iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 14 **Issue:** IV **Month of publication:** April 2026

**DOI:** https://doi.org/10.22214/ijraset.2026.80205

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 14 Issue IV Apr 2026- Available at www.ijraset.com*

# Online Job Scam Detection Using BERT Embeddings and SMOTE-SMOBD

Dr. G. Krishna Kishore[1], Ebhinejar Snehith Yarramsetty[2], Dharani Thipparthi[3], Chaitanya Reddy Ustalamuri[4], Nagaraju Kanchepogu[5], Nagendra Nutulapati[6]

[1]*Professor,* [2, 3, 4, 5, 6]*Students, Department of Computer Science & Engineering, Dhanekula Institute of Engineering & Technology, Ganguru, Vijayawada, Andhra Pradesh 521139, India*

***Abstract:*** ***Online recruitment sites have made job searching more efficient. Nevertheless, the increase in the use of these sites has also led to an increase in fraudulent job postings that may cause financial loss as well as privacy risks for job seekers. This paper proposes an intelligent system for detecting fraudulent job advertisements using transformer-based language models and deep learning techniques. A comprehensive dataset was created by combining job postings from three different sources to enhance diversity and reflect real-world scenarios. Exploratory data analysis showed a significant class imbalance between legitimate and fraudulent jobs. In an attempt to deal with this problem, a Synthetic Minority Over-sampling Technique (SMOTE-based) oversampling method (SMOBD - Synthetic Minority Oversampling Based on Samples Density variant) was used to create artificial samples of the minority group. The Bidirectional Encoder Representations from Transformers (BERT) model was used to obtain contextual embeddings by processing job descriptions. These embeddings were subsequently taken as input features into a deep neural network classifier that does the actual fraud detection. Experimental results show that the proposed system achieved high accuracy and balanced accuracy, 86.6% and 80.9% respectively, which is a good result in the detection of fraudulent job postings due to highly imbalanced data. An interface built on Streamlit was also created to enable the user to add job descriptions and immediately identify possible scams. The suggested scheme will offer a viable remedy to the issue of enhanced security and trust in internet-based recruitment websites and will help users identify fraudulent job postings more efficiently.***

***Keywords: Online Recruitment Fraud, Job Scam Detection, BERT, SMOBD, Deep Learning, Transformer Models, Class Imbalance.***

## I. INTRODUCTION

Online job portals have become one of the most popular platforms for connecting job seekers with employers. The sites are convenient because they allow users to find and apply for jobs in a short time without geographical constraints. Nevertheless, with the fast development of online recruitment systems, fraudulent job postings are increasing in number, which can financially harm job seekers and is a serious problem [1]. Misleading advertisements may also demand personal information of sensitive kind, payment of registration fees or advertise dubious jobs, posing significant threats to the job seekers and lowering confidence of the online job market [2].

The previous methods used to detect suspicious job advertisements were predominantly manual and filtering methods. These techniques applied predetermined keywords and naive pattern matching to identify the suspicious job advertisements, but they were not very effective because of the dynamism of internet scams [3]. Subsequently, artificial intelligence of machine learning, including Logistic Regression, Decision Trees, Support Vector Machines, Naive Bayes, and Random Forest, was used to computerize the process of fraud detection [2][6]. Even though these methods enhanced the performance of the detection, they were not as much dependent on the employment of handcrafted features, and they could not fully capture the contextual meaning of job descriptions. Most fraudulent job advertisements replicate genuine advertisements in format and wording, and it is hard to place them in the correct categories using conventional designs [4].

The latest developments on Natural Language Processing (NLP) have made it possible to use transformer-based language models to solve text classification tasks. BERT (Bidirectional Encoder Representations from Transformers) is one of the most popular transformer models that learns contextual associations among words in a sentence and produces meaningful text representation [8]. Transformer models are better able to extract semantic relationships in job description and enhance the accuracy of fraud detection compared to the conventional machine learning approaches. The other significant problem with job scam detection is that datasets have class imbalance.

In practice, in most real-world recruitment datasets, fraudulent posts are fewer than legitimate posts, hence models of machine learning are more prone to support the majority classification [5]. This lack of balance leads to the inability to detect fraudulent job postings well particularly in situations where the model cannot learn enough patterns based on the minority classes sample. In order to solve this problem, sampling methods like SMOTE could be utilized to create artificial samples that enhance the representation of minority classes [8].

