**InfosysSpringboard Internship 4.0 Project**

**Documentation**

**Project Report   
 Fake Job Posting**

# **Problem Statement- Fake News Detection**

## **Use Case- Develop a Deep Learning-Based System That Can Accurately Detect and Classify Fake Job Postings**

## 

* **Submitted by-**

**Mentor Name-** SUDHEER KUMAR Y

**Intern Name-** TANISHKA SHAH

## 

**MAY- JUNE 2024**

**Content-**

# Introduction

#### background

#### Problem Statement

#### Objectives

#### Significance of Study

#### Challenges

#### Scope

#### Methodology

#### Expected Outcome

# Data Preprocessing

* Data Collection
* Data Cleaning
* Data Preprocessing
* Feature Engineering
* Data Splitting

# Modelling Approch

* Logistic Regression
* LSTM model
* Bert Based Model

# Comparision of Models

# Future Scope and Use Cases

* Future Work
* Practical Applications

# Conclusion

# 

# **Introduction:-**

* Background

Fake news and online fraud, in general, have sharply grown in recent years. Now among these, some of the most knee-breaking frauds are 'fake job posting' frauds. Such fraudulent job ads will always do nothing but waste the time of job seekers - apart from wasting their valuable time and money spent on Bonds in infrastructure, Real Estate and Money Laundering. The deceits may sometimes also cross the threshold of identity and financial fraud. This problem is further catalyzed with today's internet and social media, which provide a very large audience to malicious actors with the required anonymity to carry on such activity. Thus, safeguarding people and preserving the integrity of online job markets by detecting and preventing fake job postings has become an indispensable mission.

* Problem Statement

This paper mainly makes an endeavor to propose a deep learning-based system for proper and correct detection and classification of fake job posting. This system is going to target job seekers, recruiters, and platform administrators to identify for holding fake job ads and filter them to avoid, hence enhancing the trust and safety of the online job platform.

* Objectives

1. Dataset Development: Gather a full dataset with preprocessed legitimate and fake job posts, to be used in training and testing the deep learning model.

2. Feature Engineering: Integrate important features from within the job posts to help distinguish between genuine and fake advertisements.

3. Model Development : Design any deep learning model to solve text classification problems. The model needs to train on the objectives for the recognition of patterns and indicators in tokens of fake job postings.

4. Model Evaluation : Using the developed models, estimate various metrics of classification—accuracy, precision, recall, and F1 score. Make sure this procedure is well generalized and is robust to unseen data.

5. Deployment : Tie the developed model into a user-friendly system that will allow checking the legitimacy of the posts in real-time by jobseekers and the administration of this platform.

* Significance of the Study

The rise of the fake job posts is not only destructive to the experience of a job seeker but also impacts safety-related issues on several critical levels. The proposed system will:

- Protect Job Seekers : Since they would not be getting access to fraudulent job offers, they would not fall into the traps, which would waste their time and lead them into a scam. This time and effort could have been spent by job seekers on a legitimate job.

- Enhance the Image of Job Boards and Recruitment Portals : They can look more reliable as they will be able to track more visitors and be sure of the authenticity.

- Help with Cybersecurity : The identification and exploration of fraud contribute to the broader objectives of security online and the reduction of cybercrime.

* Challenges

Several challenges lie in the task of fake job posting detection:

- Data availability: A large and well-annotated dataset is required for training the model to give sound performances. Collection and labeling of such data can be resource-consuming and time-intensive.

- Evolving tactics : Fraudsters keep evolving their tactics to outsmart detection. Therefore, this solution needs to have an adaptive model and hence remains effective over time.

- Feature Extraction : Correct features that hold the differences between real and fake job postings are very essential to have high model performance.

- Balancing Precision and Recall : There should be a balance between the minimizing of false positives, misclassifying true job postings as fake, and false negatives, failing to identify fake job postings, in order to be practically useful.

* Scope

The following are the key components within the scope of this project:

1. Data Collection and Preprocessing : This step will involve the collection of a job posting dataset and, finally, data preprocessing in readiness for the analysis to be done.

