ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2s/1452-1462
Borkar Nithya Rao *et. al.*, / International Journal of Engineering & Science Research

Online Recruitment Fraud (ORF) Detection System

M.Vineela, Borkar Nithya Rao, Kamati Poojitha
[1] Associate Professor, Department Of Cse, Bhoj Reddy Engineering College For Women, India.
[2,3]B. Tech Students, Department Of Cse, Bhoj Reddy Engineering College For Women, India.

ABSTRACT
The rise of digital platforms for job recruitment has brought with it a growing threat of fraudulent job postings, undermining the trust and safety of online hiring systems. This paper proposes an advanced fraud detection system based on the ALBERT (A Lite BERT) model to identify fraudulent job postings. The system will utilize a dataset created by merging job postings from multiple sources to better capture both legitimate and fraudulent job listings. A comprehensive pre-processing pipeline, including data cleaning and feature engineering, will be applied to prepare the data for model training. The system will incorporate various techniques, such as SMOTE (Synthetic Minority Over sampling Technique), to address class imbalance. The ALBERT model will then be fine-tuned to classify job postings, and the system will be evaluated using standard performance metrics like accuracy, precision, recall, and F1-score. The goal is to provide an effective and scalable solution for detecting fraudulent job postings in real-time, enhancing the overall security and trustworthiness of online recruitment platforms. The accuracy of the proposed technique is 99.82% which is higher than other well-known existing techniques.

1. INTRODUCTION

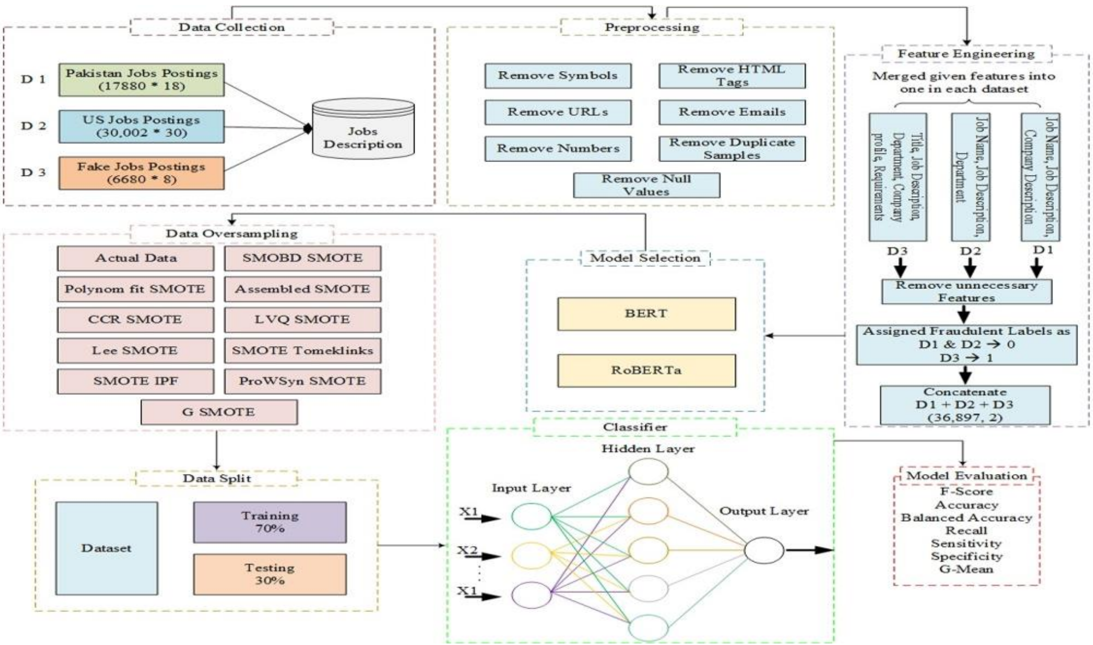
In the age of advanced technology, the internet has drastically transformed our lives in different ways. The traditional way to do any activity has now been switched online. Therefore, seeking a job and hiring employees have also switched online. An online recruitment system (E-recruitment) is an internet application, the benefits of which encompass productivity, easiness, and efficacy [1]. Most organizations prefer online recruitment systems to provide job opportunities to potential candidates [2]. Organizations publish job ads for their vacant positions through job portals, in which they mention job descriptions, including requirements, salary packages, offers, and facilities to be provided. Job seekers visit different online job advertising websites, seek job ads related to their interests, and apply for suitable jobs. The company then screens the CVs of applicants matching their requirements. The position is closed after fulfilling other formalities like interviewing and selecting potential candidates. The trend of posting online job advertisements was inflated during the global pandemic of COVID 2019. According to the World Economic Outlook Report, the International Monetary Fund (IMF) estimated that the unemployment rate increased to 13% at the peak time of the COVID-19 pandemic in 2020. These statistics were only 7.3% in 2019 and 3.9% in 2018. During the outbreak, many companies decided to post job openings online to provide facilities to job seekers [3]. But, where a facility is provided to the public, it also allows online fraudsters to take advantage of their pessimism. An employment scam is one of the considerable problems in the realm of online recruitment fraud (ORF). Although an online recruitment system benefits job seekers and recruiters, it can also be deleterious for them if it is not administered carefully. It is inauspicious for job seekers in terms of losing their privacy, money, or even their current job sometimes. Moreover, fraudsters also breach the credibility of well-reputed companies by defacing their reputation in the job market [4]. The fraudsters are using sophisticated methods to involve people in the scam, and making it very difficult for them to distinguish between real/fake job advertisements. According to the survey conducted by Flex Jobs [5], about 52% of the aspirants did not know ORFs, whereas the rest had only preliminary knowledge about them. Another survey recently accompanied by Action Fraud [6], it is investigated that more than 67% of people are now interested in looking for a job online. Still, they need to be aware of the increased number of job scams.