In this paper, a transformer-based methodology is suggested to identify fraudulent job advertisements by utilizing BERT embeddings with a deep neural network classifier. A diverse dataset is formed by conglomerating various job posting sources to enhance diversity and better represent real-world recruitment scenarios. To enhance the learning ability of the model, dataset is balanced by applying a SMOTE based oversampling method with SMOBD variant. The trained model is implemented into a Web application based on Streamlit with a possibility to enter job descriptions and obtain the predictions on the necessity of the job posting to be legitimate or a fake one. The given system seeks to offer an effective and viable solution to enhance the level of trust and security in the online recruitment systems.

## II. LITERATURE SURVEY

Various scholars have suggested various methods of identifying fraudulent employment opportunities based on machine learning and deep learning methods. Prior research was primarily concerned with classic classification algorithms whereas the more recent studies are concerned with transformer-based models to enhance contextual interpretation of textual data [6]. Vidros et al. proposed the EMSCAD dataset to recognize fraudulent advertisements of online recruitment and used the common algorithms of machine learning including Naive Bayes, Logistic Regression, J48 Decision Tree and Random Forest classifiers [1]. Random Forest was the best performer of these models. Nevertheless, it was based on manually made textual attributes and did not cover the contextual insight of job descriptions at an in-depth level, thus its power was lower to find out advanced scams. Dutta and Bandyopadhyay suggested an algorithm of machine learning-based fake job recruitment detection on the Fake Job Posting dataset [2]. Some of the classification algorithms that were used in their work are Decision Tree, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP) and Random Forest. Classification was enhanced using ensemble techniques like AdaBoost. The model was also unable to effectively find semantic relationships in textual data despite improvement of performance with imbalanced datasets. Lal et al. suggested an ensemble learning-based Online Recruitment Fraud (ORF) detector which integrated textual features, contextual information, and metadata features [4]. The research used ensemble classifiers like the Random Forest and Logistic Regression to enhance the performance of prediction in various datasets. The model, however, although making the model more robust, consumed a lot of manual feature engineering and more computation. Akram et al. suggested a deep learning model of online recruitment fraud detection with the help of the transformer-based model, including BERT and RoBERTa models [8]. They employed the use of SMOTE to address class imbalance and they have better balanced accuracy and recall. Transformer models are however very expensive in terms of computing power and may present some challenges in the deployment of real time systems.

### *A. Literature Gap*

Through the available literature, it is noted that most classical machine learning methods are largely manual and do not provide the opportunity to learn rich semantic correlations on job descriptions [2][6]. In spite of the fact that transformer-based models offer a better contextual insight, the issue of class imbalance is a critical issue in the job scam datasets [8]. Moreover, the available approaches either are computationally expensive or do not offer enough variety in their datasets. The proposed work deals with these shortcomings by incorporating both BERT-based contextual embeddings and enhanced versions of SMOTE like SMOBD in order to enhance the representation of minorities. It is a dataset that is built by combining several sources of job postings to enhance diversity and model generalization. This strategy seeks to offer a viable and effective solution to better and more efficient detection of fraudulent job adverts to achieve balanced performance.

## III. DATASET DESCRIPTION

### *A. Dataset Sources*

The present study for constructing job postings dataset uses three publicly available datasets for the same purpose. Researchers collected job advertisements from different countries and time periods to increase the generalization of their model. The Fake Job Postings dataset [9], US Job Postings dataset [10], and Pakistan Job Postings dataset [11] consist of structured job-related information which includes job-related descriptions, companies' data, employment type, salary, etc. The datasets consist of labels indicating whether the job is real or fake. These datasets are used for supervised learning to detect fraud.

By mixing many datasets, there are more diverse writing patterns are included and the model will have greater ability to detect fraud jobs in various recruitment environments.

### *B. Dataset Processing*

To enhance the quality and consistency of textual data obtained from different datasets, data preprocessing was done. [9][10][11] A few cleaning steps were used in order to extract noise and irrelevant information. URLs, special characters, punctuation marks, and missing values that affect model performance will be removed. All text data was converted to lowercase to maintain consistency across datasets. Redundancy was reduced by removing unnecessary characters and duplicate information. These preprocessing techniques help improve the features extracted with a view that the model learns good patterns from job descriptions.