2. Model Development and Training : Design and training of a deep learning model for fake job posting detection.

3. Evaluation and Optimization : Evaluation of this model's performance and optimization for better accuracy and generalization.

4. Deployment and User Interface : User interface needs to be developed that will help in real-time fake job posting detection.

* Methodology

The project will carry out a methodical process of achieving the objectives at hand by going through several key steps:

1. Data Collection : It will collect job postings from sites like job boards, recruitment websites, and datasets available to the public. Data augmentation techniques may also be applied to ensure diversity in the dataset.

2. Preprocessing : Text preprocessing steps such as tokenization, lemmatization, and removal of stop words shall form part of data preparation to be training of the model.

3. Feature Engineering : Explore the possibility of TF-IDF, Word Embedding, and other text-based features to increase the model's performance.

4. Model Design : The effectiveness of various deep learning architectures, such as RNN, CNN, and Transformer-based models, will be compared.

5. Training and Evaluation : The model will be trained and evaluated over the dataset collected with the corresponding model. Techniques regarding hyperparameter tuning and cross-validation should be used for the optimization of models.

6. Deployment : A user interface will be developed to facilitate easy interaction for inputting job posts, with the posting users receiving real-time responses indicating the authenticity of the job posts.

* Expected outcomes

By the end of the project term and deloyement, we hope to achieve the following with this model:

• High‐accuracy: The model should be able to classify job posts correctly and should have high precision and recall.

– User-Friendly Interface : An interface that is intuitive in nature and friendly to all users when trying to check the credibility of a given vacancy

– Adaptability: The system will adapt to new and changing scammer techniques through learning and updating continually.

In today's digital landscape, the detection of bogus job postings is critical. This project aims to provide an effective solution for securing job seekers and maintaining the quality of online job boards by employing deep learning techniques. It will provide a useful tool against online fraud through careful data collection, model development, and user-sensitive design.

# **1.Data Preprocessing-**

* Data Processing

Data pre-processing is one of the important processes in the development of any deep-learning-based fake job posting detection and classification system. This process is about the collection, cleaning, transformation, and arrangement of data to be used for training and evaluation. A well-done data processing stage ensures the learning and generalization of the model from quality data and can much better generalize a model to new and unseen job postings effectively. The main steps in the data processing pipeline for the current project are as follows:

**1. Data Collection**

Objective : To collect a wide and, at the same time, full diverse dataset of job postings, including both legitimate and fake entries.

Sources:

- Online Job Boards and Recruitment Websites : These would include LinkedIn, Indeed, Glassdoor, and other major players in this online employment space. These can be considered primary sources of such job postings.

- Publicly Available Datasets : Datasets sourced from research papers, competitions like Kaggle, or government job boards.

- Web Scraping : Pulling job listings from a range of online sources, including mostly utilized job boards, using automated scripts.

- Manual Collection and Annotation: The datasets of manually created content that are usually labeled by hand, including examples of fake and real job posts.

Challenges:

- The dataset has to be from a mix of different job sectors and geographic regions.

- The number of fake and real job postings should be somewhat similar, as much possible, to prevail over the class imbalance scenario.

**2. Data Cleaning**

Goal: Cleaning a dataset from irrelevant, incomplete, or noisy data to improve quality.

Procedure

• Remove Duplicates: Identify and eliminate duplicate job posts to avoid redundancy.

• Handle Missing Values : Deal with missing values either by replacing them with suitable placeholders or removing incomplete records.

• Remove Irrelevant Data : Eliminate posts not relevant to the job market, like spam or advertisements.

• Uniformity in Formatting : Data must be placed in a uniform format; for example, data should be in date formats and in a similar case, either lowercase or uppercase.

Problems:

- Processing a huge amount of data in an efficient way.

• Decide how to handle any missing or incomplete data in such a way that no bias is injected.

**3. Data Preprocessing**

Goal : Make the source data ready to be trained on the model.