Multiple studies were conducted to detect ORF. Authors in [7] and [8] applied traditional machine learning algorithms to classify job postings as fraudulent/non-fraudulent. The work

[9] and [10] used ensemble-based machine learning techniques to improve classification accuracy. The authors in [11] first performed down sampling to handle the imbalance problem and then used an Artificial Neural Network (ANN) based model for classification. Authors in [12] extracted features by using TF-IDF, and after oversampling data, applied Random Forest (RF) to improve accuracy. Researchers in [13] created their own dataset and proposed context-based behavioural features to test them on conventional machine learning algorithms to get predictions. Upon reviewing the underlying study, it is noticed that many machine learning approaches have been used for ORF detection. Nowadays, the trend is moving towards implementing transformer-based deep learning techniques to get promising results compared to traditional machine learning algorithms; however, for ORF detection advanced deep-learning approaches have yet to be explored in their full capacity to solve this problem. Therefore, this research aims to analyse and alarm people about rapidly growing employment scams and to detect ORF by implementing transformer-based deep learning models. Consequently, people would not fall into the trap of job scams anymore. So by detecting ORF, the people wasting their time and money on those fraudulent activities can be more careful. In this research, we presented a novel dataset of fake job postings labelled as ‘‘fraudulent’’ for fake job postings and ‘‘non-fraudulent’’ for legitimate job postings. The proposed data is a combination of job postings from three different sources. We use ‘‘Fake Job Postings1 as a primary dataset and add publicly available job postings of Pakistan2 and the US3 to extend the dataset with the latest job postings. We have done this because the existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the capability of existing models in detecting fraudulent jobs. After preparing the dataset, Exploratory Data Analysis (EDA) was performed on this data. Through EDA, it was identified that the dataset has an imbalanced class distribution. Imbalance class distribution can be defined as the ratio of the number of samples in the minority class to the number in the majority class [14]. It may cause high predictive accuracy for frequent classes and low predictive accuracy for infrequent classes. Class imbalance problem occurs in various real-world domains, including anomaly detection [15], face recognition [16], medical diagnosis [17], text classification [18], and many others. SMOTE[19]gained extensive popularity as an oversampling technique. Almost 85 different

SMOTE variants have been introduced in the literature and are recently used by various researchers to handle class imbalance problems in multiple domains.