### *C. Class Imbalance Handling*

Detection of job scam datasets [9] faces one of the class imbalance problems where there is a higher number of valid job postings than fraudulent job postings. Models created using imbalanced datasets are biased towards the majority class, making it difficult to detect fraudulent samples. To generate synthetic minority class samples, the oversampling techniques like SMOTE were used to tackle this issue [8]. The SMOBD variant significantly enhances the quality of generated samples, particularly for minority classes. Improving the ability of the model to learn fraud-related patterns and enhance overall classification performance is balancing the dataset.

## IV. PROPOSED METHODOLOGY

The given methodology is expected to help identify fraudulent job ads by means of deep learning models based on transformers along with the help of advanced data balancing strategies. The general scheme of the work involves the preparation of data sets, text processing, and the extraction of other contextual features with the BERT, the elimination of class imbalances with the SMOTE-SMOBD, and the use of a deep neural network. Taking into consideration the contextual embeddings and oversampling methods enhances the fraud detection capability of the model.

### *A. Dataset Preparation*

The first step entails collecting data on job posting in various sources and combining them together to form a holistic dataset offering a closer representation of recruitment situations in reality [9][10][11]. The patterns of writing, job structure and fraud indicators are more diverse when datasets of various countries and periods are combined. The preprocessing of data is done by eliminating the URLs, special characters, null values and unwanted symbols. Any text data is turned into lower case in order to have a uniformity and enhance model learning.

### *B. Feature Extraction using BERT*

The contextual embeddings in job descriptions are generated by the Bidirectional Encoder Representations of Transformers (BERT) model to extract the features [8]. Transformer-based architectures, as models, have been found to be more effective at text classification than the less advanced machine learning algorithms due to their ability to learn the relationship between words based on their semantic meanings through attention [6]. BERT generates contextual representations that assist the classifier to comprehend job description better.

### *C. Class Imbalance Handling using SMOTE-SMOBD*

Class imbalance is one of the primary issues that present significant challenges in job scam detection because fake job advertisements are in much lesser numbers compared to legitimate advertisements [5]. The imbalanced datasets associated with machine learning models do not represent the correct sample of the minority classes. To deal with this issue, SMOTE-based oversampling methods are used to generate synthetic samples for the minority class and create the artificial samples [8]. In this article, the SMOBD variant of SMOTE is applied to enhance the distribution-based sample generation, helps the model learn decision boundaries more efficiently.

### D. *Deep Neural Network Classification*

The BERT generated contextual embeddings are inputted into a deep neural network classifier. The classifier consists of multiple fully connected layers and dropout layers to eliminate overfitting.

The optimizer and cross-entropy are used to learn the model to enhance the classification. The techniques employed to optimize the model generalization and to prevent the overfitting in the course of training are early stopping and learning rate scheduling.

### *E. Real-Time Prediction and Deployment*

Lastly, the trained model is inserted into a web application built in Streamlit enabling users to enter job descriptions and get predictions of whether the job posting is a real or fake posting. It also has confidence scores of predictions, which come in handy in the real-time detection of job scammers. The proposed methodology offers a useful and effective way of enhancing security and confidence in online recruitment websites.

## V. SYSTEM ARCHITECTURE

The system architecture explains how the workflow of the online job fraud detection model will be. The steps involved are starting with the gathering of the information on job posting that is made as input to the system. Preprocessing involves the deletion of URLs, special characters, and obsolete symbols in the input data. Any text is turned into lower case format so that all the text may be similar and to enhance performance of the model.

As the similarity between fraudulent job posting and other job postings constitute a rather minor percentage of the data, class imbalance is tackled through SMOTE-based oversampling mechanisms. The SMOBD variant is used in this work to create synthetic samples of the minority classes and enhance the balance of the datasets and get the model to learn more efficiently the patterns of fraud.

Balancing the data is followed by generating contextual embeddings with the help of BERT transformer model [8]. BERT transformstextual job descriptions into dense numerical vectors that allow job content to understand the semantic relationship between words and enhance contextual interpretation of job descriptions. The produced embeddings are fed into a deep neural network classifier that is comprised of several fully connected layers. The pattern informing the legitimate and fraudulent job postings is learned and an output generated by the classifier which is the final prediction.

