National College of Ireland

APPLICATION OF NATURAL LANGUAGE PROCESSING FOR FAKE JOB POSTING DETECTION

Natural Language Processing

MSc Research Project

Vaishnavi Gadikota

Student ID: 23393777

School of Computing

National College of Ireland

Supervisor: Arjun Chikkankod

National College of Ireland
MSc Project Submission Sheet
School of Computing

Student Name: ……. ……vaishnavi gadikota ……

Student ID: ……23393777……

Programme: ……MSc Data Analytics …… Year: …2025-2026…

Module: ……Research practicum part-2……

Supervisor: ……Arjun Chikkankod……

Submission Due Date: ……28/01/26……

Project Title: Application of natural language processing for fake job posting detection
Natural Language Processing

Word Count: ……

Page Count……..…….

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……vaishnavi gadikota……

Date: ……28/01/26……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Abstract

The growing trend of counterfeit job ads on online job market sites has posed serious threats to job seekers and has drastically caused a loss of faith in online job markets. These fraudulent postings have not always had identifiable organisations, have been in imprecise language and have sought to gather personal data or deceive applicants. This project has managed this emerging threat by coming up with a Natural Language Processing (NLP)-based fake job posting detection system that can determine deceptive advertisement through linguistic and metadata-based patterns. The study aimed at evaluating NLP and machine learning models to detect fraudulent job offers, determining linguistic cues associated with deceit, and comparing the model performance to improve the classification. The research has chosen a secondary quantitative research approach that has been supported by a deductive and a positivism method, and has used the Kaggle "Real or Fake Job Posting Prediction" dataset with 17,880 job posts. Data were collected by importing the dataset and performing extensive preprocessing, which includes text cleaning, lemmatisation, stop-word removal, filtering rare-word removal and TF-IDF vectorisation. Further, the study has been exploratory data analysis, visualisation, class balancing through SMOTE, and training various models, including Logistic Regression, Naive Bayes, XGBoost, and LSTM. The strongest results have been attained with the Logistic Regression model, which exhibit high recall and F1-scores, especially with respect to the minority fraud group. The study has established that the NLP methods have been fruitfully used in order to extract meaningful linguistic patterns that can be used to identify frauds.

Chapter 1 Introduction

1.1 Introduction

The rising population in the world has created a rising demand in job demands which has been prevalent primarily to the huge advancement in modern technology like manufacturing and assembling lines. The job market has been hugely affected by the notoriously huge amount of fake job postings on online platforms like social connection applications like LinkedIn, Fiverr and many more. A report by Davis, 2025 found out that the US has been spammed with fake job listings on LinkedIn amounting to around 27.4%. Among the location centred analysis in the U.S. Los Angeles has been ranked at the top with 30.5% fake listings.

1.2 Overview

The fake job postings phenomenon has grown so much in the last year that 1 out of every 5 job listings across all major platforms is fake (Kesslen, 2025). The identified industry with the most amount of fake job listings was the construction industry where the number of fake listings was over 38%. The main problem with the fake listings is that the applicants must face misleading details and a sense of false hope where it has been found out that the details of the fake listing have been found out to be highly exploitative and sometimes outrageous. These listings are aimed to fool applicants and get their personal data or sometimes asking for advanced payments. The fraudulent activities range from just the political standpoint of organisations to full blown cybersecurity problems like phishing and bullying. The enormous number of fake listings has spammed the online job listing for too long which has resulted in almost 0.8 unemployed people per number of job times as found out through the report by (Thapa, 2024). The identification and avoiding these fake listings are a concern for individuals looking to support their livelihood in all sectors across all places, this forms the basis of this undertaking.

1.3 Research Rationale

The rationale for this study lies within the growing necessity for protecting the job seekers and maintaining trustworthiness of the online platforms for recruitment. The postings make use of the availability and reach of the platforms to reach a vast amount of people ranging from diverse types of people across the board including both employed and unemployed people. The main problem lies with the fake job identification process which majorly depends on the look and feel of the job posting rather than the technical details because job seekers are emotionally attracted to the job postings and when they come across a post with a trending company, they are eager for entrance into the organization (Jaison and Kodabagi, 2023). This makes the applicants fall into the trap of either profile hungry companies looking for market attraction rather than hiring or cybercriminals aiming to steal personal information. The financial implication of fake job postings is majorly done by huge businesses who tend to undermine the work of the current employees by showing demand and for poaching few highly skilled individuals from competitors. Rather than the financial implications the psychological implications are also present which directly impact the unemployed individuals actively looking for jobs. When a user interacts with a fake listing this consumes the time and energy with little to no positive outcome which may result in losing self confidence and trust in the job application process. The previous studies or research papers have represented basic classification techniques and keyword-based approaches. There are a few papers which have effectively used ML and advanced NLP techniques for fake job detection (Amaar et al., 2022). The main concerned type of fake jobs is first, done by companies to attract market capitalists for

investments and second, by cybercriminals who are looking to defraud personal details of people. These parties change their methods continuously and can after some time not be detected by traditional fake job listings.

1.4 Problem Statement

The digital transformation of the IT industry has brought around major changes in the information gain sector which directly leads to the easy availability and usability of online applications for daily necessities and platforms for other means for shopping and social connections. This has also attracted notoriously motivated people to defraud others for their own personal gain, which is a major problem this project aims to deal with. The traditional or conventional rule-based and manual moderation systems are struggling to keep pace with a huge amount of job posting advertisements released daily on several platforms and the constantly evolving techniques used for such postings. These fraudulent publications hamper credibility of the recruitment platforms and sometimes lead to privacy and financial risks for the applicants which has proven to be a major security breach to the point of losing confidence in such organizations and the credibility of their employment opportunities. There is a need of a sophisticated and robust technique for the identification of the fake job listings by not just text-word and feature based but a metadata related technique that will not just classify the listing in a binary manner but can handle the underlying hidden patterns using Natural Language Processing (NLP) techniques.

1.5 Research Aim

This research aims to assess Natural Language Processing (NLP) methods to identify fake job postings by creating and testing classification models, contributing to reducing the risk of employment fraud.

1.6 Research Objectives

- To perform exploratory data analysis on the data set to see the distribution and nature of real and false job postings.
- To pre-process and convert text and metadata features into usable forms for application in other NLP-based classification models.
- To compare and enhance the performance of machine learning models for identifying spurious job postings.
- To discover important linguistic patterns, features, or indicators commonly found in fake job advertisements with the help of feature importance and model interpretation methods.

1.7 Research Hypothesis

Null Hypothesis (H0): There are no presence of significant linguistic patterns that can be identified and used to determine the credibility of a job posting. Analysing Metadata using Natural Language Processing will not result in models which can accurately identify fake from real job postings.

Alternate Hypothesis (H1): There are hidden linguistic patterns which on identification can be used differentiate between genuine job opportunities and fake job postings. Using Natural Language Processing to analyse metadata will result in the creation of a powerful model which will accurately classify a job listing as genuine or fake.

1.8 Dependent and Independent Variables

Table 1: Variables

Type	Variable name/s

Independent Variable	"title", "company_profile", "description", "requirements", "benefits",”employment_type,” “required_experience,” “required_education,” “industry,” “function”
Dependent Variable	“fraudulent”

1.9 Dissertation Structure

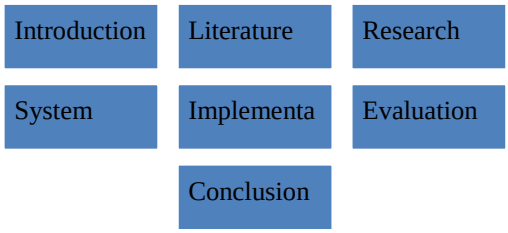


Figure 1: Dissertation structure

Chapter 2 Literature Review

2.1 Factors leading to fraudulent job posting

Fraud in online recruitment happens due to lack of firm identifiers, weak proof of content and improper job descriptions. Thus, Nanath and Olney (2023), aimed for improving detection of fake jobs by implementing content features like presence or absence of company profile or log with human signals. Implementing EMSCAD subsets and four human signs, it was observed that profile and logo of a company is the strong prediction and net-promoter judgment style are the key factors leading to fake posting of jobs. However, direct-flag questions and fuzzy-logic were not key factors due to response-burden and bias. Problems include high sensitivity of crowdsourcing bias and design shown by people during job postings. Prasanna et al. (2022), however had proposed a framework which is semi-supervised with four lightweight detectors that helps in operationalising key factors of fraud include near-duplicate postings, domains which are blocked, trusted posters and residuals multi-classifier for cases which are ambiguous. The aim was to do real-time filtering of job posts for which mixed rule-based filters at cluster or poster level were used with supervised classification. The results showed that the above-mentioned factors were key reasons leading to fake jobs; however, lack of external validation and comparative analysis was not properly explored, thus questioning the credibility of the research. Reddy et al. (2022), likewise compare Decision Tree, KNN, Naive Bayes, MLP, SVM, DNN and Random Forest on EMSCAD posts for predicting fraud. After proper preprocessing, DNN achieved 98% accuracy, ensemble models also performed well. The critical evaluation beyond accuracy revealed that fake URLs, lack of proper job description along with company profile and logo are key indicators of fake jobs posting. However, limitations include dependency on one dataset which might hamper generalisation and reduce feature level analysis.