2-PROPOSED METHODOLOGY



This section discusses the different phases involved in the underlying research. Firstly, datasets from three different sources are integrated to propose a final version of the dataset. An Exploratory Data Analysis (EDA)is performed to identify that the dataset has an imbalanced class distribution. A detailed discussion is given in the section III-C to show the importance of different features. Second, necessary steps in the preprocessing phase are performed on the proposed data. The special symbols, URLs, emails, numbers, HTML, tags, duplicate records and samples that contain null values are removed in the preprocessing phase to clean the dataset.

Thirdly at the feature engineering phase, only required and relevant features are selected and merged as a single feature named ''Job Content''. This process is repeated for each dataset D1, D2, and D3 as shown in Fig. 2. Then, fraudulent and non-fraudulent labels are assigned as D1, D2 to '0' for non-fraudulent jobs and D3 to '1' for fraudulent job posting. Later in the next step of the feature engineering phase, all three datasets D1, D2, and D3 are concatenated to generate a finalised dataset The dataset is encoded in phase four through BERT/RoBERTA to generate the contextual vectors. Then data is augmented using different SMOTE variants to get a balanced class distribution in the fifth phase. We chose to use only the encoder part of the BERT/RoBERTa model because contextual information across entire sequences is essential for ORF detection. The ability of the BERT/RoBERTa model to grasp long- range dependencies is particularly relevant for identifying subtle patterns indicative of fraudulent activities. Moreover, it has been found in the literature review that the tasks requiring contextual understanding, such as natural language processing, demonstrate the superior performance of transformer-based model Lastly, classification is performed to detect fraudulent job postings.

Data Collection

The first step in the methodology is to collect a labeled dataset of job postings. The dataset should consist of job escriptions, each labeled as either **fraudulent** or **non- fraudulent**. Ideally, the dataset should include diverse job categories, locations, and industries to ensure the model can generalize well across various domains. The fraudulent labels should be based on common fraud indicators, such as promises of unrealistic pay, vague job descriptions, or requests for sensitive personal information.

ORF DETECTION TECHNIQUES

Several studies have focused on detecting fake job postings using machine learning and deep learning techniques, primarily on the EMSCAD dataset. Vidros et al. tested multiple classifiers, with Random Forest (RF) achieving the highest precision of 91.4%, and using empirical rules further improved results. Dutta and Bandyopadhyay found RF outperformed other ensemble classifiers, achieving 98.27% accuracy, while Decision Tree (DT) achieved 97.2% among single classifiers. Alghamdi and Alharby utilized Support Vector Machine (SVM) for feature selection and RF for classification, achieving 97.2% precision. Lal et al. applied ensemble methods, reaching 95.5% accuracy with their ORF Detector. Nasser et al. tackled data imbalance with down sampling and used Artificial Neural Networks (ANN) to achieve 93.64% accuracy. Deep learning models, like the DNN used by Habiba et al., demonstrated superior accuracy, reaching 99%. Lokku et al. enhanced results by balancing the dataset and applying TF-IDF for feature extraction, achieving 99% accuracy with RF. Meanwhile, Nindyati and Nugraha introduced the IESD dataset and achieved 90% accuracy using context-based behavioral features. Across studies, RF consistently performed well, with deep learning models offering the highest accuracy. Addressing data imbalance and feature selection were key to improved outcomes.

Researchers have explored various techniques to address class imbalance in datasets. Gosain and Sardana proposed oversampling methods like SMOTE, Borderline SMOTE, ADASYN, and Safe Level SMOTE (SLSMOTE), with SLSMOTE outperforming others across six datasets using models like NB, KNN, and SVM. Akhbardeh et al. used methods such as undersampling, oversampling, and feedback loops on logbook datasets and models like BERT, LSTM, CNN.

CRITICAL ANALYSIS

Online Recruitment Fraud (ORF) detection has relied heavily on machine learning, but advanced deep learning approaches remain underutilized. Employment scams are increasing rapidly, causing significant harm to job seekers, including financial loss and privacy breaches. Detecting illegitimate job postings is essential to protect candidates and ensure access to genuine opportunities from authentic companies

Despite high classification accuracies in prior work, poor recall and class imbalance issues remain critical challenges. High accuracy often misrepresents performance by favoring the majority class while neglecting the minority class. Balanced evaluation metrics, like balanced accuracy and recall, are crucial for capturing the true scenario. Exploratory Data Analysis (EDA) reveals a significant class imbalance in the dataset, prompting the selection of top-performing SMOTE variants for experimentation. Advanced deep learning methods and robust balancing techniques will be employed to address these challenges, as detailed in the subsequent methodology section.

