e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:07/Issue:04/April-2025 Impact Factor- 8.187 www.irjmets.com

ONLINE RECRUITMENT FRAUD (ORF) DETECTION USING DEEP LEARNING APPROACHES

Dr. R. Gopi[*1], Sanshay Kumar Tiwary[*2], Raushan Kumar[*3], Suraj Kumar[*4], Sanni K Ranjan Kumar[*5]

[*1]Assistant Professor, Department Of Computer Science- Dhanalakshmi Srinivasan Engineering College (Autonomous), India.

[*2,3,4,5]Student, Department Of Computer Science-Dhanalakshmi Srinivasan Engineering College (Autonomous), India.

DOI: https://www.doi.org/10.56726/IRJMETS73923

ABSTRACT

Most companies nowadays are using digital platforms for the recruitment of new employees to make the hiring process easier. The rapid increase in the use of online platforms for job posting has resulted in fraudulent advertising. The scammers are making money through fraudulent job postings. Online recruitment fraud has emerged as an important issue in cybercrime. Therefore, it is necessary to detect fake job postings to get rid of online job scams. In recent studies, traditional machine learning and deep learning algorithms have been implemented to detect fake job postings; this research aims to use two transformer-based deep learning models, i.e., Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT-Pre training Approach (RoBERTa) to detect fake job posting sprecisely. In this research, anovel dataset of fake job postings is proposed, formed by the combination of job postings from three different sources. Existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the existing models' capability in detecting fraudulent jobs. Hence, we extend it with the latest job postings. Exploratory Data Analysis (EDA) highlights the class imbalance problem in detecting fake jobs, which tends the model to act aggressively toward the minority class. Responding to overcome this problem, the work at hand implements ten top-performing Synthetic Minority Oversampling Technique (SMOTE) variants. The models' performances balanced by each SMOTE variant are analyzed and compared. All implemented approaches are performed competitively. However, BERT+SMOBD SMOTE achieved the highest balanced accuracy and recall of about 90%.

Keywords: Class Imbalance, Data Augmentation, Deep Learning, Employment Scam, Fraud Detection, Machine Learning, Online Recruitment, SMOTE, Transformer-Based Models.

I. INTRODUCTION

In the age of advanced technology, the internet has revolutionized nearly every aspect of daily life, including how people search for jobs and how companies recruit talent. Traditional recruitment methods have largely shifted to online platforms, offering a more efficient and accessible way to connect employers with potential candidates. This shift has given rise to online recruitment systems, commonly referred to as **E-recruitment**, which allow job seekers to browse and apply for positions with ease.

However, alongside these benefits, the rise of E-recruitment platforms has also opened the door to a growing cybercrime threat: **online recruitment fraud**. Fraudsters exploit these platforms by posting fake job advertisements to deceive applicants for financial gain or to collect sensitive personal information. These scams often appear legitimate, making them difficult to detect through manual review alone. As a result, there is a pressing need for intelligent and automated systems capable of identifying fraudulent job postings effectively.

Previous research has employed traditional machine learning and some deep learning models to tackle this problem. While these approaches have shown promise, they often lack the depth of contextual understanding required to accurately analyze the complex language used in job descriptions. Furthermore, many of these models were trained on outdated or domain-limited datasets, which restrict their ability to generalize to modern and diverse job postings.

To address these limitations, this study introduces a deep learning-based framework using two state-of-the-art transformer models: Bidirectional Encoder Representations from Transformers (BERT) and Robustly

Optimized BERT Pretraining Approach (RoBERTa). These models have demonstrated superior performance in a wide range of Natural Language Processing (NLP) tasks due to their ability to capture deep semantic and contextual relationships in text.

Each transformer model was trained and evaluated using the balanced datasets produced by the SMOTE variants. Performance was assessed using balanced accuracy, precision, recall, and F1-score. Among all approaches, the combination of **BERT with the SMOBD (Borderline-SMOTE1) variant** achieved the best results, with approximately **90% balanced accuracy and recall**, indicating a strong ability to identify fraudulent job advertisements.

II. DATASET DESCRIPTION

The dataset used in this study is a curated collection of job posting records gathered from multiple online job portals. It includes both legitimate and fraudulent job advertisements, providing a comprehensive basis for developing and evaluating models for fake job detection. The presence of fake postings within the dataset enables supervised learning techniques to effectively differentiate between real and deceptive listings.