Lakshmi et al. (2024), moreover had developed a Fake job posting prediction system by using SVM and Random Forest with hyperparameter tuning, reported high accuracy and stating that factors like request for fees for getting jobs, aggressive collection of personal and bank related information and unsafe method of hiring are the major factors of fake job posting. The research provided the balanced ROC-AUC of SVM whole RF had better classification of job posting showing trade - off between the metrics. The limitations include less depth in benchmarking and external validity leading to reduced generalisation. Thus, based on the analysis, key factors including profiles, logos, duplication of content, domains and less solicitation patterns are what people need to look at while searching for real jobs.

2.2 Impact of fake job posting on organisations

Fake posting of jobs tends to hamper organisations by corrupting the pipeline of talents, hampering the trust of the brand, and disrupting the visibility of the hiring market. Thus, Yamashita, Tran and Lee

(2024) had aimed for exposing platform level harms based on analysing fake resumes and data poisoning with the help of black-box surrogate and FRANCIS generator. The research developed prediction models which depend on company-promotion/demotion and user promotion attaches. Empirically, even injections at a small level allows recruiters and candidates to see the impact which tends to harm the visibility of firms, downstream the actions of HR and prevents in understanding the quality of candidates. Impacts on organisations happen leading to decline of hiring leading the performance to hamper. Limitation of this research is related to transfer to proprietary victim models and lack of sectoral analysis. V and Vishvanath (2025), however, aimed at providing users protections and improving the credibility of platforms based on ML detectors which are lightweight. For this TF-IDF features along with Random Forest, SVM and Logistic Regression were used along with real-time prediction. Findings showed that Random Forest achieved 94% accuracy and user study showed that improved trust by reducing frauds, organisations get benefits like moderate workload, less reputational incidents, and accurate applications. However, problems of this research include dependency on a single dataset, lack of data balancing and uncertain cross-platform generalisation.

B et al. (2023), similarly aimed at comparing data mining classifiers for predicting fake jobs. For this, models like SVM, Random Forest, KNN, Decision Tree, Naive Bayse, DNN and MLP were implemented where DNN achieved 98% accuracy. This shows that proper screening of job posts helps organisations to reduce the time of recruitment in screening fake profiles and restore the confidence of organisations which is important for organisations for hiring talents. However, the feature design of the research had avoided NLP analysis, using one dataset leading to a lack of analysis of sectors and changing frauds. Hamza et al. (2021), conversely aimed at linking the practice of selection to performance of organisations by implementing quantitative survey analysis for around 60 recruiters in the telecom sector. The findings showed that selection methods like interviews, assessment centres, references, and tests properly align with effectiveness, financial viability, efficiency, and relevance, showing that fake postings tend to hamper the application pool, degrade trusts leading performance of organisations to adversely hamper. Limitations include analysis of single sector and small sample size hampering generalisation. Thus, based on the analysis, it can be observed that impact on organisation extends beyond reputational harm to improper hiring and cost of time in screening leading to negative performance for which proper detection models are needed for making hiring process smooth.

2.3 Role of Machine learning in detecting fake job postings

Machine learning is becoming a key technology for detecting fake job posts since the models can learn text and metadata at scale. Thus, Pillai (2023) had aimed to evaluate deep sequence modelling with tree-based ensemble models. For this Kaggle's 17880 post corpus dataset was used with Text and Count Vectorisation with Bi-directional LSTM with Random Forest, Gradient Boosting and LightGBM. The result showed that Bi-LSTM's accuracy was 98.71% with 0.91 as ROC-AUC, stating implementation of bidirectional context has more signalling gains in comparison to bag-of-word trees. However, problems include evaluation of single dataset, sensitivity in single dataset and improper cross-platform validation. Anbarasu, Selvakani and Vasumathi (2024), on the other hand aimed for operationalising a detector which is lightweight capable of doing real-time screening of recruitment

processes. For this SVM, Logistic Regression and Random Forest were implemented with TF-IDF features within a single triage interface. The research showed that Random Forest had 98% accuracy, stating that non-deep text classifications can provide precision which are highly appropriate while reducing manual review. The model highlighted key factors like missing identifiers and unrealistic salaries are used for fake job postings. Limitations include dependency on a single dataset, lack of proper reporting of recall, precision and F1 and no ablation analysis leading to less robustness.

Amaar et al. (2022), moreover aimed at comparing classical ML models with high imbalance paired BoW/TF-IDF with ADASYN and implemented six models (SVM, MLP, KNN, Random Forest, Logistic Regression and Extra Trees). The dataset used was Kaggle's 17000 job post data and was evaluated using accuracy, recall, F1 and precision. The results showed that ADASYN+TF-IDF, Extra Trees had achieved 99.9% accuracy without class balancing leading overfitting to happen. Limitations included less external validation, improper cost-error analysis, and oversampling - induced optimism thus reducing credibility. Afzal et al. (2023), conversely, had implemented SMOTE, PCA/Chi-square methods of feature selection along with voting classifier using Logistic Regression, Extra Trees, and Random Forest on the same dataset of Kaggle allowing to achieve 99% accuracy along with k-fold optimisation. The limitations include accuracy being more important than interpretation and calibration of the models in real-time analysis. Thus, based on the analysis, fine - tuned ensemble and deep sequence models integrated with BoW/TF-IDF text features, proper feature reduction and oversampling of minority class helps in achieving accuracy up to 99%. However, real benefits for companies depend on validating F1/precision and recall, resilience on drift and calibration on all platforms, otherwise high accuracy might not always provide high detection.

2.4 Theoretical Framework

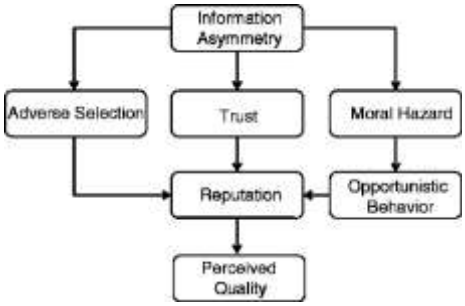


Figure 2: Information Asymmetry theory

(Source: Devos, Landeghem and Deschoolmeester, 2011)

The research has implemented Information Asymmetry Theory as the theoretical aspect on which this research is based. According to Akerlof (1970), this theory helps in explaining situations where one party has more or proper information than the other, causing improper transactions, inadequate selection, and high exploitation. The theory is highly applicant in fake job posting detection as fraudsters have high information regarding the falsity of the posting which is not known to job seekers. This type of asymmetry allows fraudsters to exploit trust, gain access to sensitive information and mislead candidates. This theory helps in understanding the importance of transparency, safeguard applicants, and reputation of organisations against fraud using latest technologies.

2.5 Conceptual Framework

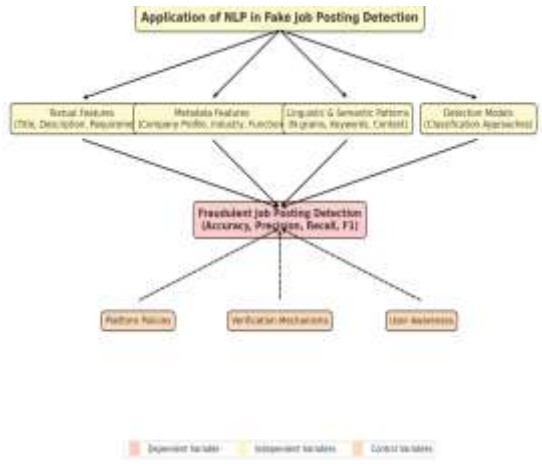


Figure 3: Conceptual Framework

2.6 Gaps

Even though much research has been done exploring fake job posting detection using ML and NLP there exist some gaps. Most of the studies give importance to accuracy (technical) but do not provide importance to real-world application across industries and platforms. There is also a problem in proper exploration of how ML models are responding in changing tactics of fraud in this field, creating problems in long-term application. Moreover, less integration of organisational factors like user awareness, policies and verification systems tend to reduce the application of the models in practice. Overall, gaps exist in developing, adaptable, holistic, and proper NLP based models helping organisations to achieve their goals.