Steps :

- Tokenization : Breaking down the job descriptions into words or tokens.

- Lowercasing : To change all text into lowercase form to standardize it.

- Stopwords Removal : These are common words like 'and', 'the', 'is' which barely add any valuable information.

- Lemmatization/Stemming : Reducing words to their base or root form so that similar words can be treated uniformly ("running" to "run").

- Removing Punctuation and Special Characters\*\*: Removing characters that are non-alphanumeric and do not add too much meaning to the text.

- Handling Contractions: Expanding contractions to their full forms ("can't" to "cannot").

Challenges :

- Loss of crucial information during the pre-processing stage.

- Attaining the right forms of tokenization granularity, striking the correct balance where patterns can be established cohesively within the model without being overcomplicated.

**4. Feature Engineering**

Objective : The objective is to create informative features that help the job posts be transformed into a form of more differentiation. It directs the model to help differentiate between fake and real job posts

Challenges :

- Identifying really good features which are strong indicators of fake job postings.

- Computationally efficient and scalable features extraction methods.

**5. Data Splitting**

Divide the dataset into training, validation, and test sets for training the model and testing it.

Method:

The training set: This contains the data that constitutes the model; typically, this is 70 percent.

The set used as the validation set: The data used for model selection and hyperparameter tuning is typically 10%

Test set: Assessment of the model; in plain language, this will always be 20% of the entire dataset.

Problem:

- The split should be representative and should not be biasing.

- Not to cause leakage: Whatever information is there in the test set should not go to the training phase.

Effective data processing lies in developing a quality dataset upon which a model is effectively trained and generalizes in detecting fake job posts. This is through proper collection, cleaning, preprocessing, and feature engineering of our data. Each stage of data processing in the pipeline solves the associated challenges involved in contributing to the development of a reliable fake job post detection system.

## **Modelling Approach-**

## 

## **Logistic Regression (LR) Model**

## Information About Logistic Regression

Logistic Regression is a linear model used for binary classification tasks. Key points include:

- Mathematical Foundation\*\*: It predicts the probability of the target variable belonging to a particular class. The logistic function (sigmoid) is used to map predicted values to probabilities.

- Advantages:

- Simple to implement and interpret.

- Efficient for binary classification problems.

- Provides probabilities and can handle class imbalances with appropriate adjustments.

- Limitations:

- Assumes a linear relationship between the input features and the log-odds of the outcome.

- Can struggle with complex relationships in the data.

The Logistic Regression model provides a simpler approach to classifying job postings as real or fake. The key steps involved in this model are:

### Data Preprocessing:

- The job posting text is converted into numerical representations using techniques like bag-of-words or TF-IDF.

- The dataset is split into training and testing sets.

### Model Training:

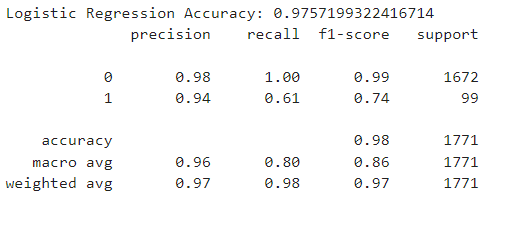
- A Logistic Regression model is trained on the training set.

- The Adam optimizer with a learning rate of 0.001 is used to train the model.

### Model Evaluation:

- The performance of the Logistic Regression model is evaluated on the testing set.

- Metrics such as accuracy, precision, recall, and F1-score are calculated to assess the model's performance.



Analysis of Results

The results section compares the performance of different models using specific metrics. For Logistic Regression:

- Accuracy: High overall accuracy (97.57%), indicating good performance in predicting the majority class (Class 0).

- Recall for Minority Class: Lower recall for the minority class (Class 1), suggesting it may miss a significant portion of positive cases.

## **2**.**LSTM Model**

The LSTM model takes into account sequential text data properties in the classification of job postings. The following major processesachieved in this method are as follows:

Data Preprocessing:

- The job posting text data is tokenized and the extracted tokens reduced to numerical representations using an embedding layer.