The bar chart in Figure 3 illustrates the class distribution of real versus fake job posts in a dataset. The x-axis represents the two classes: "0" for real job posts and "1" for fake job posts, while the y-axis indicates the count of posts in each category. The chart reveals a significant class imbalance, with the majority of posts classified as real (class 0) and only a small fraction as fake (class 1). This imbalance highlights the rarity of fraudulent job posts compared to legitimate ones, which is a common challenge in fraud detection systems. Addressing this imbalance is crucial for building an effective and unbiased fraud detection model.

3-IMPLEMENTATION

The implementation detail of BERT/RoBERTa models, as done in our case, is given below: First, the input jobs description is preprocessed to match the required input format expected by the BERT/RoBERTa model. For this purpose, the Byte Pair Encoding (BPE) algorithm is used for tokenizing the text into sub words, and then the sub words are converted to their correspond ing numerical representations. Next, a fully connected dense neural network is added on top of the pre-trained BERT/RoBERTa model for the classification task. The flow diagram of the dense network is shown in Fig. 8. A linear layer on top of the final hidden state of the BERT/RoBERTa model passed the feature vector to the following dense layers for training the classifier to predict whether a job is fraudulent or not.

TABLE 1. Parameter description of the BERT and RoBERTa.

Model	Parameters	Value
BERT/ RoBERTa	No. of transformer layers	12/24
	No. of attention heads	12/16
	Hidden units	768
	Total parameters	110M
	Max sequence length	512
	Vocabulary size	30,000 (default)
	Dropout rate	0.1
Dense Network	No. of Neurons	128
	Hidden layers	02
	Batch size	32
	No. of Epochs	10
	Learning rate	0.001
	Activation function	Sigmoid
	Loss function	Binary cross entropy
	Optimizer	Adam

For fine-tuning the hyperparameters, a default learning rate of 0.001, a batch size of 32, and an early stopping mechanism are used during fine-tuning to avoid overfitting or underfitting. The selected parameter values for the BERT/RoBERTa are adopted from default configurations that are empirically validated for broad applicability in the natural language processing domain The choice of transformer layers, attention heads, and dropout settings optimizes both convergence and generalization, , making these parameters ideal for diverse applications without extensive customization.

After running a series of experiments, it has been found that the provided parameters give better performance So, we have chosen the parameters for BERT/RoBERTA model sand Dense net work as specified in Table 1. The pre-trained BERT/RoBERTa model is then fine-tuned on our dataset. During fine-tuning, the weights of the pre-trained model are updated based on the labelled data, while the pre-trained weights are used as a starting point to minimize the loss function on the task-specific labelled dataset. After BERT/RoBERTa model has been fine-tuned on the classification task, it is tested for inference on test data and outputs a predicted label based on the learned representation of the input text.

The process is visually supported by diagrams depicting the data preprocessing pipeline, model architecture, and evaluation results. This comprehensive approach ensures effective identification of fraudulent job postings while maintaining fairness across class predictions. Future enhancements may include the integration of hybrid oversampling techniques, multilingual datasets, and advanced explainable AI methods to broaden the system's applicability and accuracy.

The implementation of online recruitment fraud detection using deep learning involves preprocessing job descriptions to prepare them for model input. Tokenization methods like Byte Pair Encoding (BPE) convert text into subwords and numerical representations. Pretrained models such as BERT and RoBERTa are fine-tuned with task-specific labeled data to classify job postings. A dense neural network is added atop these models for classification. Hyperparameters, including learning rate and batch size, are fine-tuned to optimize performance. To address class imbalance, Synthetic Minority Oversampling Techniques (SMOTE) variants are applied, enhancing model recall and reducing Type II errors. The models are evaluated for accuracy, recall, and balanced metrics, revealing BERT combined with SMOBD SMOTE and RoBERTa with G SMOTE as particularly effective configurations for fraud detection. These approaches demonstrate the potential of deep learning and data augmentation to identify fraudulent job postings accurately and mitigate risks for job seekers.