Chapter 3 Methodology

3.1 Research Approach

The research approach that has been implemented for detecting fake job posting using NLP and ML is deductive approach. According to Sundqvist (2024), this approach starts with theories and concepts which are already established, develops hypotheses and then evaluates empirically through data analysis for conforming, refining, or challenging the existing theories. This approach is justified for this research since it helps in evaluating the predefined assumptions of theories about the impact of semantic, textual, and structural features in fake job posts. By aligning the hypotheses and literature-based variables, this approach makes sure that outcomes are measurable using NLP. This provides structural validation of whether the patterns which are recognised accurately predicts fraud, ensuring objectivity, reliability, and practical application in organisational aspect.

3.2 Justification of the software

The software that has been chosen for this research is Python due to its wide scale adoption, flexibility and powerful range of libraries helping to do NLP and ML. Libraries like scikit-learn, NLTK and spaCy helps in doing proper preprocessing of text, training of model and extraction of features, while frameworks like PyTorch and TensorFlow provide development of complex deep learning models. Python also includes libraries like Numpy and Pandas for doing proper data handling allowing for proper management of data. The open-source nature, cross-platform application and good community support allows Python to be cost-effective, scalable, and accessible for developing ML models which are NLP based for fake job detection.

3.3 Dataset description

The dataset that has been selected for this research is Kaggle's "Real or Fake Job Posting Prediction" data which has 17880 job postings and 18 variables including department, salary_range, company_profile, description, requirements, telecommunication, industry, required_education, function, required_experience and so on (Kaggle, 2020). The target variable is fraudulent (binary where real=0 and fake = 1). There are mixed features like metadata and textual help to develop classification models using NLP. The balance between fake and real job posts along with excellent quality features makes it appropriate for creating fake - job prediction models, testing classification models and analysing how metadata and text signals allows in proper prediction of fake jobs.

3.4 Machine Learning models

XGBoost Classifier: This model is selected during the capability of the models for outperforming traditional models. As stated by Kumar, Yasmeen, and Kumar (2025), boosting methods like XGBoost

helps in reducing bias, handle imbalance datasets and help in improving reliability of prediction by sequentially correcting previous errors. The research revealed that XGBoost achieved 94% precision, F1-score around 0.93 and 92% recall showing that the model perfectly balances false negative and positive, making sure that model is robust.

Logistic Regression: This has been selected since the model is simple, easily interpreted, and effective in doing binary classification like classifying fake and real jobs. According to Goud and Reddy (2025), the model is based on a probabilistic framework which estimates likelihood of jobs being fraudulent, providing a proper decision boundary. The research showed that the model achieved 91.5% accuracy, recall of 91%, precision of 90.2% and F1-score of 90.6% showing that the model can capture linear relationships in structural and textual feature while remaining properly efficient and easy to deploy, thus is an appropriate baseline model for predicting fake job posting.

Naive Bayes: Naive Bayes has been chosen in this study due to its simplicity, speed, and applicability in a task of text classification like detection of fraudulent job posting. Dutta and Bandyopadhyay (2020) stated that Naive Bayes uses the Bayes theorem and conditional probability to deliver the results of a specific classification in case of features independence. Their experiment indicated that they were 72.06% accurate, 0.72 F1-score, and Cohen, KS of 0.12 and MSE of 0.52. Although its performance is worse than the ensemble models, its computational efficiency and its capability in high-dimensional text information make it a worthy base model in detecting fraudulent job posting.

LSTM: LSTM has been chosen as it has the best ability to reason sequential relationships in text data, which is necessary to identify fake job postings. According to Habiba, Islam and Tasnim (2021), deep neural models with recurrent network models like LSTM were more accurate than their traditional counterparts because they were able to remember contextual information on long word sequences. Their relative analysis showed that these models performed better in comparison to those algorithms such as Random Forest and SVM and can reach up to 99% accuracy and higher recall and precision. Even though LSTM is a computationally intensive model, the capability to comprehend contextual meaning and long-term relationships enables it to be a very trustworthy model of detecting deceptive and contextually misleading job postings.

3.5 Evaluation metrics

Accuracy: Determines the rate of total correct prediction by the model (Walker et al., 2020). Applicable in broad performance but cannot be trusted on its own, owing to imbalance of classes in fraudulent job detection.

$$\text{Accuracy} = \frac{TP + TN}{TP - TN + FP + FN}$$

F1-score: Harmonic average of accuracy and recall, which is a balance between false positives and false negatives in the classification (Hicks et al., 2022).

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Precision: Represents the number of the forecasted fake job placements that were real (Vu et al., 2024). Minimises false positives, which is essential to prevent false job ads being treated as scams, and the organisation reputation is safe.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures the number of real fake job postings that the model detects (Pillai, 2023). Critical to reduce false negativity, making sure that the system is not deprived of fraudulent listings, which would be detrimental to the job seekers.

$$\text{Recall} = \frac{TP}{TP + FN}$$

ROC-AUC: Assesses the model in terms of real and fake job posting differentiation at various levels (Salunkhe et al., 2025). Provides strong classification in diverse conditions that helps to ensure that the models can be used to recruit in the real world.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

Chapter 4 System Design

4.1 Introduction

This chapter helps in proposing the design of the system for predicting fake job posting by implementing advanced text analytics, ML and DL models. Based on using techniques like data preprocessing, TF-IDF vectorisation and supervised classification models, the design makes sure that scalable and accurate detection of fake job postings are done. The system tends to understand the tools used for visualisation, class balancing and performance metrics for proper assessment of models thus providing an efficient framework that uses linguistic analytics and predictive modelling for proper detection of fake job posts.

4.2 System Design

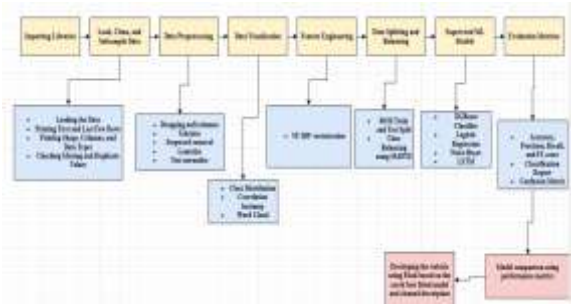


Figure 4: System Design for Fake job posting

The above figure provides the system design for Fake job posting detection using Natural Language Processing which shows a methodical and structured approach using text analytics and ML, DL models for detecting recruitment activities which are fraud. The workflow starts with importation and cleaning of the data to visualise, extraction of features, training of models and evaluating the performance. The inclusion of preprocessing tasks like removal of stop words, lemmatisation and tokenisation makes sure that consistency in language and semantic in text data are met, which is important for predicting language-based deception patterns common in fake job posts (Vu et al., 2024). The incorporation of TF-IDF vectorisation improves the ability of ML and DL models for capturing term importance for documents, enhancing discrimination between fake and real job postings. The implementation of class balancing using SMOTE helps in solving the problem of class imbalance which is a key problem in fraud detection, helping to enhance the generalisability and fairness of the models. The visualisation stage with word clouds and heatmaps provides proper interpretation of the data, helping to understand key textual cues like inappropriate description of jobs or salary. This systematic implementation of preprocessing and exploratory steps shows adherence to implementing NLP and improving the robustness of the models for deployment in the real-world.

The evaluation phase, including Naive Bayes, Logistic Regression, LSTM and XGBoost shows a hybrid approach using classical ML and DL models. This comparative methodology is beneficial, since simple linear models provide interpretability, while LSTM helps in understanding contextual semantics for each sequence, improving performance for long-text description of jobs. The implementation of F1-Score, confusion matrix and accuracy make sure that the assessment of the models is done holistically rather than only focusing on accuracy, showing attention given to recall and precision for each class. This is important for reducing false alarms which might overlook fake job postings. Moreover, the deployment of the best - performing model using Flask shows practical implementation, promoting scalability for online recruitment companies. Conversely, the design could have been improved

by adding explainability models like LIME or SHAP for ethical AI assurance, improving transparency and trust of the HRs. Overall, the architecture properly aligns with advanced NLP pipelines, balancing analytical rigor and feasibility of deployment in predicting frauds in online recruitment.

