.3 Feature Extraction. . … . . . . . . . . . . . . . . . . . . . . . . . . . . . ix

4.4 Text Analysis . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ix

4.5 Fraud Detection Model . . . . . . . . . . . . . . . . . . . . . . . . . … ix

4.6 Model Evaluation . .... . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x

5) Results & Discussion. . . . . . . . . . . . . . . . . . . . . . . .. .. . . . . . . . . .. ..x

6) Conclusion and Future Work. . . . . . . . . . . . . . . . . . .. .. . . . . . . . . ..xii

7) References. . . . . . . . . . . . . . . . . . .. .. . . . . . . . . .. .. . . . .. .. . . . . . . .xiii

1. **Introduction**

**1.1 Objective and goal of the project**

The objective of this project is to develop a machine learning model using natural language processing (NLP) to identify and prevent online job recruitment scams.

The primary goal of the project is to protect job seekers from fraudulent activity and create a safer digital environment for them. The project aims to build a robust and accurate model by employing various ML algorithms and enhancing its performance through hyper-parameter tuning. The CRISP-DM technique will be used to complete the project, including phases such as feature selection, pre-processing, and data sampling. Ultimately, the project aims to contribute to resolving the issue of fraudulent practices in online job recruitment and provide job seekers with genuine and secure job opportunities.

**1.2 Problem Statement**

The rise of online job recruitment scams has made job seekers vulnerable to fraudulent practices, where scammers present false job offers and request money or personal information. This issue needs urgent attention to protect job seekers from financial and emotional harm. Therefore, this research aims to develop a machine learning model using natural language processing to identify and prevent such scams and create a safer digital environment for job seekers.

**1.3 Motivation**

The most important factor that served as a motivation during research and also during the production phase is the number of fraudulent job advertisements which are all over the internet and tricking many job-seekers. The increased accessibility has also made job seekers more susceptible to online job scams, which may present false employment opportunities to individuals and then ask for money or personal information. This type of fraudulent activity can have severe consequences for job seekers, including financial loss, identity theft, and emotional distress. Our predication systems aims to create a reliable and accurate machine learning model that can protect job seekers from online job scams and create a safer digital environment for them.

**1.4 Challenges**

NA values in the dataset huge in number and iterative imputing is implemented to assign the NA values. Outliers in the dataset caused less accuracy in prediction since some data values may be entered wrong or the for unknown reason.

**2. Literature Survey**

The project visualizes the data to better comprehend the dataset before starting the model-building process and to assure a better model performance, the dataset is partitioned and changed to fit the required model.

Shawni Dutta, Prof.Samir Kumar Bandyopadhyay paper titled ‘Fake Job Recruitment Detection Using Machine Learning Approach’ aims to help job seekers by identifying fake job postings and focusing on legitimate job opportunities. The study predicts that the Decision Tree Classifier is the most effective predictor for identifying fake job postings compared to other classifiers such as the Multi-Layer Perceptron Classifier, Naïve Bayes, and K-Nearest Neighbour Classifier. Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. paper titled ‘ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection’ propose an ensemble based ORFDetector model for online fraud detection. The model is based on three baseline classifiers which include J48, Logistic Regression, and Random Forest, and three ensemble techniques (Majority Vote, Average Vote, and Maximum Vote) are applied to create the ORFDetector framework. The ORFDetector gives us a very accurate prediction result. It indicates that in spite of achieving a good accuracy, the specificity is low which classifies many legitimate companies as fraud. Mehboob, A., & Malik, M. S. I paper titled ‘Smart fraud detection framework for job recruitments’ aims to determine the role of the characteristics of an organization, job description, and compensation type as a stand-alone model and their combined effect on detecting fraud. The paper analysis salary\_range, company\_profile, organization\_name, required\_education, and has\_multiple\_jobs as the most important features in detecting fraud. Anita, C. S., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepakkumar, G. paper titled ‘Fake job detection and analysis using machine learning and deep learning algorithms’ aims to use machine learning and deep learning algorithms may be used to detect bogus jobs once the data has been cleaned and analyzed. The simplest and most straightforward machine learning classification approach is logistic regression. The result tells us that by using the classifiers, we can determine that, when compared to other classification algorithms, the Bi-Directional LSTM provides the best accurate result for identifying bogus jobs. Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V., & Ludi, S. paper titled ‘Detection of fake job postings by utilizing machine learning and natural language processing approaches’ aims to use machine learning and natural language techniques to determine bogus job advertisements on the internet.

**3. Requirements Specification**

* 1. **Hardware Requirements**

3.1.1 Memory space: 1GB Ram

3.1.2 Disk space: 6GB of disk

3.1.3 CPU Processor: 8 CPU core with 64-bit, 2.5 GHz.

* + 1. Server overhead: 2-4GB or 10% system overhead CPU cores.

**3.2 Software Requirements**

3.2.1 Installation of Jupyter notebook and python.

3.2.2 Windows 10 or 11

3.2.3 Python 3.11.0 (64-bit) amd.exe.

1. **Implementation of System**

**4.1 Data Collection:**

The first step is to collect data related to job postings from various online job portals. The data collected includes job titles, descriptions, company names, locations, and other relevant information. The data is collected in a structured format to facilitate analysis.