Each record in the dataset consists of several attributes that offer rich contextual information about the job posting. The key features include:

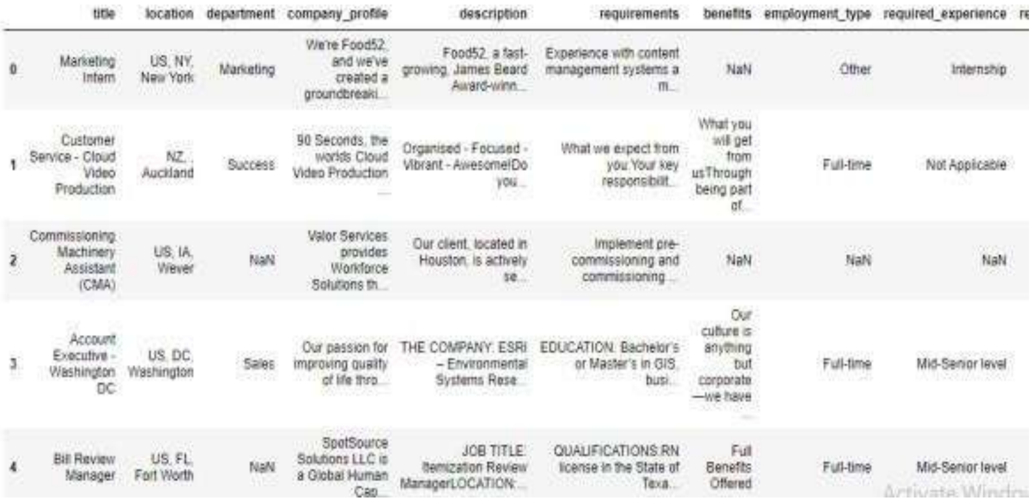
- ✓ **ID**: A unique identifier for each job posting.
- ✓ **Job Title**: The title or designation of the job position.
- ✓ **Location**: The geographic location where the job is based.
- ✓ **Description**: A detailed narrative of the job role and responsibilities.
- ✓ **Requirements**: The qualifications and skills required for the job.

This dataset provides both structured and unstructured features. Text fields such as **description**, **requirements**, and **employer profile** are particularly valuable for deep learning models like BERT and RoBERTa, which excel at interpreting natural language data.

Language Used: Python

III. DATA IMPORTING

In the initial phase of the Job Scam Detection project, importing the dataset is a fundamental and critical step. It lays the foundation for building, training, and evaluating machine learning models aimed at identifying fraudulent job postings. The dataset, stored in **CSV (Comma-Separated Values)** format, contains a rich and diverse collection of job listings gathered from multiple online recruitment platforms, including both legitimate and fake entries.

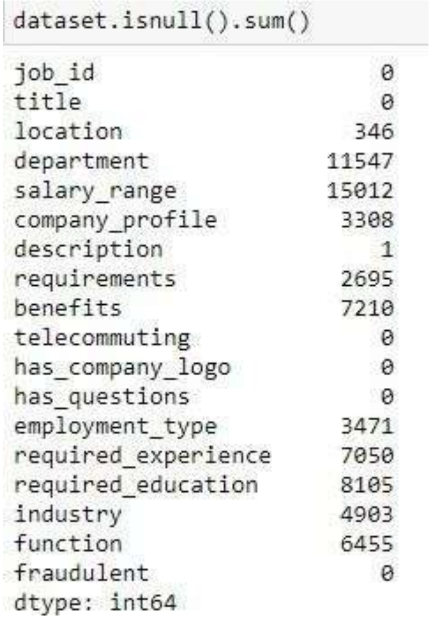


	title	location	department	company_profile	description	requirements	benefits	employment_type	required_experience	re
0	Marketing Intern	US, NY, New York	Marketing	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN	Other	Internship	
1	Customer Service - Cloud Video Production	NZ, , Auckland	Success	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - Awesome!Do you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...	Full-time	Not Applicable	
2	Commissioning Machinery Assistant (CMA)	US, IA, Wever	NaN	Valor Services provides Workforce Solutions th...	Our client, located in Houston, is actively se...	Implement pre-commissioning and commissioning ...	NaN	NaN	NaN	
3	Account Executive - Washington DC	US, DC, Washington	Sales	Our passion for improving quality of life thro...	THE COMPANY: ESRI – Environmental Systems Rese...	EDUCATION: Bachelor's or Master's in GIS, busi...	Our culture is anything but corporate—we have ...	Full-time	Mid-Senior level	
4	Bill Review Manager	US, FL, Fort Worth	NaN	SpotSource Solutions LLC is a Global Human Cap...	JOB TITLE: Itemization Review ManagerLOCATION:...	QUALIFICATIONS:RN license in the State of Texa...	Full Benefits Offered	Full-time	Mid-Senior level	