4.3 Summary

The system shows a proper, end-to-end workflow for predicting fake job postings with the help of ML, DL and NLP techniques. This properly shows cleaning of data, engineering of features, and balancing the training of models and visualisations which can be interpreted. The comparative analysis of the models like Naive Bayes, XGBoost, Logistic Regression and LSTM, improves performance and reliability in detecting fake job posts. Deployment using Flask improves applicability in the real world. The implementation of TF-IDF vectorisation shows textual importance, implementing explainable AI can improve precision and transparency even better.

Chapter 5 Implementation/Solution Development

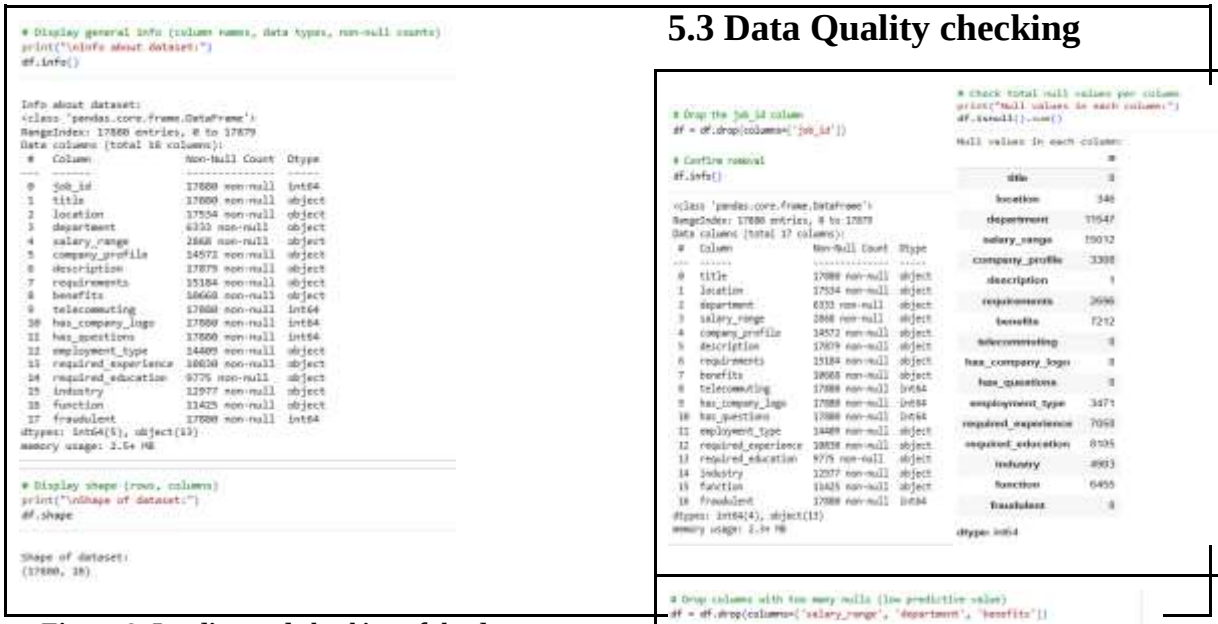
5.1 Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
nltk.download('punkt', quiet=True)
nltk.download('stopwords', quiet=True)
nltk.download('wordnet', quiet=True)
nltk.download('omw-1.4', quiet=True)
nltk.download('punkt_tab', quiet = True)
from wordcloud import WordCloud

from scipy.sparse import hstack, csr_matrix
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from xgboost import XGBClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.naive_bayes import ComplementNB
from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score, roc_auc_score,
    classification_report, confusion_matrix
)
from joblib import dump
import json
import os
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Dropout, BatchNormalization
from tensorflow.keras.callbacks import EarlyStopping
from sklearn.preprocessing import StandardScaler
from scipy import sparse
```

Figure 5: Importing key libraries

The above figure shows importing of key libraries for developing a fake job posting detection system. The figure shows that pandas and numpy have been implemented for wrangling of data, NLTK and regex for lemmatisation, stopwords, tokenisation, TF-IDF and SciPy for sparse features. SMOTE has also been implemented for class balancing, models like Logistic Regression, XGBoost, Complement Naive Bayes and deep LSTM with TensorFlow and Keras along with BatchNorm, EarlyStopping and Dropout. Evaluation metrics like F1, Precision, ROC-AUC, Recall, Accuracy from scikit-learn have also been imported. Lastly, joblib has been imported for saving the models which have been implemented for web deployment.

5.2 Data loading and checking

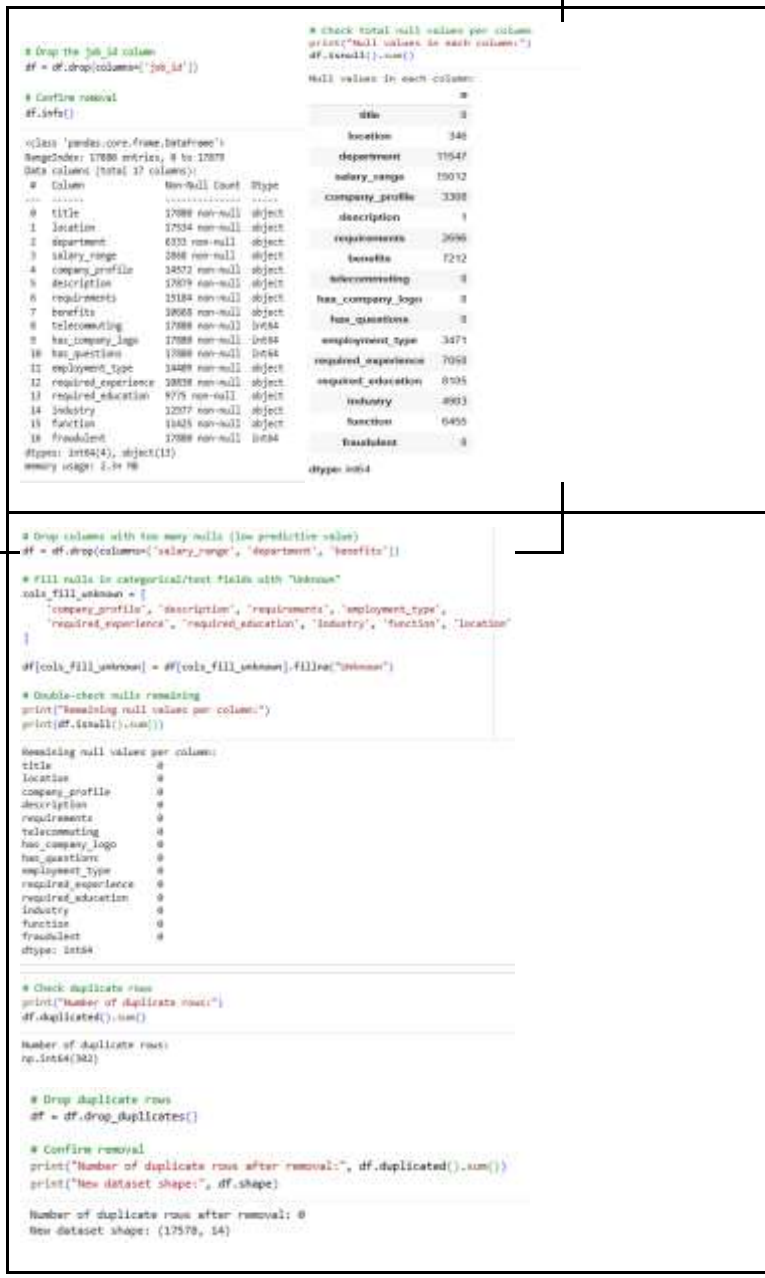


```
# Display general info (column names, data types, non-null counts)
print("\nInfo about dataset:")
df.info()

Info about dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15184 non-null  object
 8   benefits             10668 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
memory usage: 2.5+ MB

# Display shape (rows, columns)
print("\nShape of dataset:")
df.shape

Shape of dataset:
(17880, 18)
```

Figure 6: Loading and checking of the dataset

The above figure shows loading and checking the dataset for getting an idea about the dataset. The dataset has 18 fields and 17880 rows with different data types. There are 5 integers and 13 text based columns showing real-world cases. High missing values in benefits, salary_range, department requires exclusion or target-based imputation. Rich-free texts like requirements, company_profile and description will be included in NLP analysis where clean and concatenate analysis of these columns will be done. This helps in understanding that the dataset has been loaded properly and is ready for doing further cleaning and analysis which will be helpful in creating an unbiased fake job posting detection system.

5.3 Data Quality checking



```
# Drop the job_id column
df = df.drop(columns=['job_id'])