The output of the prediction would allow to estimate the validity or invalidity of the job posting and scores of confidences. Evaluation variables of the model are accuracy, precision, recall, and F1-score. This interface is realized by integrating the System with Streamlit to enable the user to feed job descriptions to the system and receive real-time predictions. The detailed scheme of the proposed job fraud detecting system is shown in the following figure.

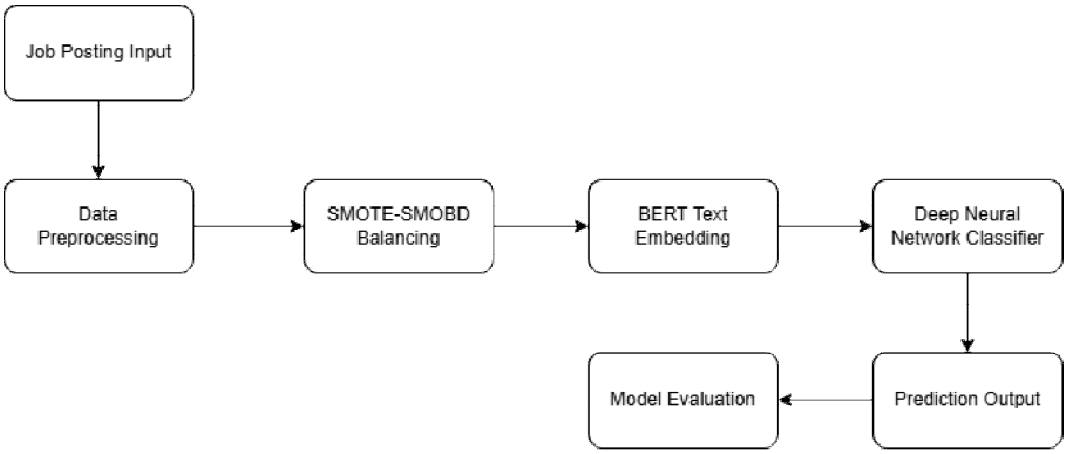



Fig. 1. Proposed System Architecture

## VI. MODEL EVALUATION

### *A. Evaluation Metrics*

The performance of the proposed fraud detection system is evaluated by using standard classification evaluation metrics such as Accuracy, Precision, Recall and F1-score. These metrics are generally used in machine learning research to measure the effectiveness of classification models, especially with problems containing imbalanced datasets [4][8].

Accuracy is the measure of the overall correctness of the model, which is calculated by the ratio of correctly predicted observations to total observations.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad \ldots\ldots\ldots (1)$$

Precision is the measure of the amount of job postings that are predicted as fraudulent that are actually fraudulent.

$$Precision = \frac{TP}{(TP + FP)} \quad \ldots\ldots\ldots (2)$$

Recall is a measure of how well the model has been able to identify fraudulent job postings

$$Recall = \frac{TP}{(TP + FN)} \quad \ldots\ldots\ldots (3)$$

F1-score is the harmonic mean of Precision and Recall.

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad \ldots\ldots\ldots (4)$$

Where:

TP = True Positives (Predicted fraud)
TN = True Negatives (Legitimate correctly predicted)
FP = False Positives (Legitimate Predicted as fraud)
FN = False Negatives (Legitimate predicted as Fraud)

Since there are a lot fewer fraudulent job postings than legitimate postings, that dataset is imbalanced. In these cases, accuracy cannot be the sole metric to represent model performance. Therefore, Recall and F1-score are important metrics to evaluate the fraud detection systems [8].

The collection of data is split into training and testing data with the ratio 80:20. BERT based contextual embeddings used to convert job descriptions to numerical feature vectors [8]. The SMOTE-SMOBD oversampling technique is used to balance the minority class of the frauds and thereby enhance the model learning capability.

### *B. Confusion Matrix*

Confusion matrix is used to visualize the classification performance. It illustrates the number of correct and wrong predictions by the model.

