International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 05 | May 2026 www.irjet.net p-ISSN: 2395-0072

# ONLINE RECRUITMENT FRAUD (ORF) DETECTION USING DEEP LEARNING APPROACHES

**P. Hima Bindhu[1] Mr.S Muni Kumar [2]**

*[1]student, Mca [2]nd Year Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India*
*[2]Associate professor, Dept Of Mca, Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India*

---

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 Existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the existing models' capability in detecting fraudulent jobs.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

***Keywords: Recruitment, Online, Fraud, Deep Learning* Approaches**

## 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 no weens witches d 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. Most organizations prefer online recruitment systems to provide job opportunities to potential candidates .Organizations publish job ads for their vacant positions through job portals, in which they mention job descriptions, including requirements, salary packages, job advertisements were 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. 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.

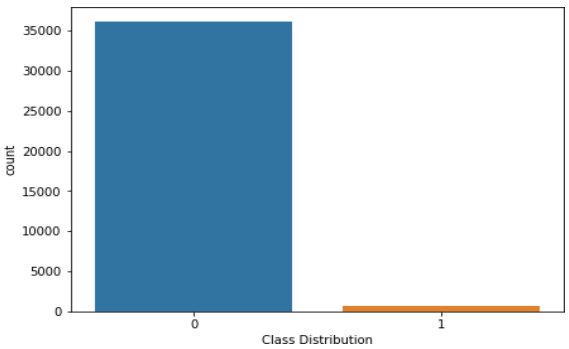


**Chart 1.** Amount of real vs. fake job posts

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. 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

## 2. RELATED WORK

This section reviews multiple studies related to Online Recruitment Fraud (ORF) detection. Moreover, as it is mentioned earlier that the collected dataset for this

research 1https: //www .kaggle.com/shivamb/real- r-fake-fake job posting prediction https:// www.kaggle.com/ datasets/ zusmani /pakistans-job-market https :// www.kaggle.com/ datasets/ prompt cloud/ indeed-job-posting dataset

## 3. ORFDETECTION TECHNIQUES

To detect fake job postings, Varietal. officially released the first dataset, ‘‘Employment Scam Aegean Dataset’’ (EMSCAD), and applied traditional machine learning class sifters on it to detect ORF. They performed two types of experiments and compared their results. The first experiment consists of six different classifiers, Naive Bayes (NB), Zero Rule (ZeroR), One Rule (Oner), Logistic Regression (LR), J48, and Random Forest (RF). The best classifier of this experiment is RF, with the highest precision of 91.4%. For the second experiment, the empirical ruleset model is used. LR, J48, and RF classifiers gave a precision of 90.6% for the empirical ruleset modelling. Dutta and Bandyopadhyay also applied machine learning algorithms to the ‘‘fake job postings’’ dataset. NB, Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), and Decision Tree (DT)are used as single classifier-based predictions. RF, Adaptive Boosting (Adobos), and Gradient Boosting (GB) classifiers are used as ensemble classifier-based predictions. DT achieved the highest accuracy of 97.2% among single classifier-based predictions, whereas, the RF classifier outperforms with an accuracy of 98.27% among ensemble classifier-based predictions. Another work to detect ORFs was published by Alghamdi and Alharbi. They applied Support Vector Machine (SVM) for the determination of relevant features present in the dataset. For the classification task, they used an ensemble-based RF classifier.

## 4. DATAAUGMENTATION TECHNIQUES

To balance class distribution in data, Gosling and Sir dana proposed four oversampling techniques; Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, ADASYN, and Safe Level SMOTE with various classification models, NB, KNN, and SVM. These over sampling techniques and models were implemented on six different datasets. The performance of different oversaw piling techniques on various datasets has been evaluated. SLSMOTE is considered to be the outperformer in this study. Akhba reds et al. in experimented with seven logbook datasets from the domain off acidity, aviation, and automotive. They used four methods to handle the class imbalance problem: under sampling, oversampling, feedback loop, and random down sampling loop., and Borderline-SMOTE.

## 6. CRITICAL ANALYSIS

Many machine learning approaches have been used for Online Recruitment Fraud (ORF) detection; however, advanced deep-learning approaches have yet to be explored in their full capacity to solve this problem. Employment scam been achieved but with poor recall. However, due to the class imbalance problem, accuracy does not represent the accurate picture of the story. It can be misleading that we get high predictive accuracy for the majority class and fail to seize the minority class, so we cannot rely only upon it as an evaluation metric. There is a need to improve balanced accuracy and recall to capture the situation truly. Furthermore, it is identified from Exploratory Data Analysis (EDA) that our collected dataset has a class imbalance problem, so the literature review helped us identify some top-performing SMOTE variants to be selected for experimenting in this regard. The methodology we will follow to handle the issues mentioned above is presented in the next section in detail.