The combination of advanced deep learning models, effective preprocessing, and robust class balancing techniques demonstrates a reliable framework for online recruitment fraud detection, ensuring safer experiences for job seekers and enhancing trust in online job platforms.

To enhance the detection capability, a dense neural network is added to the output layer of the pretrained models for classification purposes. During fine-tuning, the models' weights are adjusted using task-specific data to optimize the loss function, ensuring that the model learns to differentiate between real and fake job postings effectively. Hyperparameter optimization, including adjustments to the learning rate, batch size, and the use of early stopping mechanisms, is performed to balance convergence speed and generalization.

The implementation of an online recruitment fraud detection system involves a systematic approach to data collection, preprocessing, class imbalance handling, model architecture development, and deployment. Data is gathered from reliable sources, analyzed for imbalances, and preprocessed by cleaning, tokenizing, and converting text into numerical formats. To address class imbalance, various SMOTE variants, including Borderline-SMOTE and SMOBD, are applied. Transformer-based models like BERT and RoBERTa are fine-tuned, incorporating dense layers and optimized with hyperparameters such as a learning rate of 0.001 and early stopping. Evaluation metrics like balanced accuracy and recall ensure unbiased performance across classes, identifying the optimal combination of SMOTE variants and models. The fine-tuned system is deployed for real-time fraud detection via APIs, supported by diagrammatic representations of the process and performance. Future enhancements may include hybrid models and multilingual datasets for scalability and improved efficiency

Transformer-based models like BERT and RoBERTa are fine-tuned, leveraging pre- trained weights, with additional dense layers for classification. Hyperparameters like learning rate, batch size, and early stopping ensure optimal training performance. Evaluation metrics such as balanced accuracy, recall, F1-score, and confusion matrix

are used to compare imbalanced and balanced datasets, identifying the best SMOTE variant and model combination. Finally, the model is deployed as a real-time fraud detection service, complete with visual representations of process flows, architectures, and evaluation metrics to ensure transparency and reliability. For future scalability, hybrid models and multilingual datasets are recommended to enhance the system's performance further.

4-EXPERIMENTAL RESULTS

TABLE 2. Confusion metrics of BERT+SMOTE variants.

Methodology	Actual Real vs. Predicted Real	Actual Real vs. Predicted Fake	Actual Fake vs. Predicted Real	Actual Fake vs. Predicted Fake
BERT + Actual Data	10827 99.79%	22 0.21%	**152** **68.77%**	69 31.23%
BERT + Polynom fit SMOTE	10586 97.57%	263 2.43%	57 25.79%	164 74.21%
BERT + Assembled SMOTE	10643 98.10%	206 1.90%	46 20.81%	175 79.19%
BERT + LVQ SMOTE	10802 99.56%	47 0.44%	114 51.58%	107 48.42%
BERT + CCR SMOTE	10808 99.62%	41 0.38%	117 52.94%	104 47.06%
BERT + SMOBD SMOTE	10560 97.33%	289 2.67%	**39** **17.64%**	182 82.36%
BERT + ProWSyn SMOTE	10515 96.92%	334 3.08%	54 24.43%	167 75.57%
BERT + Lee SMOTE	10820 99.73%	29 0.27%	108 48.86%	113 51.14%
BERT + G SMOTE	10637 98.04%	212 1.96%	47 21.26%	174 78.74%
BERT + SMOTE Tomeklinks	10590 97.61%	259 2.39%	46 20.81%	175 79.19%
BERT + SMOTE IPF	10742 99.01%	107 0.99%	79 35.74%	142 64.26%

ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2s/1452-1462
Borkar Nithya Rao *et. al.*, / International Journal of Engineering & Science Research

This
section is divided into two parts. The first part contains type error analysis performed to study the impact of using SMOTE variants on the predictive models. Transformer based classification models, i.e., BERT and RoBERTa were implemented on imbalanced and balanced data, and the achieved results are compared in the second part. Different evaluation metrics, as discussed in the previous section, have been used to measure the performance of implemented frameworks. All implemented approaches showed uptomark performances.