Fig 1: Data Importing

IV. DATA CLEANING

Following the successful import of the dataset, the next crucial step in the Job Scam Detection pipeline is **data cleaning.** This phase is essential to ensure the integrity, reliability, and accuracy of the data used to train machine learning and deep learning models. A clean and consistent dataset not only improves model performance but also reduces the risk of biased or misleading results.



```
dataset.isnull().sum()

job_id                     0
title                      0
location                 346
department             11547
salary_range           15012
company_profile         3308
description                1
requirements            2695
benefits                7210
telecommuting              0
has_company_logo           0
has_questions              0
employment_type         3471
required_experience     7050
required_education      8105
industry                4903
function                6455
fraudulent                 0
dtype: int64
```

Fig 2: Data Cleaning

V. DATA VISUALIZATION

the Job Scam Detection project advances to an essential step—**data visualization.** This phase transforms raw data into intuitive and informative visual formats that reveal underlying patterns, correlations, and distributions within the dataset. Visualization not only enhances understanding but also guides the development of machine learning models by highlighting the most relevant features.

Using Python's powerful visualization libraries—**Matplotlib** and **Seaborn**—a variety of plots and charts are generated to explore the structure and nature of the job postings data. These visualizations play a pivotal role in distinguishing between fraudulent

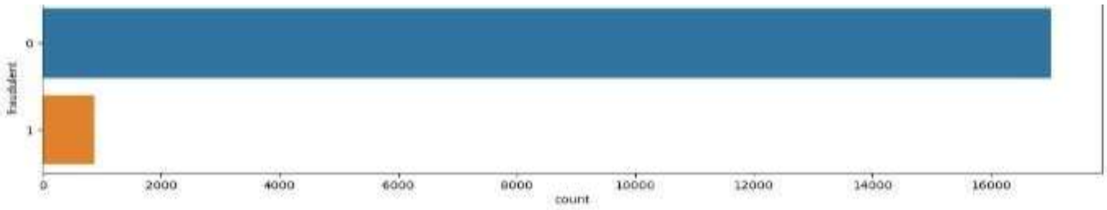


Fig 3: Fraudulent Postings in the Dataset

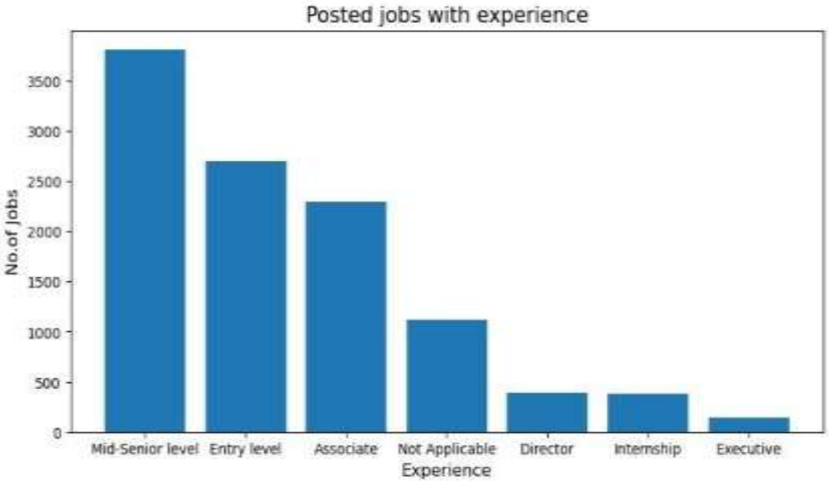


Fig 4: Posted Job with Experience

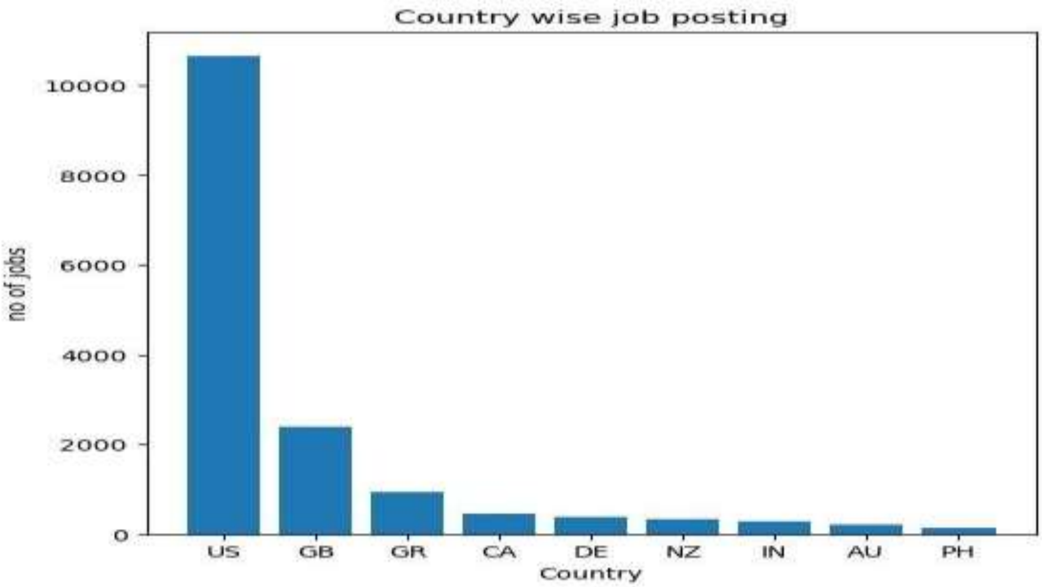
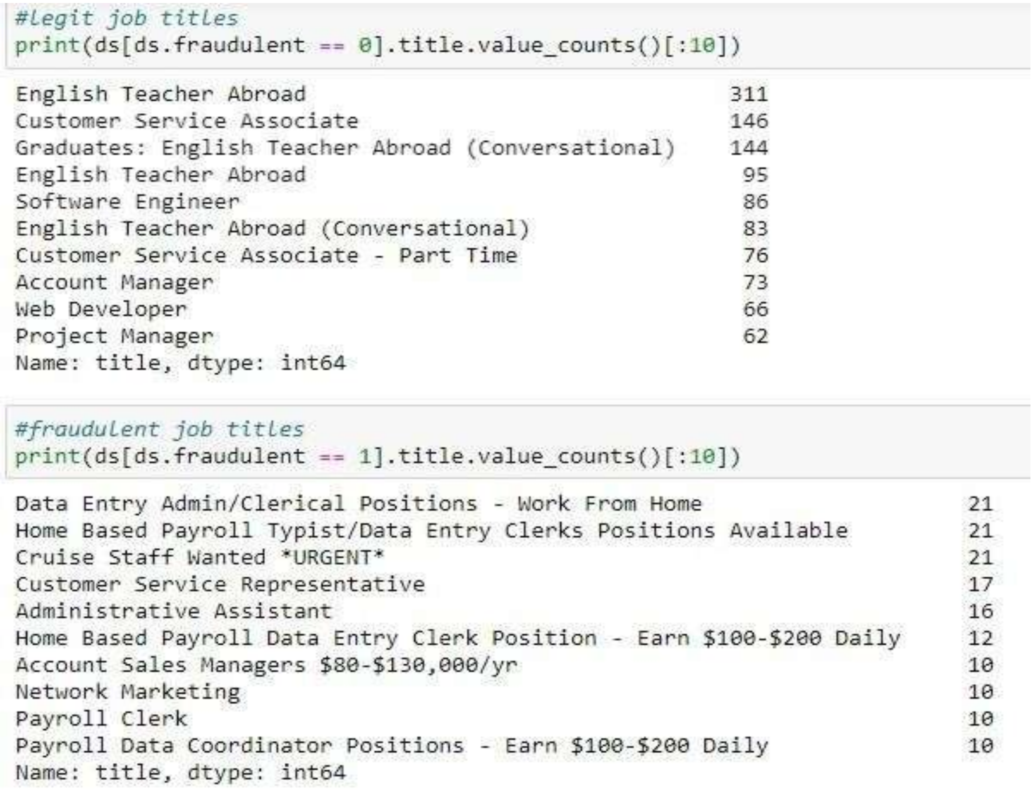