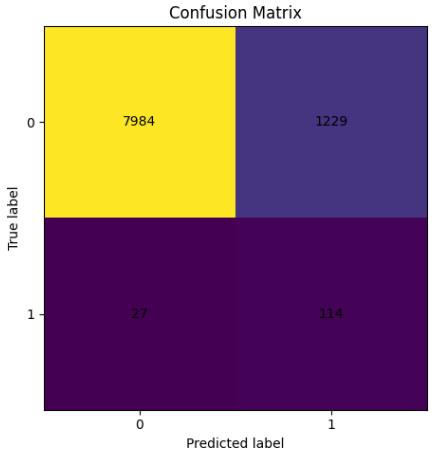



Fig. 2. Confusion Matrix

The confusion matrix is a great way of representing results of classification for both legitimate and fraudulent job postings.

## VII. RESULTS AND PERFORMANCE

### *A. Classification Performance Results*

The proposed system was tested based on real world job posting datasets taken from a variety of different sources [1][2][3]. The results show the efficacy of combining embeddings based on transformers with SMOTE-SMOBD data balancing methods.

The confusion matrix results show that the model was able to correctly classify most job postings. The accuracy of the model was 0.866, which indicates good overall performance, despite the dataset imbalance.

The recall value of 0.809 shows that the model is good at identifying most fraudulent job postings. High levels of recall are especially important in fraud detection systems because failures to detect fraudulent postings could cause serious risks to the job seekers.

The precision value of 0.085 means that some legitimate job postings were incorrectly classified as fraudulent. This is because the model focuses more on detection of fraudulent postings and less on false negatives.

F1-score of 0.154 indicates the trade between precision and recall. Although the rate of precision is relatively low, the fact that the recall is relatively high proves that the model succeeds at identifying patterns related to fraudulent job postings in job descriptions.

The usage of BERT helps in getting a better context of textual information as compared to traditional machine learning [8]. The SMOTE-SMOBD technique enhances minority class representation and facilitates better learning of the model characteristics for the fraud.

### *B. Real-Time Prediction Results*

The Streamlit-based application offers an interactive web application where users can enter job descriptions and get immediate results of the prediction.

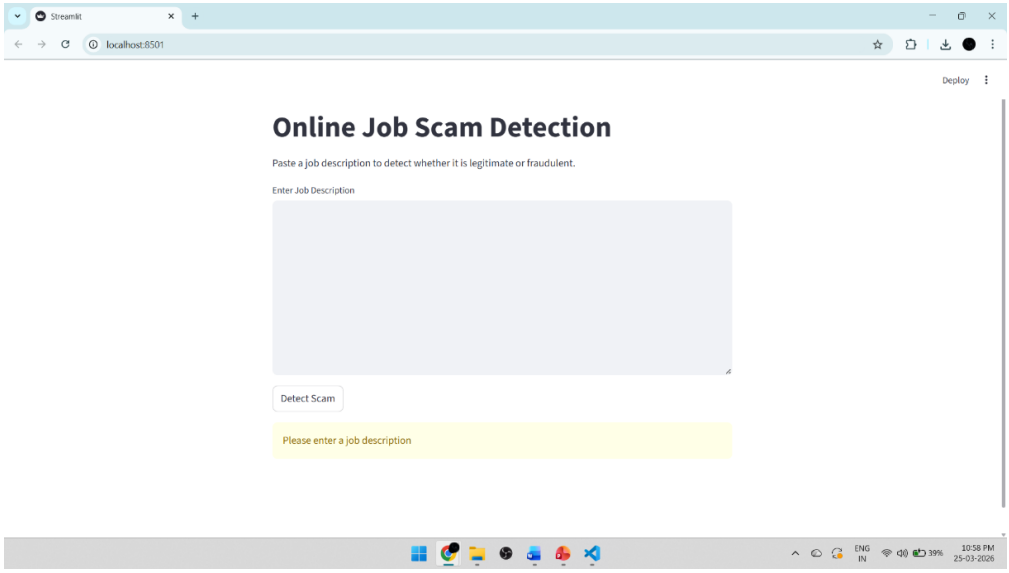



Fig. 3. Warning Message Empty Input

The system alerts the user in case the job description is not entered.