**4.2 Data Pre-processing:**

The collected data needs to be cleaned and pre-processed to remove irrelevant information, duplicates, and inconsistencies. This involves techniques such as data cleaning, data transformation, and data normalization. The pre-processing step is important to ensure that the data is consistent and accurate.

**4.3 Feature Extraction:**

The relevant features are extracted from the pre-processed data. Features such as job title, job description, company name, location, and other relevant information are extracted to be used in the analysis. Feature extraction is an important step in machine learning as it helps to identify the most important features for analysis.

**4.4 Text Analytics:**

The extracted features are analyzed using natural language processing techniques to identify patterns and trends. The job descriptions are analyzed for specific keywords and phrases that indicate fraudulent job postings. Text analytics helps to identify patterns in the data that are indicative of fraudulent job postings.

**4.5 Fraud Detection Model:**

A machine learning model is developed using the extracted features and text analytics results to detect fraudulent job postings. The model is trained on a dataset of known fraudulent job postings and non-fraudulent job postings. Machine learning models are used to identify patterns in the data that are indicative of fraudulent job postings.

indicative of fraudulent job postings.

**4.6 Model Evaluation:**

The developed model is evaluated using various metrics such as accuracy, precision, recall, and F1 score.The evaluation helps to determine the effectiveness of the model in detecting fraudulent job postings. Model evaluation is an important step in machine learning as it helps to identify the strengths and weaknesses of the model.

1. **Results and Discussion**

Evaluation is a crucial step in any data mining project, as it determines the effectiveness of the model in terms of various metrics. Commonly used evaluation methods include accuracy, F1-score, precision, and recall. Accuracy calculates the overall performance of the model, while F1-score measures the harmonic mean of recall and precision. Precision assesses the percentage of positive instances out of total predicted positive instances, and recall calculates the percentage of positive instances out of the total actual positive instances. In our project, Gaussian Naive Bayes was found to be the best performing algorithm with high accuracy and one of the least execution times.

The following tables show the scores of various models for the different models applied on the dataset:

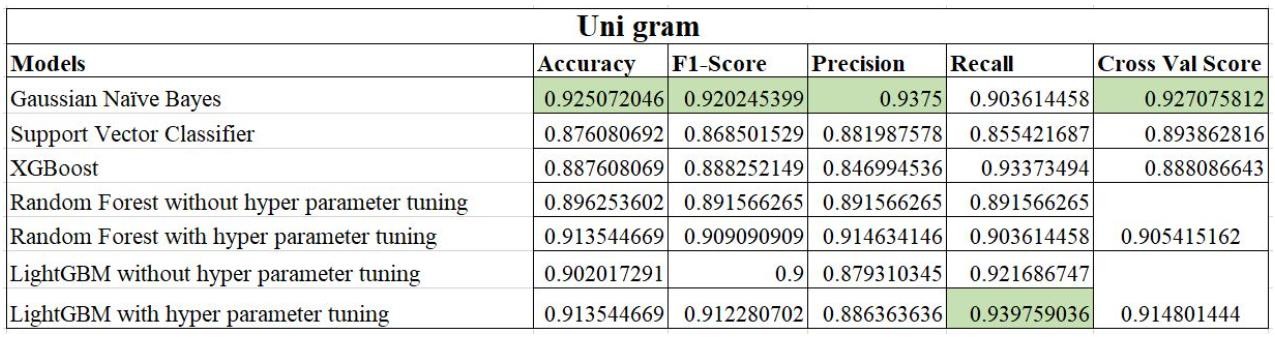
1. Uni-Gram

Fig.1. Uni Gram

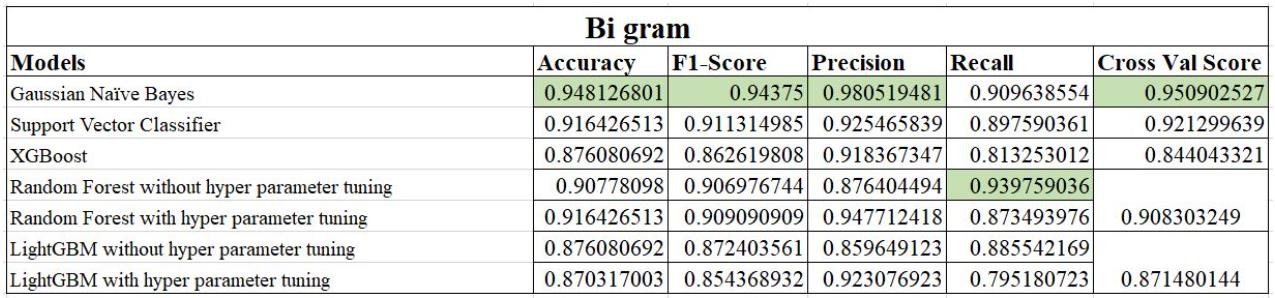
1. Bi-Gram:

Fig.2. Bi-Gram

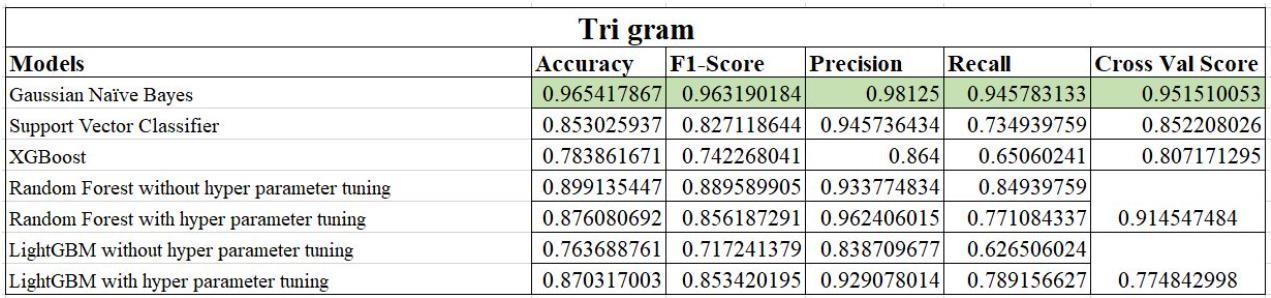
1. Tri-Gram:

Fig.3. Tri-Gram

* + 1. TF-IDF:

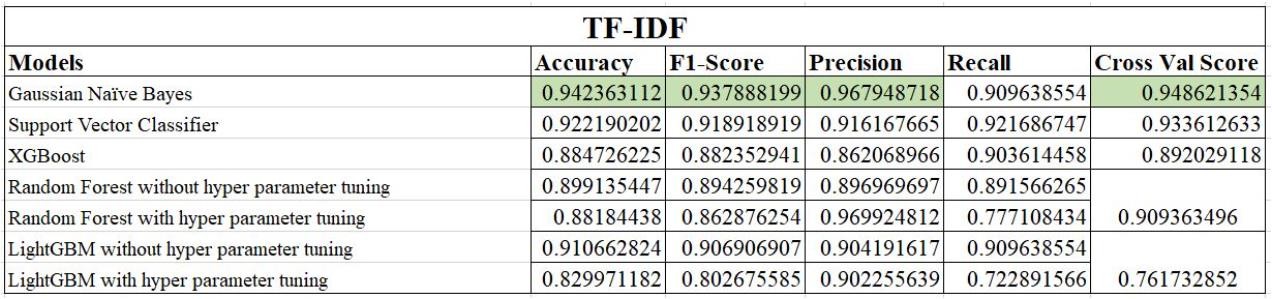


Fig.4. TF-IDF

Thus, from the above results it can be observed that for Gaussian Naive Bayes for tri gram we get the best model for detecting whether the job offers are genuine or not.

**6. Conclusion and Future Work**

The goal of this research was to create a machine learning model with high accuracy and low latency. Several steps were taken to accomplish this. First, low-correlation categorical variables were removed, reducing the model's dimension and improving its manageability. The data was then processed using a bag of words model. However, due to high-dimensionality issues with a sparse matrix, the model's accuracy may have suffered. To address this issue, the ChiSquare feature selection method was used to select relevant features and eliminate noisy data, resulting in an increase in the accuracy of the machine learning algorithm. Gaussian Naive Bayes was found to be the best performer among the machine learning algorithms tested. It is a quick and simple algorithm that is extremely effective. Due to its ability to manage sparse data and low complexity, it is an excellent choice for this problem. Overall, the research was aimed at developing a model with high accuracy and low latency.

In future, could include investigating other feature selection techniques, such as Recursive Feature Elimination or LASSO regularization, to improve the model's accuracy while reducing its dimensionality. Incorporating more advanced natural language processing techniques, such as word embeddings or neural networks, may also aid in the capture of more complex word relationships and improve the model's performance on more nuanced text data. Furthermore, additional testing and optimization of hyperparameters may be required to ensure the model consistently performs well on a wide range of data sets and use cases. By exploring these options, the project will be able to continue to refine and improve its machine learning model, making it more effective for a wide range of applications.

**7**. **REFERENCES**

1. Shawni Dutta, Prof.Samir Kumar Bandyopadhyay. Fake Job Recruitment Detection Using Machine Learning Approach -International Journal of Engineering Trends and Technology, 68(4),48-53.
2. Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection. In S. S. Iyengar, & V. Saxena (Eds.), *2019 Twelfth International Conference on Contemporary Computing (IC3)* [8844879] (International Conference on Contemporary Computing). IEEE.
3. Mehboob, A., & Malik, M. S. I. (2021). Smart fraud detection framework for job recruitments. *Arabian Journal for Science and Engineering*, *46*(4), 3067-3078.
4. Anita, C. S., Nagarajan, P., Sairam, G. A., Ganesh, P., & Deepakkumar, G. (2021). Fake job detection and analysis using machine learning and deep learning algorithms. Revista Geintec-Gestao Inovacao e Tecnologias, 11(2), 642-650.
5. Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V., & Ludi, S. (2022). Detection of fake job postings by utilizing machine learning and natural language processing approaches. Neural Processing Letters, 1-29