Fig 5: Country wise Job Posting



```
#legit job titles
print(ds[ds.fraudulent == 0].title.value_counts()[:10])
```

```
English Teacher Abroad                                311
Customer Service Associate                            146
Graduates: English Teacher Abroad (Conversational)    144
English Teacher Abroad                                 95
Software Engineer                                      86
English Teacher Abroad (Conversational)                83
Customer Service Associate - Part Time                 76
Account Manager                                        73
Web Developer                                          66
Project Manager                                        62
Name: title, dtype: int64
```

```
#fraudulent job titles
print(ds[ds.fraudulent == 1].title.value_counts()[:10])
```

```
Data Entry Admin/Clerical Positions - Work From Home                 21
Home Based Payroll Typist/Data Entry Clerks Positions Available      21
Cruise Staff Wanted *URGENT*                                         21
Customer Service Representative                                      17
Administrative Assistant                                             16
Home Based Payroll Data Entry Clerk Position - Earn $100-$200 Daily  12
Account Sales Managers $80-$130,000/yr                               10
Network Marketing                                                    10
Payroll Clerk                                                        10
Payroll Data Coordinator Positions - Earn $100-$200 Daily            10
Name: title, dtype: int64
```

Fig 6: Legit and Fraudulent Job Titles

VI. WORD CLOUD

In the realm of Job Scam Detection, **word clouds** serve as a powerful visual tool for extracting and understanding key textual patterns from job postings. After the data cleaning process, word clouds are used to highlight the most frequently occurring words within the dataset, offering immediate insights into the language commonly associated with both **legitimate** and **fraudulent** job listings.

This technique aids in visually distinguishing between scam and genuine job offers based on the vocabulary used. For instance, fraudulent job postings often include persuasive language, vague role descriptions, or too-good-to-be-true incentives, which can be captured through the frequency of certain terms.