TYPE ERROR ANALYSIS

The type error analysis conducted in this study focused on evaluating the impact of different SMOTE oversampling techniques on the performance of BERT and RoBERTa models for online recruitment fraud detection. The analysis specifically targeted reducing **Type II errors**, where fake job postings are incorrectly classified as real, as these errors pose significant risks to job seekers and also there will be financial losses for the unemployment job seekers.

For the **BERT model**, Type II error rates on the original imbalanced dataset were high, reaching 68.77%. However, employing SMOTE variants drastically reduced these errors. Among the variants, **SMOBD SMOTE** achieved the best performance, reducing the error rate to 17.64%, thanks to its ability to generate synthetic data points based on actual data density and distribution. Other notable variants, such as **G SMOTE** and **Assembled SMOTE**, also significantly reduced to Type II errors to 21.26% and 20.81%, respectively.

For the **RoBERTa model**, the original dataset resulted in a 100% Type II error rate, indicating complete misclassification of fake job postings. After applying SMOTE variants, substantial improvements were observed. **G SMOTE** reduced the error rate to 15.38% by generating diverse and balanced synthetic instances, while **SMOTE IPF** and **Assembled SMOTE** also achieved low error rates of 15.85% and 15.83%, respectively.

TABLE 3. Confusion matrices of RoBERTa with various SMOTE variants.

Methodology	Actual Real vs. Predicted Real	Actual Real vs. Predicted Fake	Actual Fake vs. Predicted Real	Actual Fake vs. Predicted Fake
RoBERTa + Actual Data	10849 100%	0	**221 100%**	0
RoBERTa + Polynom fit SMOTE	9892 91.17%	957 8.83%	75 33.93%	146 66.07%
RoBERTa + Assembled SMOTE	8546 78.77%	2303 21.23%	35 15.83%	186 84.17%
RoBERTa + LVQ SMOTE	10342 95.32%	507 4.68%	145 65.61%	76 34.39%
RoBERTa + CCR SMOTE	9021 83.15%	1828 16.85%	86 38.91%	135 61.09%
RoBERTa + SMOBD SMOTE	8974 82.71%	1875 17.29%	40 18.09%	181 81.91%
RoBERTa + ProWSyn SMOTE	9703 89.43%	1146 10.57%	63 28.50%	158 71.50%
RoBERTa + Lee SMOTE	9607 88.57%	1240 11.43%	57 25.79%	164 74.21%
RoBERTa + G SMOTE	8820 81.29%	2029 18.71%	**35 15.83%**	186 84.17%
RoBERTa + SMOTE Tomeklinks	9160 84.43%	1689 15.57%	51 23.07%	170 76.93%
RoBERTa + SMOTE IPF	8550 78.80%	2299 21.20%	36 15.85%	187 84.62%

MODELS CLASSIFICATION RESULTS

The classification results highlight the performance of BERT and RoBERTa models for online recruitment fraud detection. Initially, the models achieved high accuracies on imbalanced data (98.42% for BERT and 98% for RoBERTa) but with low recall values (65.50% and 50%, respectively), indicating class bias. To address this, SMOTE variants were applied to balance the dataset.

BERT's superior results stem from its ability to capture contextual and semantic relationships bi-