## 5. 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 per for med 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 pre-processing 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 pre-processing 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 as shown in

## 7. DATA ACQUISITION

To address the under lying problem, we present an ovel data asset of fake job postings labelled as ‘‘fraudulent’’ for fake and ‘‘non-fraudulent’’ for legitimate job postings. The proposed data is a combination of job postings from three different sources mentioned as follows: ‘‘Fake Job Postings’’ dataset containing almost 17,880 real-life job postings advertised between 2012 and 2014 in different countries was collected. Eighteen features represented a particular job posting in this data. ‘‘US Job Postings’’ dataset containing almost 30,000 job advertisements published from July 2019 to August 2019 and belonging to different cities in the United States was collected. Thirty features represented a particular job posting in this data. ‘‘Pakistan Job Postings’’ dataset containing about 7000 job advertisements published during COVID-19 from

December 2019 to March 2021 and belonging to different cities in Pakistan was collected. Nine features represented a particular job posting in this data. We add publicly available job postings of Pakistan and the Using the ''Fake Job Postings'' dataset. The reason to extend the ''Fake Job Postings'' dataset is that its job postings are pretty outdated, and limited due to knowledge of specific job postings, which limits the capability of existing models in detecting fraudulent jobs. Therefore, we enhance this dataset with the latest job postings of Pakistan and the US to get a better realization of this problem. All textual columns of the aforementioned datasets are combined into a single column to get a prediction. The shape of the final data is now changed, shown in Fig. 1. It has only two columns. The first is ''job-content,'' representing the job description, whereas the second column, ''fraudulent,'' represents the class label. It can either be ''0'' for non-fraudulent or ''1'' for fraudulent. The rest of the columns do not take part in making predictions. They have been kept for analysis purposes only. In the next section, the prepare cessing steps performed to clean our data are discussed in detail.

## 8. DATA PREPROCESSING

Data pre-processing is a crucial step to transform raw data in a way suitable for any machine learning and deep learning task. In this phase, we only keep a useful portion of data and remove unnecessary data. We used neattext4 python library for pre-processing task. Various pre-processing steps are performed, which include the extraction of hashtags, HTML tags, URLs, email addresses, special characters, and duplicate and null values from the data because such words do not affect the orientation of the text. Lowercasing all available text is also necessary to preserve the consistent flow of the text. After getting cleaned data, it is split into training and testing sets with a ratio of 80:20. Exploratory Data Analysis Job Description/Function, Company Profile, Job Require meant, Department, and Employment Type. These features are crucial as they provide comprehensive insights into job postings, which are essential for accurately detecting fraudulent listings. Job requirements are further categorized into required education and required experience, providing granular details that enhance model precision. We have created comparison graphs as shown in Fig. 4 for the four major features that significantly impact the detection of fraudulent job postings, illustrating their importance in our analysis. To analyse illegitimate job postings, we explored our proposed data to understand it well and extract different patterns. From Fig. 4(a), it has been analysed that the job advertisements specified as ''part-time'' w.r.t employment type are more likely to be fraudulent, with a fraudulency rate of more than 90%. The job postings in which no employment type is mentioned have a fraudulency rate of about 70%. In contrast, those in which employment type is specified as ''temporary,'' ''contract,'' and ''full-time,'' are less likely to be fraudulent, having fraudulency rates of about 10%, 30%, and 50%, respectively. From Fig.

## 10. RESULTS

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 uproar performances. We have conducted an analysis of type errors to investigate the impact of employing different SMOTE oversampling techniques on predictive models. Our analysis involved a comparison with the performance of models that do not utilize any SMOTE variant. In this context, two distinct types of errors may arise: Type I and Type II. Type I errors occur when a genuine job posting (representing the majority class) is mistakenly classified as a fake job posting (representing the minority class). Type I errors are generally less consequential for job seekers. Conversely, Type II errors arise when a fake job posting (minority class) is erroneously classified as a real job posting (majority class). Type II errors present a greater challenge for us, as considering any fake job posting as real can lead to numerous significant problems as discussed in Section I. Hence, our primary focus was on mitigating TypeII errors Table 2 represents that the Type II error rate produced on BERT actual data was 68.77% as it predicts 152 incorrect samples against 221 fake job samples.

## 9. CONCLUSION

In this research, the problem of ORF detection is analysed 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. To rectify this class distribution imbalance, the top ten highly effective SMOTE variants were implemented on the imbalanced data. Subsequently, a type error analysis was conducted to investigate the impact of employing SMOTE variants on predictive models. Transformer-based classification models, BERT and Roberta, were implemented on both the imbalanced and balanced data, and the results were compared to derive more comprehensive insights from the experiments. Diverse evaluation metrics were employed to compare the performance of the implemented techniques. enhancing balanced accuracy and recall as evaluation metrics. All implemented approaches exhibited commendable performance.

## 11. REFERENCES

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

[2] C. S. Anita, P. Nagaraj an, G. A. Sai ram, P. Ganesh, and G. Deepak Kumar, ‘‘Fake job detection and analysis using machine learning and deep learning algorithms,’’ Revisit Gestapo Inovio e Technologies, vol. 11, no. 2, pp. 642–650, Jun. 2021.

[3] A. Raza, S. Ubaid, F. Yuan’s, andF.Akhtar,‘‘Fakee job posting prediction based on advance machine learning approaches,’’ Int. J. Res. Publication Rev., vol. 3, no. 2, pp. 689–695, Feb. 2022.

[4]Online Fraud. Accessed: Jun.19,2022. [Online]. Available: https://www.cyber.gov.au/acc/report

[5]Harington, ‘‘Survey: More millennials than seniors victims of job scams,’’ Flex jobs, CO, USA, Sep.2015. Accessed: Jan.2024[Online]. Available :www.flexjobs.com/blog/post/survey-results-millen nails-seniors-victims-job-scams

[6] Report Cyber. Accessed: Jun.25,2022.[Online]. Available : https://www.actionfraud.police.uk/

[7] S. Videos, C. Koloa’s, G. Kambouris, and L. Akol, ‘‘Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,’’ Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.

[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.

[9] B. Alghamdi and F. Alharbi, ‘‘An intelligent model for online recruitment fraud detection,’’ J. Inf. Secured., vol. 10, no. 3, pp.