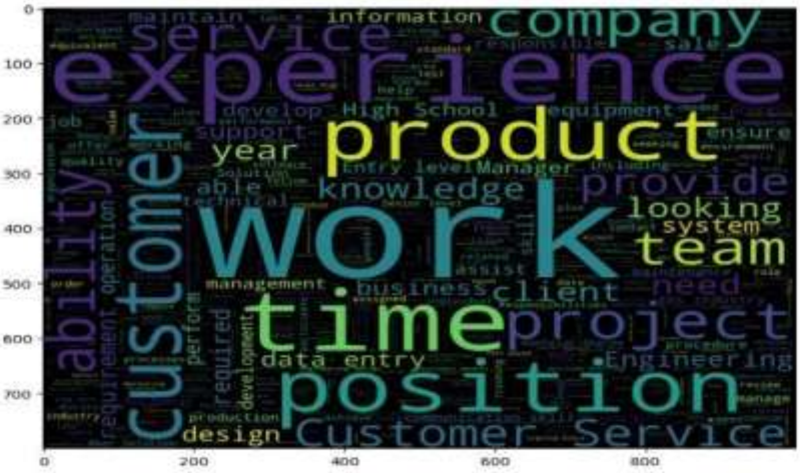


Fig 7: Word Cloud of Fraudulent Jobs

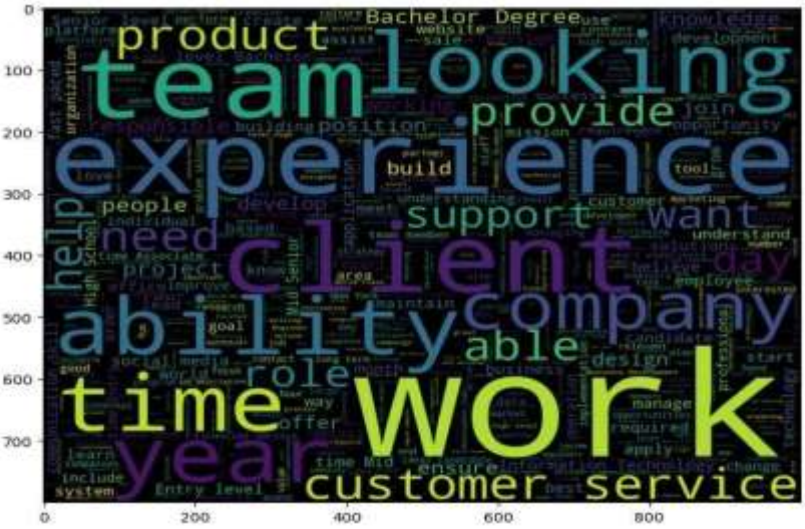
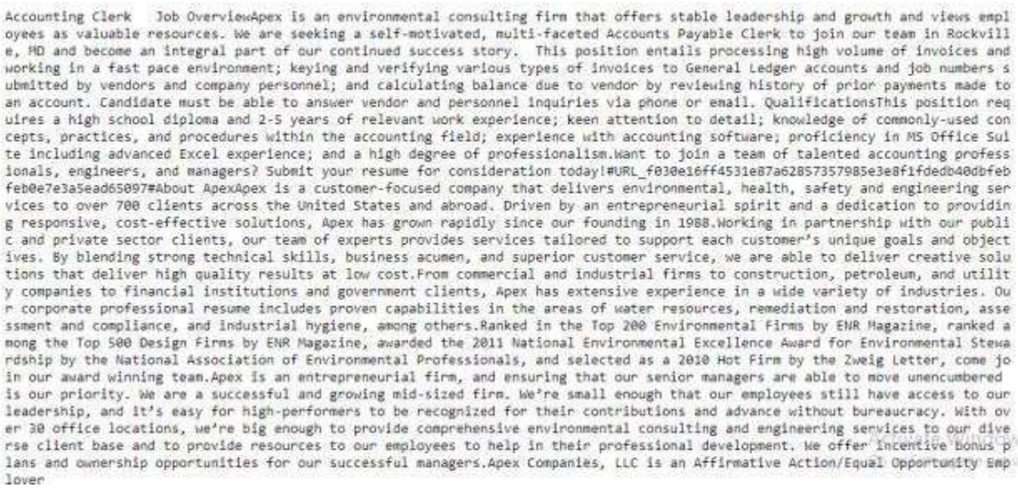


Fig 8: Word Cloud of Real Jobs

VII. PRE-PROCESSING

In the Job Scam Detection project, **data pre-processing** is a foundational step that significantly influences the performance and reliability of machine learning models. The goal of this phase is to transform raw, inconsistent, and unstructured data into a clean, uniform, and well-formatted structure suitable for effective training and evaluation.

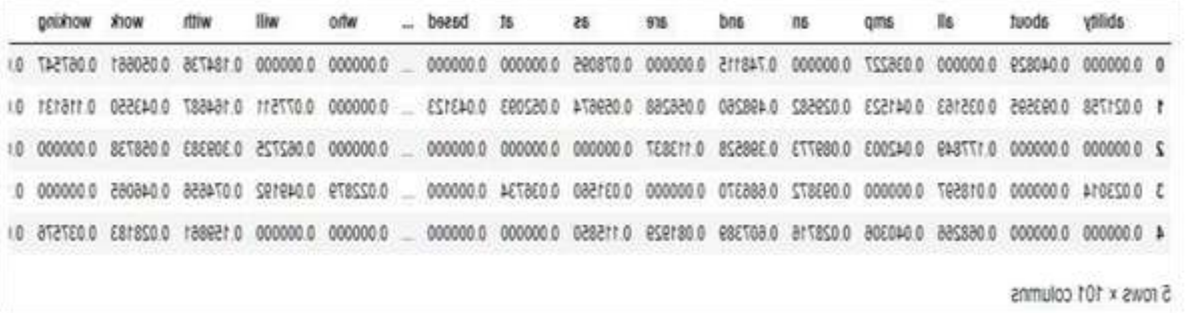


```
Accounting Clerk   Job OverviewApex is an environmental consulting firm that offers stable leadership and growth and views empl
oyees as valuable resources. We are seeking a self-motivated, multi-faceted Accounts Payable Clerk to join our team in Rockvill
e, MD and become an integral part of our continued success story.  This position entails processing high volume of invoices and
working in a fast pace environment; keying and verifying various types of invoices to General Ledger accounts and job numbers s
ubmitted by vendors and company personnel; and calculating balance due to vendor by reviewing history of prior payments made to
an account. Candidate must be able to answer vendor and personnel inquiries via phone or email. QualificationsThis position req
uires a high school diploma and 2-5 years of relevant work experience; keen attention to detail; knowledge of commonly-used con
cepts, practices, and procedures within the accounting field; experience with accounting software; proficiency in MS Office Sui
te including advanced Excel experience; and a high degree of professionalism.Want to join a team of talented accounting profess
ionals, engineers, and managers? Submit your resume for consideration today!#URL_f030e16ff4531e87a62857357985e3e8f1fdedb40dbfeb
feb0e7e3a5ead65097#About ApexApex is a customer-focused company that delivers environmental, health, safety and engineering ser