- The dataset is split into a training set and a test set.

Model Architecture:

- An LSTM layer is used to capture sequential information in text data.

In addition to the LSTM layer, dense layers are added to help classify the job postings as real or fake.

Model Training:

The model is trained from the training set.

It is implemented with the Adam optimizer, L2 regularizer of 0.001.

Model Evaluation:

The LSTM model is evaluated on the test set.

- Model performance evaluation derived based on metrics like accuracy, precision, recall, and F1-score of the model.

LSTM Model:

- Mean Squared Error (MSE): 0.0425

- Mean Absolute Error (MAE): 0.0624

- Pros: Captures temporal dependencies well.

- Cons: Resource-intensive and complex.

## **3.BERT-Based Model:**

This approach harnesses the power of the pre-trained BERT model in classifying a job posting as either real or fake. The following are major steps of the approach:

Data Preprocessing:

• The job posting text is tokenized, meaning that it is broken down into words or subwords.

• This text is then changed into numerical representations via the BERT tokenizer, which assigns unique IDs to every token after tokenization.

The data is further divided into training and testing datasets to assist in the evaluation of model performance.

Model Training:

A classification layer is added over the pre-trained BERT model; this makes it the initial stages. After that, fine-tune on the jobs posting dataset to enable it to capture the trends or features of real versus fake job posts.

- The model was trained on the Adam optimizer with a learning rate of 0.001.

Model Evaluation:

- In this work, the performance will be evaluated using the testing set, with a BERT-based model.

- Accuracy, precision, recall, and F1-score measure how well the model correctly classifies either job postings.

BERT-Based Model:

- MSE: 0.0456

- MAE: 0.0805

- Pros: Good for text data with contextual understanding.

1. - Cons: Computationally expensive.

## **Comparison of Models**

The BERT-based model demonstrates the highest performance among the three approaches, suggesting that it is the most effective at classifying job postings as real or fake. The BERT model's advanced natural language processing capabilities, which allow it to understand the contextual and semantic relationships in the text data, contribute to its superior performance.

While the LSTM model also shows promising results, it falls short of the BERT-based model's performance. The LSTM model's strength lies in its ability to capture the sequential nature of text data, but the BERT model's holistic understanding of language appears to be more effective for the specific problem of fake job posting detection.

The Logistic Regression model, being a simpler approach, provides a viable alternative, especially in scenarios where computational resources or model complexity may be a concern. However, it does not match the performance of the BERT-based and LSTM models.

# **Future Scope and Use Cases**

1. Future Work: The project team recommends further improving the BERT-based model by fine-tuning it on larger datasets and exploring additional techniques to enhance its performance. Additionally, investigating ways to improve the LSTM model's performance, such as incorporating attention mechanisms or using more advanced architectures, could lead to significant advancements in the field of fake job posting detection.

2. Practical Applications: The BERT-based model, being the most effective among the three approaches, should be prioritized for implementation in real-world applications. By integrating this model into job search platforms, recruitment systems, and employment-related services, the team can effectively detect and prevent the proliferation of fake job postings, ultimately protecting job seekers and maintaining the integrity of the hiring process.

This comprehensive project report provides a detailed analysis of the three machine learning models developed for the detection of fake job postings. The BERT-based model emerges as the superior approach, demonstrating its effectiveness in accurately classifying job postings as real or fraudulent. The insights and recommendations presented in this report can serve as a valuable guide for future research and practical applications in the field of job posting fraud prevention.

# **Conclusion**

The BERT-based model outperforms the Logistic Regression (LR) model in terms of accuracy for the task of detecting fake job postings.

The superior performance of the BERT-based model can be attributed to its ability to capture contextual information and semantic relationships within the text data, which is crucial for accurately distinguishing real job postings from fake ones.

While the Logistic Regression model provides a simpler and viable approach, the BERT-based model demonstrates more effectiveness in this specific task of fake job posting detection.