# Confirm removal
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 17880 entries, 0 to 17879
Data columns (total 17 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   title                17880 non-null  object
 1   location             17534 non-null  object
 2   department           6333 non-null   object
 3   salary_range         2868 non-null   object
 4   company_profile      14572 non-null  object
 5   description          17879 non-null  object
 6   requirements         15184 non-null  object
 7   benefits             10668 non-null  object
 8   telecommuting        17880 non-null  int64
 9   has_company_logo     17880 non-null  int64
 10  has_questions        17880 non-null  int64
 11  employment_type      14409 non-null  object
 12  required_experience  10830 non-null  object
 13  required_education   9775 non-null   object
 14  industry             12977 non-null  object
 15  function             11425 non-null  object
 16  fraudulent           17880 non-null  int64
dtypes: int64(4), object(13)
memory usage: 2.3+ MB

# Check total null values per column
print("Null values in each column:")
df.isnull().sum()

Null values in each column:
                        0
title                   0
location              346
department          11547
salary_range        15012
company_profile      3308
description             1
requirements         2696
benefits             7212
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
```

```
# Drop columns with too many nulls (low predictive value)
df = df.drop(columns=['salary_range', 'department', 'benefits'])

# Fill nulls in categorical/text fields with "Unknown"
cols_fill_unknown = [
    'company_profile', 'description', 'requirements', 'employment_type',
    'required_experience', 'required_education', 'industry', 'function', 'location'
]

df[cols_fill_unknown] = df[cols_fill_unknown].fillna("Unknown")

# Double-check nulls remaining
print("Remaining null values per column:")
print(df.isnull().sum())

Remaining null values per column:
title                  0
location               0
company_profile        0
description            0
requirements           0
telecommuting          0
has_company_logo       0
has_questions          0
employment_type        0
required_experience    0
required_education     0
industry               0
function               0
fraudulent             0
dtype: int64

# Check duplicate rows
print("Number of duplicate rows:")
df.duplicated().sum()

Number of duplicate rows:
np.int64(302)

# Drop duplicate rows
df = df.drop_duplicates()

# Confirm removal
print("Number of duplicate rows after removal:", df.duplicated().sum())
print("New dataset shape:", df.shape)

Number of duplicate rows after removal: 0
New dataset shape: (17578, 14)
```

Figure 7: Data Quality checking

The above figure shows the steps of checking data quality for properly deploying ML models for classifying fake job postings. The preprocessing proper includes removing non-informative identifiers, treating missing values and removing duplicate records, providing a clearer foundation for modelling fake job postings. As stated by Gupta et al. (2021) cleaning of the data in this phase is important since performance of models, generalisability and fairness are only proper when

the data provided is cleaned. Thus, dropping job_id is suitable in this case since this provides no semantic information and could hamper the order of the rows. Checking null values shows columns like benefits, department and salary_range has high missing values which has been dealt with by dropping these to prevent noisy imputation but might discard key signals about salaries which are unrealistic or fake departments that often show scams. Filling the rest with categorical value "Unknown" helps in standardising the missing information and prevents errors when vectorisation is implemented, yet it implicitly groups many different cases, like genuinely unspecified fields and posts which provide partial information, which might dilute subtle cases of fraud. After this it was observed that there are 302 duplicate values which are removed for reducing biasness of the models leading the remaining rows and columns to be 17578 and 14 respectively, ready for further analysis.

5.4 Text based preprocessing using NLP

Figure 8: Text based preprocessing using NLP

The above figure shows the process of conducting text-based preprocessing using NLP which includes normalisation, tokenisation, and filtering job-post text before extraction of features. As stated by Chai (2022), this step is important for ML and DL since performance of the models directly are dependent on how well the raw data is shown and cleaned. Thus, the function, normalize_and_clean helps in converting to lowercase texts, removes URLs and HTML, numbers and punctuations, reducing noise and superficial variations. Lemmatisation and stopping words are also implemented to compress inflected forms into core meanings, allowing the model to focus on informative tokens. This is important for fake job post detection, where subtle linguistic cues like adjectives which are exaggerated or job descriptions which are vague can help to

differentiate fake listings for real ones. Corpus-level rare-world filtering with minimum threshold of frequency further helps in idiosyncratic tokens, reducing the size of vocabulary and enhancing computational efficiency for down streaming TF-IDF vectorisation, however, this process risks removing rare but highly indicators of frauds. Implementing this pipeline for company_profile, requirements, title and description and saving with cleaned suffix, preserves the structural context while making sure no missing values are there. Thus, this NLP is robust and systematic which makes deployment of ML and DL models easy.

5.5 Data Visualisation

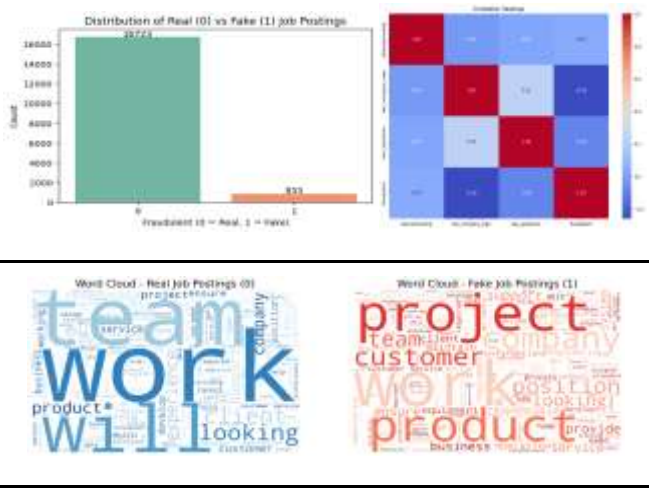


Figure 9: Data Visualisation

Data visualisation has been implemented for checking trends and patterns of fake job postings. The bar chart shows that the data is imbalance, since there are only 855 fake job posts with 16726 real ones, thus proper steps for class balancing is needed before training the models. The correlation heatmap shows that having a company logo has a negative correlation with fraud of -0.26 showing a key indicator of authenticity, while has_questions have a weak negative correlation, showing that fraudsters are not using screening questions. The word clouds provide key qualitative insights based on language used like for real job postings words like team, service and work are used, showing clarity of the role and structure of communication among the corporates. However, fake postings mostly use vague or promotional terms like equipment, provide and customer showing attempts to entice candidates through promises.

5.6 Feature engineering, data splitting and balancing

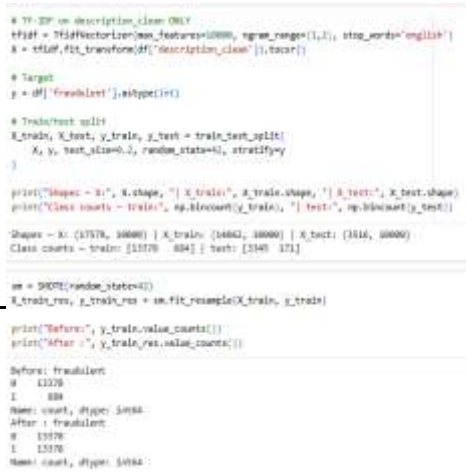


Figure 10: Feature engineering, data splitting, and balancing

The above figure shows the key steps of feature engineering, data splitting and balancing for detecting fake job posting. According to Tatchum et al. (2020) implementation of vectorisation of text using TF-IDF helps in changing proper descriptions into weighted numbers, which is important for ML as this helps in changing unstructured text into vectors which can be measurable for ML to interpret. In this case, by reducing the features to 10000 and implementing bigram combination, the model captures association between words properly while reducing dimensionality. After this the dataset is split into 80:20 along with stratification, making sure proportional representation of fake and real classes in both sets. However, since the class is imbalanced, SMOTE has been implemented to synthetically generate minority samples and equalise class counts.

Chapter 6 Results and Critical Analysis

6.1 Implementation of Models

6.1.1 XGBoost classifier

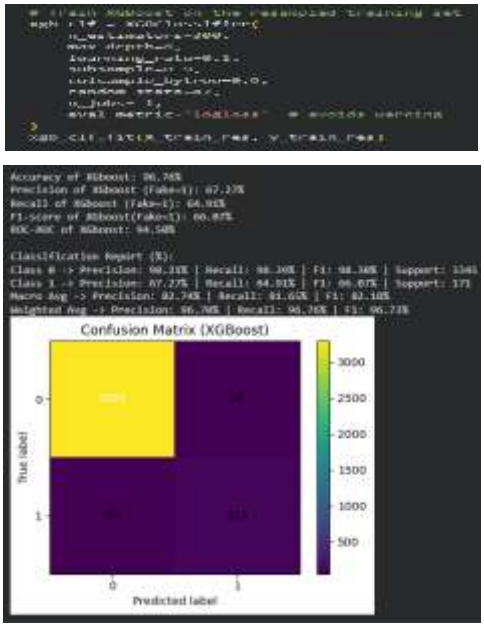


Figure 11: Performance of XGBoost

XGBoost in ***Figure 11*** made one of the best performances that saw it attain an accuracy of 96.76, which is the highest amongst the classic models. It showed great accuracy (67.27%), recall (64.91%), and F1-score (66.07) on fake postings, which means that it exhibits a high degree of balanced detection even in minority-class situations. The real postings were very highly classified and a confusion matrix was provided to show that there were very few misclassifications. The boosted-tree structure of XGBoost effectively modeled the nonlinear features and feature interactions thus outperformed in terms of identifying vague or deceitful language. The model had a ROC-AUC of 94.50 hence the model was extremely discriminatory and therefore, it was very reliable in detecting fraud in job advertisements.

6.1.2 Logistic Regression