vices to over 700 clients across the United States and abroad. Driven by an entrepreneurial spirit and a dedication to providin
g responsive, cost-effective solutions, Apex has grown rapidly since our founding in 1988.Working in partnership with our publi
c and private sector clients, our team of experts provides services tailored to support each customer's unique goals and object
ives. By blending strong technical skills, business acumen, and superior customer service, we are able to deliver creative solu
tions that deliver high quality results at low cost.From commercial and industrial firms to construction, petroleum, and utilit
y companies to financial institutions and government clients, Apex has extensive experience in a wide variety of industries. Ou
r corporate professional resume includes proven capabilities in the areas of water resources, remediation and restoration, asse
ssment and compliance, and industrial hygiene, among others.Ranked in the Top 200 Environmental Firms by ENR Magazine, ranked a
mong the Top 500 Design Firms by ENR Magazine, awarded the 2011 National Environmental Excellence Award for Environmental Stewa
rdship by the National Association of Environmental Professionals, and selected as a 2010 Hot Firm by the Zweig Letter, come jo
in our award winning team.Apex is an entrepreneurial firm, and ensuring that our senior managers are able to move unencumbered
is our priority. We are a successful and growing mid-sized firm. We're small enough that our employees still have access to our
leadership, and it's easy for high-performers to be recognized for their contributions and advance without bureaucracy. With ov
er 30 office locations, we're big enough to provide comprehensive environmental consulting and engineering services to our dive
rse client base and to provide resources to our employees to help in their professional development. We offer incentive bonus p
lans and ownership opportunities for our successful managers.Apex Companies, LLC is an Affirmative Action/Equal Opportunity Emp
loyer
```