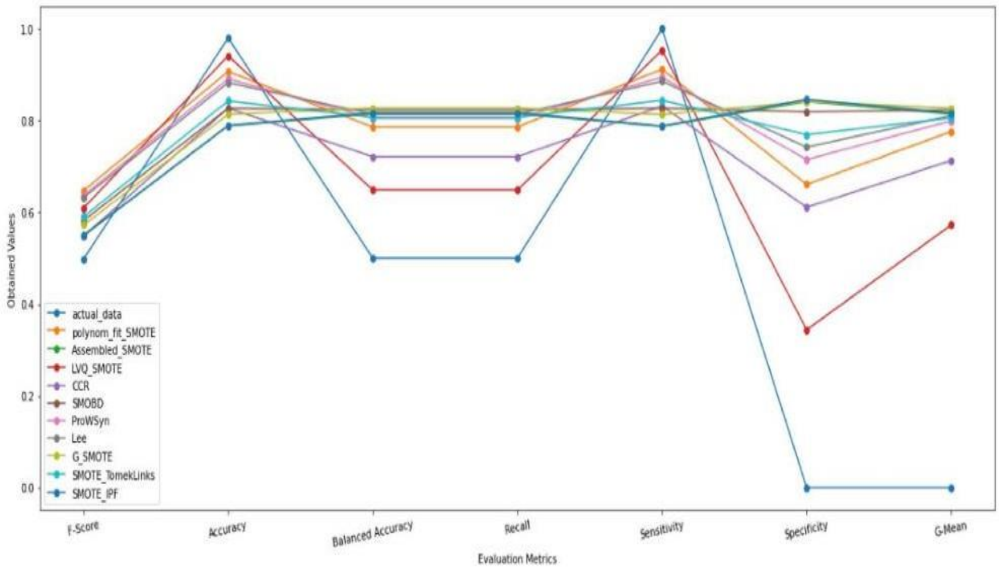


FIGURE 10. Combined evaluation metrics graph of RoBERTa+SMOTE variants.

Balancing the data improved recall significantly across all variants, though accuracies slightly decreased. Notably, **BERT with SMOBD SMOTE** achieved the highest recall of **90%**, while **RoBERTa with G SMOTE** achieved a recall of **82.73%**, demonstrating optimal performance after balancing.

directionally, making it particularly effective for this task.

These findings confirm that balancing data using appropriate SMOTE techniques mitigates class bias and enhances the model's capability to identify fraudulent job postings effectively.

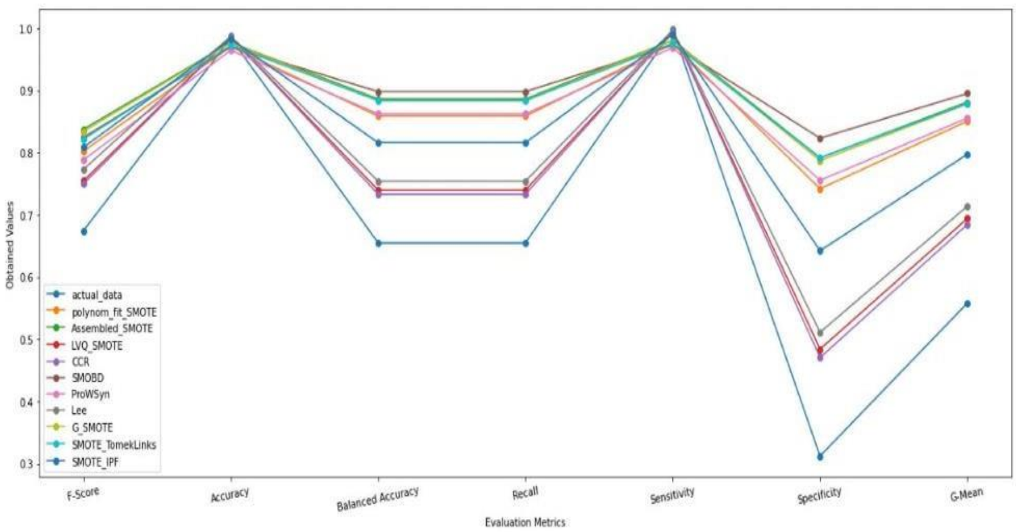


FIGURE 9. Combined evaluation metrics graph of BERT+SMOTE variants.

CRITICAL REVIEW DISCUSSION

The critical review highlights that while initial results showed high accuracies, they were misleading due to poor recall values, indicating class bias. The use of SMOTE variants significantly improved recall, achieving a better balance between precision and sensitivity for both BERT and RoBERTa models. However, some SMOTE techniques led to reduced accuracy, especially for RoBERTa, underscoring the trade- off between accuracy and recall. Despite improvements, certain techniques still fell short of achieving desired recall levels, suggesting scope for further optimization. This emphasizes the importance of addressing class imbalance in predictive modeling to ensure reliable and unbiased results.