```
# Train Logistic Regression on the resampled training set
log_clf = LogisticRegression(
    solver='liblinear',
    max_iter=1000,
    random_state=42
)
log_clf.fit(X_train_res, y_train_res)
```

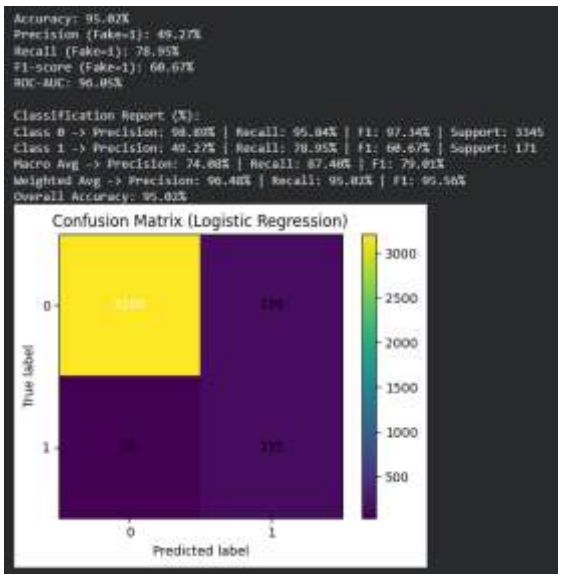


Figure 12: Performance of Logistic Regression

The Logistic Regression model in ***Figure 12*** was 95.02% accurate and has good and consistent performance as a baseline classifier. It provided accuracy of 49.27, recall of 78.95 and F1-score of 60.67 to fake postings, that is, it was able to identify most of the fraudulent entries, but produced some false positives. It was much more effective in textual complexity than the Naive Bayes. The confusion matrix indicates the misclassifications are less, particularly in the case of actual posts. Compared to both the XGBoost and the LSTM, the Logistic Regression was not as complex, yet its interpretability, computational efficiency and consistency were excellent.

6.1.3 Naive Bayes