Fig 9: Pre-Processing

VIII. MODELLING

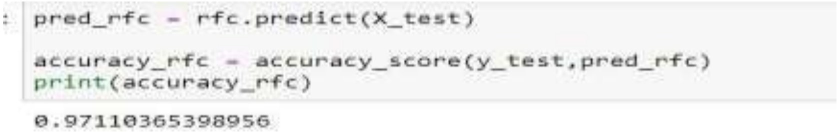
The **modelling phase** is the core of the Job Scam Detection project, where advanced machine learning algorithms are employed to classify job postings as either **fraudulent** or **legitimate** based on patterns learned from the preprocessed dataset. This phase translates insights derived from exploratory data analysis and feature engineering into actionable predictions.



	ability	about	all	amp	an	and	are	as	at	based	...	who	will	with	work	working
0	0.000000	0.040829	0.000000	0.036227	0.000000	0.748115	0.000000	0.078095	0.000000	0.000000	...	0.000000	0.000000	0.184736	0.050661	0.061547
1	0.021758	0.093595	0.035163	0.041523	0.029582	0.498260	0.056268	0.059674	0.052093	0.043123	...	0.000000	0.077511	0.164587	0.043350	0.116131
2	0.000000	0.000000	0.177849	0.042003	0.089773	0.398528	0.113837	0.000000	0.000000	0.000000	...	0.000000	0.062725	0.309383	0.058738	0.000000
3	0.023014	0.000000	0.018597	0.000000	0.093872	0.686370	0.000000	0.031560	0.036734	0.000000	...	0.022879	0.049192	0.074656	0.046095	0.000000
4	0.000000	0.000000	0.068296	0.040306	0.028716	0.607389	0.081929	0.115850	0.000000	0.000000	...	0.000000	0.000000	0.159861	0.028183	0.037576

5 rows × 101 columns

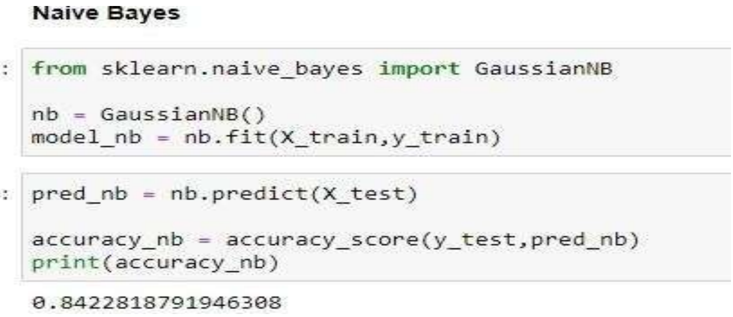
Fig 10: Modelling



```
pred_rfc = rfc.predict(X_test)

accuracy_rfc = accuracy_score(y_test,pred_rfc)
print(accuracy_rfc)
```

```
0.97110365398956
```