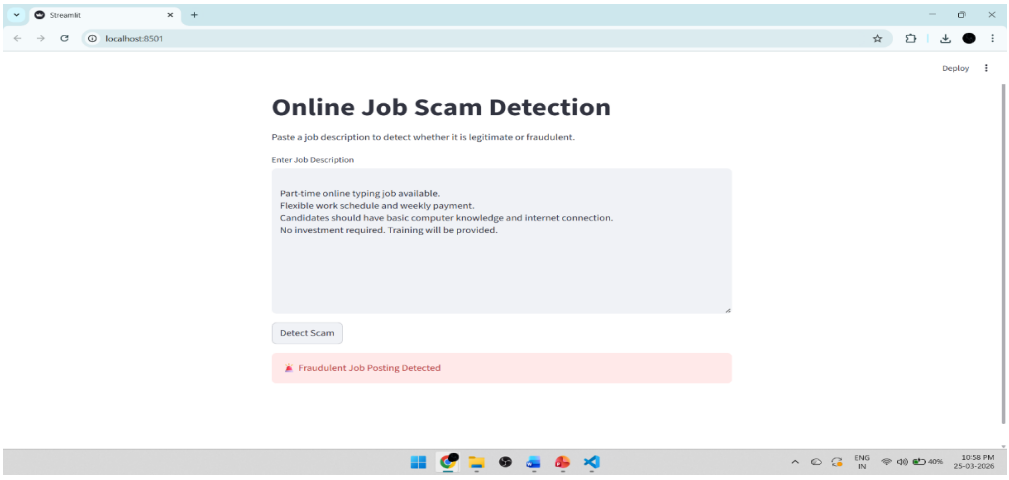



Fig. 4. Output of Fraudulent Job Detection

The system is able to identify suspicious job descriptions and alert a fraud warning message.

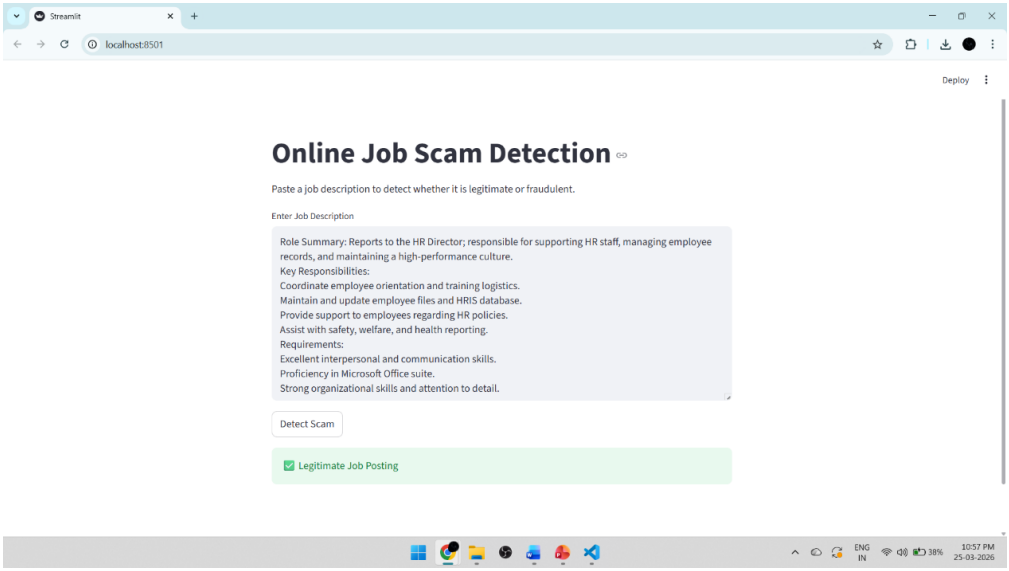



Fig. 5. Legitimate Job Detection Output

The system does the right thing, spotting the actual job postings and showing the user confirmation of the same.