```
# Naive Bayes

# Train Naive Bayes on the resampled training set
nb_clf = ComplementNB()
nb_clf.fit(X_train_res, y_train_res)
```

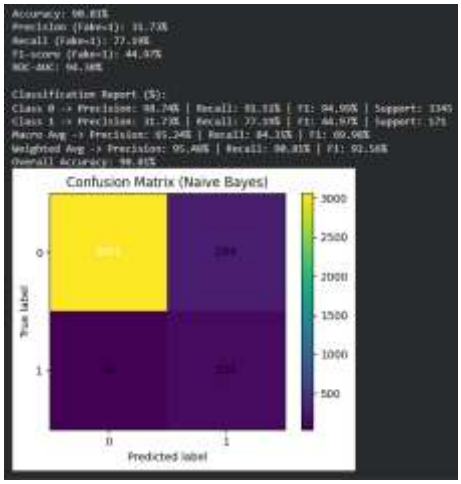


Figure 13: Performance of Naive Bayes

The overall performance of Naive Bayes in ***Figure 13*** was reasonably good with 90.81% accuracy but its results have revealed that the system possesses some definite weak points in identifying fake postings. The model yielded a low precision value (31.73%), and a recall value (77.19%), which implies that the model predicted many fake entries and a high number of real posts as fake. The F1-score of the fake-class was 44.97 which was low precision and recall. This problem is identified in the confusion matrix where 284 real posts were wrongly classified as fake. Naive Bayes was also efficient, fast but could not handle complex linguistic patterns and not the best among all models.

6.1.4 LSTM

Figure 14: Performance of LSTM

The LSTM model was found to be very strong with the accuracy at 96.47 and well-balanced classification measures. It produced an F1-score of 64.16% and a recall of 64.91% in fake postings, which indicates that it can pick up the sequencing lingual features. The confusion level indicates that the LSTM hardly classifies most of the cases of frauds as opposed to the simpler models. It is powerful with the ability to read into contextual patterns within long job descriptions and therefore is efficient in natural language tasks.

6.1.5 Best Model

Figure 15: Best model

Based on recall and precision, F1-score, ROC-AUC and accuracy, it was determined that the most effective model is the Logistic Regression. Although XGBoost demonstrated high levels of accuracy, the Logistic Regression gave the most proportioned results in terms of critical fraud-detection variables. It attained a recall of 78.95% which guaranteed that more fraud posts were reliably detected and that its precision and F1-scores remained stable. It has a high ROC-AUC of 96.05% indicating a good capacity to discriminate. Moreover, the model is computationally efficient, easy to interpret and deploy thus it is very suitable in real world application. Logistic Regression was

the most valid and generalisable in terms of performance in all metrics of evaluation ***(Refer to Figure 15)***.

6.1.6 Flask implementation for real-world fake job deployment

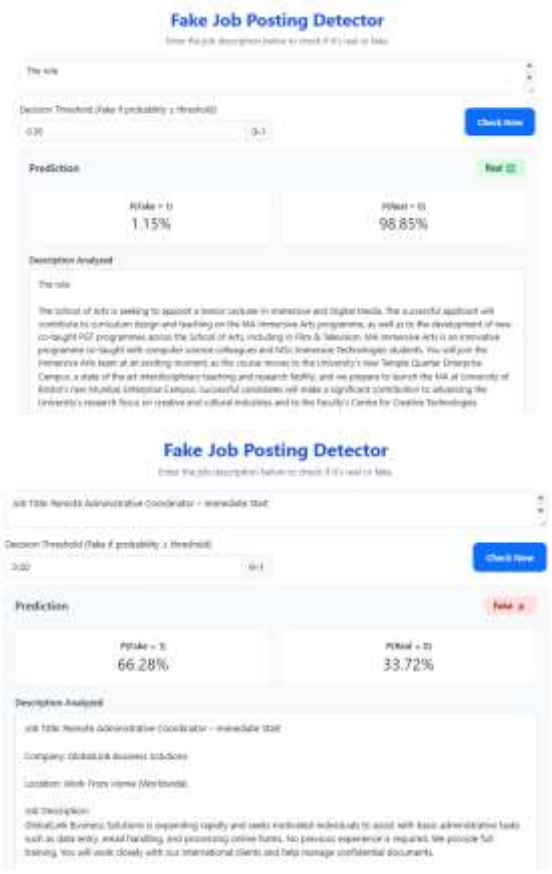


Figure 16: Flask-based web application

A Flask-based web application in ***Figure 16*** was constructed to show real-life implementation of the selected model. The interface enables the user to key in the job description and automatically get forecasts on whether the job advertisement is genuine or counterfeit. The system has probability scores of both classes, and decision thresholds that can be adjusted to have a higher level of control. The trained Logistic Regression model and preprocessing pipeline are combined in the backend and make stable and precise predictions. It displays analysed textual features in the application, which enhances transparency to the users. This efficient use demonstrates the possibility to introduce NLP-based fraud detection into the recruitment platforms without the disruption to promote safer and more reliable experience.

6.2 Findings

The results of the present research have identified the usefulness of an NLP-based methodology in identifying fake job advertisements by conducting a thorough analysis, preprocessing and model compassion, and linguistic patterns identification. The exploratory data analysis showed that the data was imbalanced with many genuine postings compared to the fraudulent ones. Fake job postings have tended to be poorly organised with respect to income benefits and company logos that validated definite distributional traits. Text with metadata was converted to clean consistent semantics though systematically preprocessing with normalising, tokenising lemmatisation, removing stop-words, filtering rare-works and vectorising using TF-IDF. These steps ensured that transformation of all features to be machine readable and applicable to the effective use of advanced NLP-based classification models. While comparing several models, Logistic Regression is found to have the best predictive performance especially in terms of recall and F1-score that makes it more effective in single minority fraudulent cases. There were some important linguistic indicators in the case of fake postings, including vague descriptions, excessive promotion, lack of role specification and the use of enticing terms.

6.3 Compare results with previous work

6.3.1 Exploratory data analysis of real and fake job postings

To gain a sense of distribution and concept of real and fake job postings, exploratory data analysis was carried out based on class imbalance, structural features, and language forms. As per literature, fake job advertisements were missing organisational

identifiers, were poorly described and with indiscriminate content templates (Prasanna et al., 2022). In this aspect, results found that the exploratory data analysis revealed comparable patterns within the study, with fraudulent posts not having company logos, salary ranges, and role description, being represented much less than authentic ones that support the imbalanced nature. Visual analysis also revealed that fake postings had more promotional and vague words, as compared to real postings that had more specific job-related language (Nanath and Olney, 2023; Reddy et al., 2022). These results showed that the distribution and textual behaviour of fake posting are like existing studies and thus can be considered as effective indicators to detect fake job advertisements.

6.3.2 Text and metadata pre-processing for NLP-based classification

The pre-processing and transformation of text and metadata to machine readable formats was critical in allowing the NLP based classification models to capture and interpret the linguistic structures and structural properties efficiently. The previous literature has underscored the significance of cleaning text, cleaning up noise, processing missing organisational identifiers and standardising descriptive fields to identify fraud-related signals (Nanath and Olney, 2023; and Reddy et al., 2022). The systematic preprocessing of the raw text has included normalisation, tokenisation, lemmatisation, removal of stop-words, filtering of rare-words and TF-IDF vectorisation in the current results to convert raw text into numerical forms and normalise the metadata fields to ensure consistency. These results have provided more linguistic insight and better feature stability compared to previous works. Thus, null hypothesis has been rejected and alternate hypothesis has been accepted which proves that meaningful linguistic patterns have been derived to correctly discriminate between fake and real job posts.

6.3.3 Machine learning model comparison and performance enhancement

In a comparison to optimize the results of machine learning models in detecting opportunistic job postings, this research study has compared several classifiers on a balanced and engineered dataset. As described in the literature, it has been established that the models, which include the Random Forest, SVM, Decision Tree, DNN and ensemble models have worked well with clean text with 98% accuracy, TF-IDF features and proper imbalance treatment (Reddy et al., 2022; Lakshmi et al., 2024). The studies have also highlighted the need of focusing on lightweight detectors and mixed rule-based filtering to enhance the performance (Prasanna et al., 2022). Ensemble-based models that are trained on TF-IDF vectors using SMOTE balancing have shown better stability in classification in the current results. In comparison to the previous research, results have been consistent, which proves that optimized preprocessing and optimised ensemble models have offered better potential to detect fake job postings.

6.3.4 Identification of linguistic patterns and key features in fake job advertisements

To find significant linguistic patterns, features or indicators that are recurrent in fake job advertisements, the most significant patterns and indicators have been distinguished between authentic job postings and the fake ones. It has been demonstrated by the existing literature that the language patterns of fraudulent posts are usually inconsistent, there is imposition of inappropriate formatting and imbalance in focus on incentive or a sense of urgency (Nanath & Olney, 2023; Prasanna et al., 2022; Lakshmi et al., 2024). In the current findings, TF-IDF based interpretations and correlation patterns have identified cues, including

use of promotional language, deficient role definitions, inconsistent structure, and absence of screening questions as valuable indicators of fraud. This signifies that usage of linguistic ambiguity and absence of organisational attributes has always been a central indicator of fake job advertisements.

Chapter 7 Discussion and Conclusions

7.1 Discussion of results in relation to objectives, validity, scope, and generalisability

Based on the secondary quantitative data analysis, it has been concluded that the structured dataset, systematic preprocessing, and multi-model analysis have effectively generated legitimate and interpretable outcomes to detect fake job postings. The study also found that the strength of TF-IDF and SMOTE-based modelling and the convergence of linguistic indicators across models have been used to impose confidence in the results. Despite the constraining nature of the dataset to a broader scope, and external validity across platforms, the internal validity has been high and the external validity to other recruitment-fraud settings has been moderate.

The findings have demonstrated that the distributional discrepancy between genuine and fraud job postings has been successfully demonstrated using exploratory data analysis. The lack of class balance, missing organisational identifiers, unspecified salary ranges and ambiguous descriptions correlated with the patterns highlighted in study results. Visualisations also proved that fraudulent advertisements contained more promotional and vague words whereas authentic posts had more definite role-specific words. *This signifies that the first objective has been achieved with success after the revelation of structural and linguistic inconsistencies.*

The findings also have shown that extensive preprocessing to clean, tokenise, lemmatise, remove stop-words, filter rare-words, and vectorise TF-IDF has yielded clean, stable and machine-readable representations of job-post text and metadata. These processes have also enhanced feature stability and linguistic clarity. *Thus, the designed pipeline has successfully supported the second objective as it turns unstructured raw and noisy data into structured numerical data that has been used to detect fake-jobs with high accuracy and reliability*. Comparison of models have demonstrated that ensemble models, especially LR, have demonstrated more consistent and predictable performance in fraudulent posting detection compared to the baseline classifiers. Balanced TF-IDF with SMOTE features helped to achieve higher recall and F1-scores, in minority fraud cases. *Ensemble and deep-learning methods have verified these findings and show that the third objective has been achieved with optimisation of models and improved performance assessment.*