Fig 11: Final Model Random Forest Accuracy Report



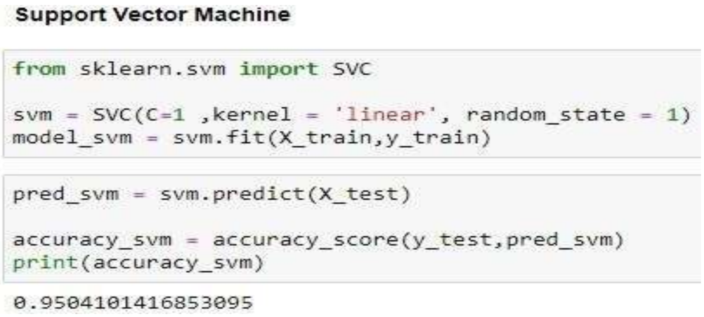
Naive Bayes

```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()
model_nb = nb.fit(X_train,y_train)
```

```
pred_nb = nb.predict(X_test)

accuracy_nb = accuracy_score(y_test,pred_nb)
print(accuracy_nb)
```

```
0.8422818791946308
```

Fig 12: Final Model Naive Bayes Accuracy Report



Support Vector Machine

```
from sklearn.svm import SVC

svm = SVC(C=1 ,kernel = 'linear', random_state = 1)
model_svm = svm.fit(X_train,y_train)
```

```
pred_svm = svm.predict(X_test)

accuracy_svm = accuracy_score(y_test,pred_svm)
print(accuracy_svm)
```

```
0.9504101416853095
```

Fig 13: Final Model Support Vector Machine Accuracy Report

IX. CONCLUSION

This research addresses the **pressing issue of Online Recruitment Fraud (ORF)** by designing and implementing a machine learning-based Job Scam Detection system. With the rising number of job seekers and the increasing shift toward online hiring platforms, scammers are exploiting digital spaces through deceptive job postings. This study responds to that threat by leveraging machine learning algorithms capable of learning complex patterns and distinguishing between legitimate and fraudulent listings.

Throughout the project, a structured workflow was followed—starting with the careful **importing and pre-processing** of a diverse dataset collected from multiple sources. Pre-processing included handling missing values, outliers, categorical encoding, text normalization, and feature scaling, ensuring that the dataset was optimized for learning and pattern recognition. Furthermore, **exploratory data analysis (EDA)** and **visualization techniques** such as bar plots and word clouds played a key role in understanding the data, detecting trends, and highlighting the presence of biases or imbalances.

X. REFERENCES

[1] B. Alghamdi and f. Alharbi, "an intelligent model for online recruitment fraud detection," journal of information security, vol. 10, no. 03, pp. 155–176, 2019. Doi: 10.4236/jis.2019.103009

[2] X. Zhang and v. S. Sheng, "an empirical study of the naive bayes classifier," international journal of computer applications, pp. 41–46, 2014.

[3] D. E. Walters, "bayes's theorem and the analysis of binomial random variables," biometrical journal, vol. 30, no. 7, pp. 817–825, 1988. Doi: 10.1002/bimj.4710300710

[4] B. Biggio, i. Corona, g. Fumera, g. Giacinto, and f. Roli, "bagging classifiers for fighting poisoning attacks in adversarial classification tasks," lecture notes in computer science, vol. 6713, pp. 350–359, 2011. Doi: 10.1007/978-3-642-21557-5_37

[5] S. M. Vieira, u. Kaymak, and j. M. C. Sousa, "cohen's kappa coefficient as a performance measure for classification models," proceedings of the 2010 ieee world congress on computational intelligence (wcci), 2010. Doi: 10.1109/fuzzy.2010.5584447

[6] B. Biggio, i. Corona, g. Fumera, g. Giacinto, and f. Roli, "bagging classifiers for fighting poisoning attacks in adversarial classification tasks," lecture notes in computer science, vol. 6713, pp. 350–359, 2011. Doi: 10.1007/978-3-642-21557-5_37 (duplicate entry removed or merged)

[7] N. Hussain, h. T. Mirza, g. Rasool, i. Hussain, and m. Kaleem, "spam review detection techniques: a systematic literature review," applied sciences, vol. 9, no. 5, pp. 1–26, 2019. Doi: 10.3390/app9050987