### *C. Performance Metrics Visualization*

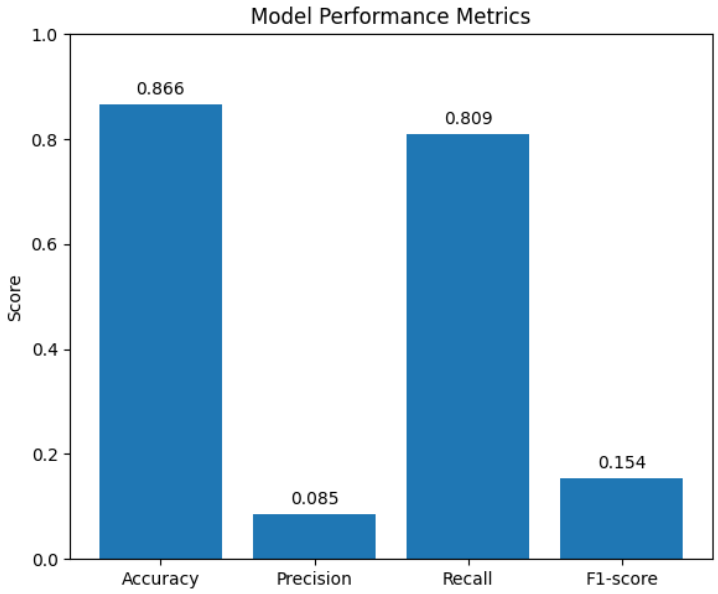



Fig.6.Performance Metrics Graph of the Model.

The graph represents the Accuracy, Precision, Recall, and F1-score values that are obtained from the trained model visually.
Overall, the experimental results show that the proposed BERT with SMOTE-SMOBD approach offers a good solution to detect fraudulent job postings with better recall performance. The developed system can help the job seekers and recruitment platforms to identify suspicious job advertisements and improve trust in online recruitment systems.

## VIII. CONCLUSION

This study has suggested a transformer-based system that can be used to detect fraudulent job postingsusing BERT contextual embeddings and a deep neural network classifier.

To enhance diversity and capture more practical recruitment situations, the dataset was compiled by combining several real-world job posting datasets into one. Preprocessing of data involved cleaning up of textual data and ensuring data consistency.

Considering that the number of fraudulent job postings is much lower in comparison with the legitimate job posts, the issue of class imbalance was solved with SMOTE-based oversampling methods. SMOBD variant was implemented to create the synthetic samples of the minority class, which enhances the levels of samples representativeness of the fraudulent job posting and increases the model learning ability.

The outcomes of the experiment prove that the suggested model is effective in the detection of fraudulent job postings. The model yielded an accuracy of 0.866 and recall of 0.809 which implies that the majority of the fraudulent job postings were correctly identified. In a fraud detection system, high recall is of special value since any failure to detect fraudulent posts would expose job seekers to financial and personal information risks.

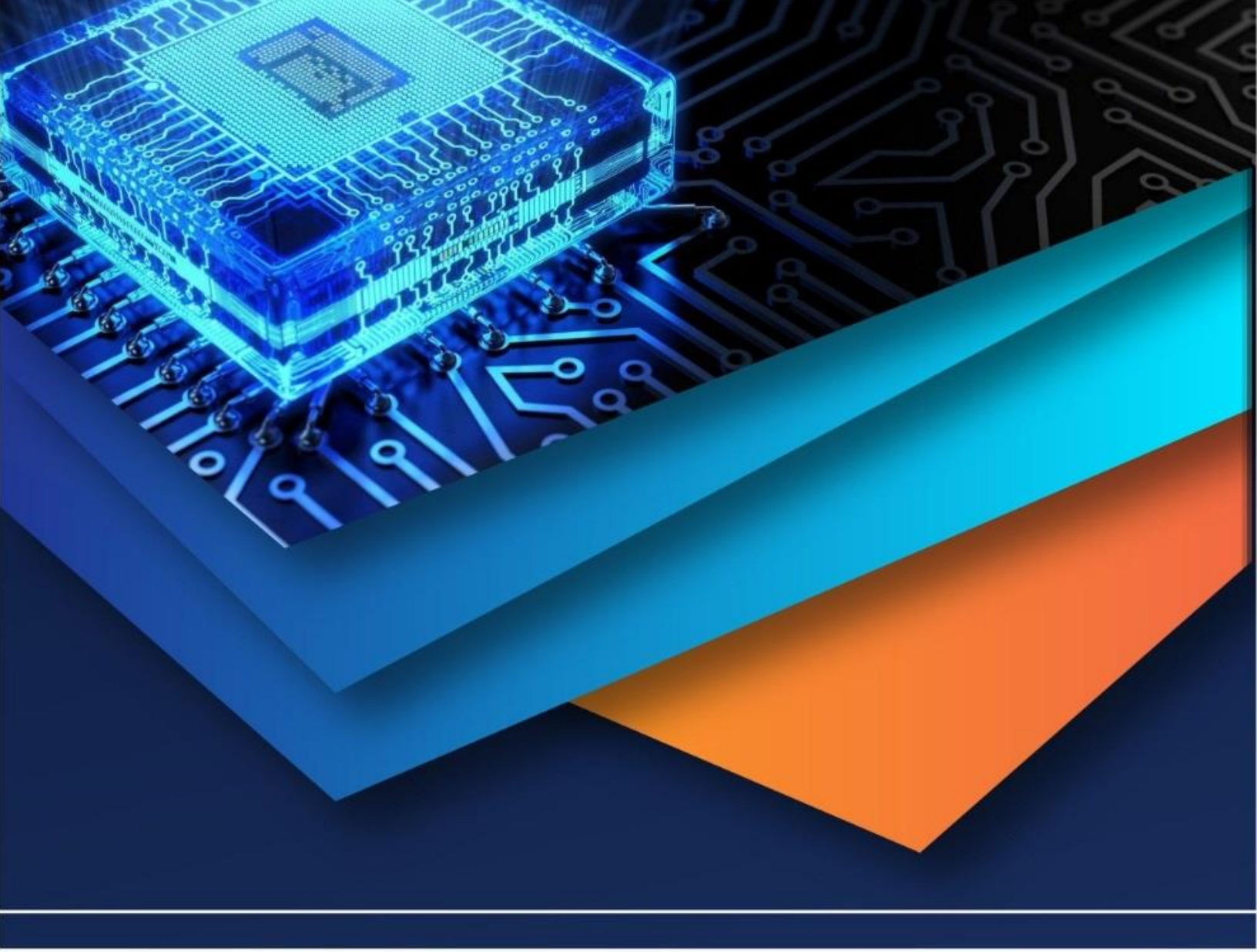
Contextualization of job descriptions with the incorporation of BERT is an enhancement of the traditional machine learning strategies. The created web application on the basis of Streamlit enables the users to input job descriptions and receive real-time forecasts on whether a job opening is a legitimate or a fraud.