The findings have revealed that there are recurring linguistic cues of deceit such as the use of vague descriptions, over use of promotional language, no screening questions, and no organisational specificity. These features have been identified by the interpretation of the TF-IDF as significant predictors of fraudulent postings. These results have demonstrated incoherent structure and urgency-based phrasing in counterfeit listings. *Therefore, the fourth goal has been achieved by finding valuable linguistic indicators to distinguish between fraudulent advertisements.*

7.2 Implications of research findings

The results have significant implications on enhancing digital safety and curbing employment fraud. The study has shown that automated screening systems can be reliable in identifying

suspicious contents that have reached job seekers by affirming that NLP techniques can identify meaningful linguistic and structural patterns in job advertisements. The findings have also suggested that regular preprocessing, and feature engineering, is needed to come up with scalable detectors that are also robust in diverse datasets. The success of the classification models has also shown that organisations, recruitment sites, and verification agencies can combine such pipelines to build real time fraud detection. The conclusion is that the research has helped to implement proactive, data-driven protection that limits the exposure to fraudulent job offers, enhances user confidence, and leads to a safer online recruitment experience.

7.3 Limitations of the study

The analysis has been based on only the dataset that provides a limited richness, diversity, and context depth of signals to use in modelling. The model is limited by small feature space that limits its capability of modeling elaborate fraud indicators that are manifested in real recruitment ecosystems. Despite multiple ML models having been considered, classical algorithms have been the main approach of the study and only one deep learning algorithm. Sophisticated transformer-based and multimodal systems have not been incorporated that has restricted the capacity to detect small semantic and behavioural hints in counterfeit job adverts. The research has only concentrated on secondary quantitative evidence, that is, primary information has not been included where job seekers, HR professionals and cybersecurity experts are concerned. This has restricted the ability to authenticate whether model identified trends are associated with actual user experiences.

7.4 Recommendations for future work

Future studies can be well served to expand data samples by incorporating larger, more varied and multimodal data, such as pictures, site snapshots of company websites and more enriched metadata, to better see the deeper indicators of recruitment fraud. In addition to that more developed transformer-based and hybrid models comprising classical machine learning and deep learning should be investigated to achieve a better semantic understanding and robustness. Research may be performed in the context of the mixed-method approach, that is including primary data, including interviewing or surveying job-seekers, HR professionals, and cybersecurity experts, and secondary analysis. This can enable the researcher to confirm the presence of model-identified patterns in the real-world experiences.

References

Afzal, H., Rustam, F., Aljedaani, W., Siddique, M.A., Ullah, S. and Ashraf, I. (2023). Identifying fake job posting using selective features and resampling techniques. *Multimedia tools and applications*, 83(6), pp.15591–15615. doi:https://doi.org/10.1007/s11042-023-15173-8.

Akerlof, G.A. (1970). The Market for ‘Lemons’: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), pp.488–500. doi:https://doi.org/10.2307/1879431.

Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V. and Ludi, S. (2022). Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. *Neural Processing Letters*, 54. doi:https://doi.org/10.1007/s11063-021-10727-z.

Anbarasu, V., Selvakani, S. and Vasumathi, K. (2024). Fake Job Prediction Using Machine Learning. *International Journal of Darshan Institute*

on Engineering Research and Emerging Technologies, 13(1), pp.12–20. doi:https://doi.org/10.32692/ijdi-eret/13.1.2024.2403.

B, S.U., Singh, V.R., P, S., and Dhage, A. (2023). Fake Job Post Prediction Using Data Mining. *Journal of Scientific Research and Technology*, [online] pp.39–47. doi:https://doi.org/10.5281/zenodo.7954261.

Chai, C.P. (2022). Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3), pp.1–45. doi:https://doi.org/10.1017/s1351324922000213.

Davis, E. (2025). *1 in 4 Job Listings on LinkedIn Are Likely 'Ghost Jobs,' According to a New Report. Here Are the Cities Impacted the Most.* [online] Entrepreneur. Available at: https://www.entrepreneur.com/business-news/one-quarter-of-jobs-posted-online-are-fake-ghost-jobs-study/496683 [Accessed 5 Oct. 2025].

Devos, J., Landeghem, H.V. and Deschoolmeester, D. (2011). The Theory of the Lemon Markets in IS Research. *Information Systems Theory*, pp.213–229. doi:https://doi.org/10.1007/978-1-4419-6108-2_11.

Dutta, S. and Bandyopadhyay, S.K. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. *International Journal of Engineering Trends and Technology*, [online] 68(4), pp.48–53. doi:https://doi.org/10.14445/22315381/ijett-v68i4p209s.

Goud, T.N. and Reddy, Dr.N.R. (2025). A Machine Learning Approach for Detecting Fraudulent Job Postings in Online Recruitment Platforms. *Journal of Sensors, IoT & Health Sciences*, 3(3), pp.14–28. doi:https://doi.org/10.69996/jsihs.2025012.

Gupta, N., Mujumdar, S., Patel, H., Masuda, S., Panwar, N., Bandyopadhyay, S., Mehta, S., Guttula, S., Afzal, S., Sharma Mittal, R. and Munigala, V. (2021). Data Quality for Machine Learning Tasks. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. doi:https://doi.org/10.1145/3447548.3470817.

Habiba, S.U., Islam, Md.K. and Tasnim, F. (2021). *A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICREST51555.2021.9331230.

Hamza, P.A., Othman, B.J., Gardi, B., Sorguli, S., Aziz, H.M., Ahmed, S.A., Sabir, B.Y., Ismael, N.B., Ali, B.J. and Anwar, G. (2021). Recruitment and Selection: the Relationship between Recruitment and Selection with Organizational Performance. *International Journal of Engineering, Business and Management*, [online] 5(3), pp.1–13. doi:https://doi.org/10.22161/ijebm.5.3.1.

Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P. and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, [online] 12(1), p.5979. doi:https://doi.org/10.1038/s41598-022-09954-8.

Jaison, S.S. and Kodabagi, M. (2023). *View of Identifying Real and Fake Job Posting using Machine Learning*. [online] Jazindia.com. Available at: https://jazindia.com/index.php/jaz/article/view/2266/1791 [Accessed 2 Oct. 2025].

Kaggle (2020). *Real / Fake Job Posting Prediction*. [online] Kaggle.com. Available at: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data [Accessed 10 Oct. 2025].

Kesslen, B. (2025). *1 of every 5 job postings are fake, study says*. [online] Quartz. Available at: https://qz.com/one-in-five-job-postings-is-fake-ghost-jobs-1851738201 [Accessed 3 Oct. 2025].

Kumar, S., Yasmeen, S., and Kumar, P. (2025). The Growing Threat of Fake Job Postings: A Review of Machine Learning and NLP-Based Detection

Approaches. *Revolutionary Advances in Computing and Electronics: An International Journal*, 1(1), pp.41–51. doi:https://doi.org/10.1807/6rw1fg49.

Lakshmi, L.S., Helen, V.B.J., Thanuja, R. and Noushin, SK. (2024). Fraudulent Job Post Recognition. *IARJSET*, 11(3). doi:https://doi.org/10.17148/iarjset.2024.11328.

Mishra, S. (2023). Ethical Implications of Artificial Intelligence and Machine Learning Ethical Implications of Artificial Intelligence and Machine Learning in Libraries and Information Centers: A Frameworks, Challenges, in Libraries and Information Centers: A Frameworks, Challenges, and Best Practices and Best Practices Ethical Implications of Artificial Intelligence and Machine Learning in Libraries and Information Centres: A Frameworks, Challenges, and Best Practices. [online] Available at: https://digitalcommons.unl.edu/context/libphilprac/article/14939/viewcontent/auto_convert.pdf [Accessed 4 Nov. 2025].

Nanath, K. and Olney, L. (2023). An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud. *International Journal of Information Management Data Insights*, 3(1), p.100167. doi:https://doi.org/10.1016/j.jjimei.2023.100167.

Pillai, A.S. (2023). DETECTING FAKE JOB POSTINGS USING BIDIRECTIONAL LSTM. *International Research Journal of Modernization in Engineering Technology and Science*, 5(3). doi:https://doi.org/10.56726/irjmets35202.

Prasanna, Mrs.T.L., Mehartaj, Mrs.SK., Kinnera, Miss.S. and Mounika, Miss.N. (2022). A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT DATA MINING TECHNIQUES. *International Journal of Research in Engineering, IT and Social Sciences*, [online] 12, pp.43–49. Available at: https://www.indusedu.org/pdfs/IJREISS/IJREISS_4057_46806.pdf [Accessed 10 Oct. 2025].

Reddy, K.S.S., Dwarakamai, J., Rahul, M., Chowdary, C.Y. and Reddy, Mr.M.S. (2022). A Comparative Study on the Prediction of Fake Job Posts using Various Data Mining Techniques. *International Journal of Advanced Research in Science, Communication and Technology*, pp.621–627. doi:https://doi.org/10.48175/ijarsct-5733.

Salunkhe, A., Taware, A., Golande, S., Khandagale, R. and D.Shelar, M. (2025). Fake Job Post Detection using Machine Learning and Deep Learning. 7(7), pp.577–585. doi:https://doi.org/10.35629/5252-0707577585.

Sundqvist, P. (2024). Extramural English as an individual difference variable in L2 research: Methodology matters. *Annual Review of Applied Linguistics*, [online] pp.1–13. doi:https://doi.org/10.1017/S0267190524000072.

Tatchum, G.W., Nzekon, J., Makembe, F.S. and Djam, X.Y. (2020). Class-Oriented Text Vectorization for Text Classification: Case Study of Job Offer Classification. *Journal of Computer Science and Engineering (JCSE)*, [online] 5(2), pp.116–136. Available at: https://icsejournal.com/index.php/JCSE/article/view/870 [Accessed 4 Nov. 2025].

Thapa, A. (2024). *Ghost jobs: What the rise in fake job listings says about the current job market*. [online] CNBC. Available at: https://www.cnbc.com/2024/08/22/ghost-jobs-why-fake-job-listings-are-on-the-rise.html [Accessed 6 Oct. 2025].

V, V. and Vishvanath, Prof.A. (2025). FAKE JOB POSTING DETECTION USING ML. *International Research Journal of Modernization in Engineering Technology & Science*, 7(9). doi:https://doi.org/10.56726/irjmets82465.

Vu, D.-H., Nguyen, K., Tran, K.T., Vo, B. and Le, T. (2024). Improving fake job description detection

using deep learning-based NLP techniques. *Journal of Information and Telecommunication*, pp.1–13. doi:https://doi.org/10.1080/24751839.2024.2387380.

Walker, S., Khan, W., Katic, K., Maassen, W. and Zeiler, W. (2020). Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. *Energy and Buildings*, 209, p.109705. doi:https://doi.org/10.1016/j.enbuild.2019.109705.

Yamashita, M., Tran, T., and Lee, D. (2024). Fake Resume Attacks: Data Poisoning on Online Job Platforms. *ACM Reference Format*, [online] 12. doi:https://doi.org/10.1145/3589334.3645524.

Bibliography

Prasath S, R. and Nachappa, M.N. (2024). Fake Job Posting Detection. *International Journal of Advanced Research in Science, Communication and Technology*, pp.283–287. doi:https://doi.org/10.48175/ijarsct-15950.

Sonkar, S., Yadav, S. and R, S. (2025). A Hybrid Machine Learning Approach for Fake Job Posting Detection: Integrating Naive Bayes and Logistic Regression Models. *International Journal of Innovative Science and Research Technology*, pp.323–329. doi:https://doi.org/10.38124/ijisrt/25jun458.

Vu, D.-H., Nguyen, K., Tran, K.T., Vo, B. and Le, T. (2024). Improving fake job description detection using deep learning-based NLP techniques. *Journal of Information and Telecommunication*, [online] pp.1–13. doi:https://doi.org/10.1080/24751839.2024.2387380.