COMPARATIVE ANALYSIS

The comparative analysis reveals that different deep learning architectures, when enhanced with SMOTE variants, exhibit varied yet notable improvements in detecting Online Recruitment Fraud (ORF). CNN models excel with SMOTE variants like Assembled SMOTE and G SMOTE, achieving the highest G-mean (89.21%) and F- score (85.51%), indicating their ability to handle class imbalances effectively in structured data. RNN models, particularly with ProWsyn SMOTE and SMODB SMOTE, displayed exceptional recall (89.01%) and balanced accuracy (89.54%), emphasizing their strength in minimizing false negatives. Transformer models, such as BERT and RoBERTa, benefit significantly from SMOTE variants. For example, BERT + ProWsyn SMOTE provides robust performance.

The comparative analysis of deep learning models (RNN, CNN, RoBERTa, and BERT) with various SMOTE variants for online recruitment fraud (ORF) detection shows significant performance differences. CNN models, particularly with Assembled SMOTE and G SMOTE, achieved the highest G-mean (89.21%) and F-score (85.51%), respectively. RNN models improved greatly with ProWsyn SMOTE (highest recall of 89.01% and F-score of 80.25%) and SMODB SMOTE (highest G- mean of 85.13%). Transformer-based models like BERT and RoBERTa showed diverse performance, with BERT + ProWsyn SMOTE offering balanced accuracy and robust F-scores. Overall, specific SMOTE variants significantly enhance model performance, with CNNs excelling in balanced metrics and RNNs effectively capturing sequential dependencies, while BERT and RoBERTa outperform in handling long-range dependencies.

5-CONCLUSION

In this research, the problem of ORF detection is analyzed thoroughly. This paper presented a novel dataset of fake job postings. The proposed data is a combination of job postings from three different sources. Upon conducting EDA, it was discovered that the class distribution within the collected dataset was highly imbalanced. The experiments performed in this research can provide valuable directions to job seekers and reputed organizations to better understand fact-based insights about employment scam and their effects on society. Conventional fraud detection without considering class imbalance problems can lead to misleading conclusions for both job-seekers and organizations. To get a true set of results, it is necessary to handle this problem as well. In this research, we extensively improved the system's performance and gained valuable results based on balanced data; still, it has many gaps that can be covered in the future. The experiments performed in this research can provide valuable directions to job- seekers and reputed organizations to better understand fact-based insights about employment scam and their effects on society. Consequently, people would not fall into the trap of employment scams anymore. By distinguishing ORF, the people who were wasting their time and money on those fraudulent activities can be vigilant now. Conventional fraud detection without considering class imbalance problems can lead to misleading conclusions for both job- seekers and organizations.

To get a true set of results, it is necessary to handle this problem as well. In this research, we extensively improved the system's performance and gained valuable results based on balanced data; still, it has many gaps that can be covered in the future. All sets of analyses are performed on the job postings advertised in the English language only.

REFERENCES

[1]P. Kaur, ''E-recruitment: A conceptual study,'' Int. J. Appl. Res., vol. 1, no. 8, pp. 78–82, 2015.

[2]A.Raza, F.Younas, F.Akhtar,''Fake e-job posting prediction based on advance machine learning approachs,'' Int. J. Res. Publication Rev., vol. 3, no. 2, pp. 689–695, Feb. 2022.

[3]J. David, J. Cui, and F. Rahimi, ''Classification of imbalanced dataset using BERT embeddings,'' Dalhousie Univ., Halifax, Canada, Jan. 2020. Accessed: Jan. 2024.

[4]J. Howington, ''Survey: More millennials than seniors victims of job scams,'' Flexjobs,
CO, USA, Sep. 2015. Accessed: Jan. 2024

[5]G. Othman Alandjani, ''Online fake job advertisement recognition and classification using machine learning,'' 3C TIC, Cuadernos de Desarrollo Aplicados a las TIC, vol. 11, no. 1, pp. 251–267, Jun. 2022

[6]B. Alghamdi and F. Alharby, ''An intelligent model for online recruitment fraud detection,''
J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[7]Kanika, J. Singla, A. Kashif Bashir, Y. Nam, N. UI Hasan, and U. Tariq,''Handlingclassimbalanceinonlinetransactionfrauddetection,'' Comput., Mater. Continua, vol. 70, no. 2, pp. 2861–2877, 2022.

[8]S. Dutta and S. K. Bandyopadhyay, ''Fake job recruitment detection using machine learning approach,'' Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, Apr. 2020.