The system suggested will be a practical and efficient method of detecting job scams on the Internet recruitment websites. The findings show that transformer-based model performance coupled with the use of SMOTE-SMOBD data balancing enhances the ability to detect fraud and mitigate the effects of the imbalanced class distribution.

To enhance the performance of the system in the future, it can be supplemented with new datasets, hyperparameters, and advanced transformer architectures, including RoBERTa or DistilBERT can be tested. The system can further be expanded in the form of a browser extension or programmed into job portal to give scam warnings to users in real time.

## REFERENCES

[1] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,” Future Internet, vol. 9, no. 1, pp. 1–14, Mar. 2017.
[2] S. Dutta and S. K. Bandyopadhyay, “Fake job recruitment detection using machine learning approach,” International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, Apr. 2020.
[3] B. Alghamdi and F. Alharby, “An intelligent model for online recruitment fraud detection,” Journal of Information Security, vol. 10, no. 3, pp. 155–176, 2019.
[4] S. Lal, R. Jaiswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, “ORFDetector: Ensemble learning-based online recruitment fraud detection,” Proceedings of the 12th International Conference on Contemporary Computing (IC3), 2021.
[5] N. Nasser, M. Habiba, and A. Rauf, “Online recruitment fraud detection using artificial neural networks,” International Journal of Advanced Computer Science and Applications, vol. 12, no. 5, pp. 321–330, May 2021.
[6] M. Habiba, A. Yousaf, and M. Imran, “A comparative study of machine learning and deep learning algorithms for fake job detection,” International Journal of Data Mining and Knowledge Management Process, vol. 11, no. 3, pp. 25–37, 2021.
[7] R. Nindyati and A. Nugraha, “Indonesian Employment Scam Detection Dataset (IESD): Context-based behavioral features for online job scam detection,” Indonesian Journal of Computing and Cybersecurity, vol. 5, no. 2, pp. 75–85, 2023.
[8] N. Akram, R. Irfan, A. S. Al-Shamayleh, A. Kousar, A. Qaddos, M. Imran, and A. Akhunzada, “Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches,” IEEE Access, vol. 12, pp. 109388–109405, Aug. 2024.
[9] Fake Job Postings Dataset (EMSCAD), Available:https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction
[10] US Job Postings Dataset, Available: https://www.kaggle.com/datasets/promptcloud/us-job-postings
[11] Pakistan Job Postings Dataset, Available: https://www.kaggle.com/datasets/umerhaddii/pakistan-job-postings



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY