International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

ONLINE RECRUITMENT FRAUD (ORF) DETECTION USING DEEP LEARNING APPROACHES

TALARI TEJESH, Mr. S. SUBAHAN
MCA Student, Assistant Professor
DEPT OF MCA
PVKK INSTITUTE OF TECHNOLOGY(AUTONOMOUS), Anantapuramu – 515001 (A.P)
tejutejes9@gmail.com, shaik.subahan06@gmail.com

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 (BERT) and Robustly Optimized BERT-Pretraining 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 Existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the existing models' capability in detecting fraudulent jobs.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 analyzed and compared. All implemented approaches are performed competitively. However, BERT+SMOBD SMOTE achieved the highest balanced accuracy and recall of about 90%.

1. INTRODUCTION

1.1. ABOUT THE PROJECT

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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.

OBJECTIVE

The objective of this research is to investigate Online Recruitment Fraud (ORF) and to overcome the possible issues in implementing the system. The significant contributions of this study are mentioned as follows:

- Job postings from three different sources are collected and combined to present a novel dataset.
- It is observed from Exploratory Data Analysis (EDA) that the class distribution from the collected dataset is highly imbalanced. Ten top-performing SMOTE variants are implemented to balance the class distribution
- ratio.
- Transformer-based deep learning models are implemented on the dataset to detect whether a job posting is fraudulent or non-fraudulent.
- Comparative analysis of implemented models is conducted on both imbalanced and balanced datasets.

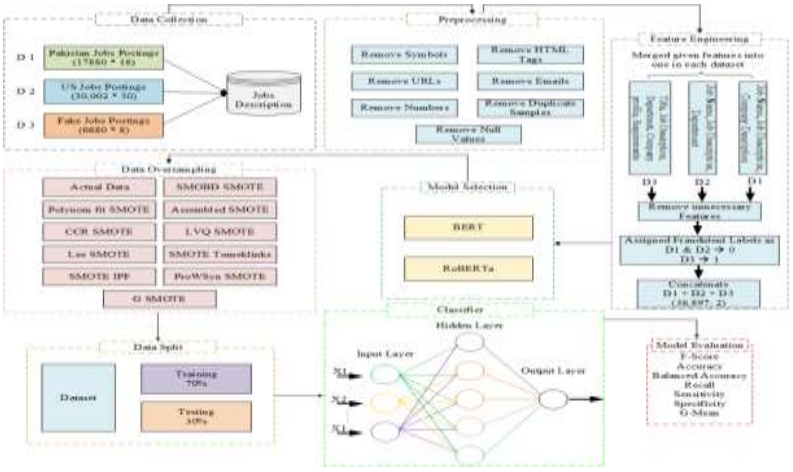


Fig.1.1: System Architecture

1.2 MODULES

1.3.1. Data Collection Module:

Combines job postings from three diverse sources to create a novel dataset.

1.3.1. Preprocessing Module:

Cleans, processes, and prepares data for analysis, addressing inconsistencies.

1.3.2. MUSIC RECOMMENDATION:

Exploratory Data Analysis (EDA) Module: Identifies patterns, trends, and class imbalance in the dataset.

1.3.4. Performance Analysis Module:

Compares model performance across SMOTE variants, focusing on accuracy, recall, and balanced metrics.

2. LITERATURE SURVEY

A literature survey or literature review is the study of references and old algorithms that we have read for designing the proposed methods. It also helps in reporting summarization of all the old references papers, and their drawbacks. The detailed literature survey for the project helps in comparing and contrasting various methods, algorithms in various ways that have implemented in the

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

research.

Online recruitment fraud (ORF) has emerged as a significant concern, leading to extensive research into detection methods utilizing deep learning techniques. Here are five notable studies in this area:

"Fraud-BERT: Transformer-Based Context-Aware Online Recruitment Fraud Detection"

This study introduces Fraud-BERT, a model leveraging transformer architectures to detect fraudulent job postings. By capturing contextual information within job descriptions, Fraud-BERT enhances detection accuracy. The authors have made their code publicly available at https://github.com/GJU-CSE/Online-Recruitment-Fraud-Detection.

"A Machine Learning Approach to Detecting Fraudulent Job Types"

This research develops a machine learning system to identify various categories of fraudulent job advertisements, such as identity theft and pyramid schemes. The study evaluates lexical, syntactic, semantic, and contextual features, finding that word embeddings and transformer-based features outperform traditional handcrafted features.

"Fraudulent Jobs Prediction Using Natural Language Processing and Deep Learning Sequential Models"

This project proposes a methodology combining natural language processing (NLP) techniques with deep learning sequential models to address fraudulent job postings. Through experimentation with text preprocessing, word vectorization, and oversampling techniques, the study achieves high accuracy, precision, recall, and F1 scores, demonstrating the effectiveness of deep learning in this context.

"ORF Detector: Ensemble Learning Based Online Recruitment Fraud Detection"

ORF Detector employs ensemble learning techniques to detect online recruitment fraud. Tested on a dataset of 17,860 annotated job postings, the model achieves an average F1-score of 94% and an accuracy of 95.4%, demonstrating improved specificity compared to baseline classifiers.

"Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches"

This study explores the application of deep learning models, specifically BERT and RoBERTa, for ORF detection. By fine-tuning these pre-trained models on a dataset of job postings, the research demonstrates significant improvements in identifying fraudulent listings.

3. ANALYSIS

3.1. EXISTING SYSTEM

To detect fake job postings,Vidros et al. [7] officially released the first dataset, ‘‘Employment Scam Aegean Dataset’’ (EMSCAD), and applied traditional machine learning classifiers 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 modeling.

3.1.1 Limitations in Existing System

- Limited generalization due to outdated datasets.
- Poor handling of class imbalance, leading to biased model predictions.
- Less accurate and precise detection of fraudulent job postings.
- Inadequate capability to adapt to evolving scam patterns.

3.2. PROPOSED SYSTEM

We presented a novel dataset of fake job postings labeled 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

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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.

3.2.1. Features of the Proposed System

- Comparative analysis of implemented models is conducted on both imbalanced and balanced datasets
- Higher accuracy and precision in detecting fraudulent job postings (90% balanced accuracy and recall using BERT+SMOBD SMOTE).
- andles class imbalance effectively, improving predictions for minority classes.
- Adapts to the latest trends in fraudulent job postings using a comprehensive dataset.
- Demonstrates competitive performance across various SMOTE variants and deep learning models.

ALGORITHMS

Step 1: Data Collection

Gather job posting data from three diverse sources.
Combine the datasets to create a comprehensive and updated dataset of job postings.

Step 2: Data Preprocessing

Remove noise such as special characters, unnecessary spaces, and stopwords.
Normalize text by converting to lowercase and stemming/lemmatizing.
Tokenize text data into meaningful units for analysis.

Step 3: Exploratory Data Analysis (EDA)

Analyze the dataset for trends, patterns, and class distribution.
Identify the class imbalance issue where fraudulent job postings form the minority class.

Step 4: Model Training

Train the BERT+SMOBD model on the balanced dataset.
Use optimized hyperparameters for learning rate, batch size, and number of epochs.

Step 5: Model Evaluation

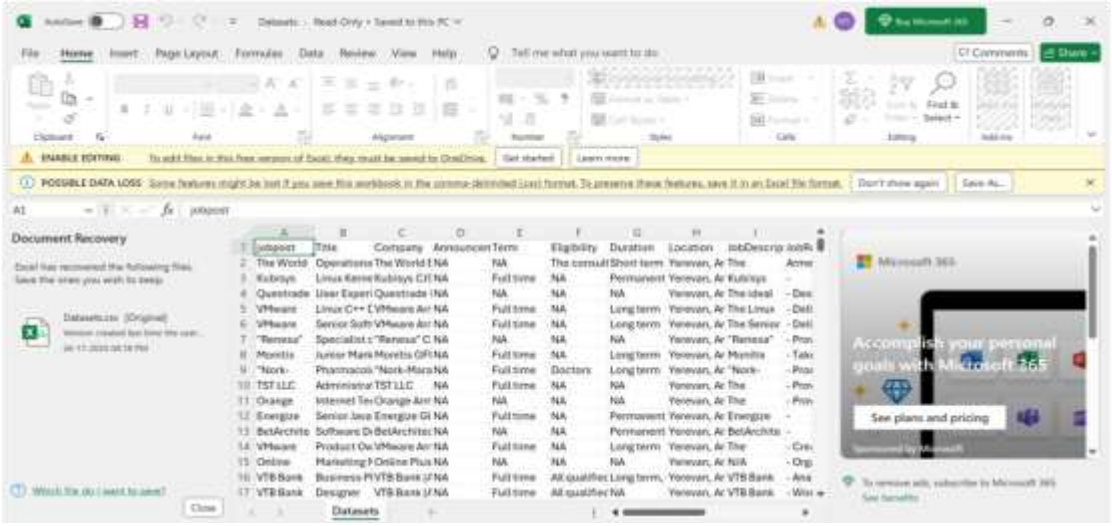
Test the trained model on the validation and testing datasets.
Compare the performance of BERT+SMOBD with other SMOTE variants and models.
Evaluate metrics like accuracy, precision, recall, F1-score, and balanced accuracy.

Step 6: Deployment and Monitoring

Deploy the final model for real-world use in detecting fake job postings.
Monitor the model's performance periodically and retrain with updated data as necessary.

9. SAMPLE SCREENS

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



Screen : Datasets

Screen :User Registration Page

Screen : User Menu Page

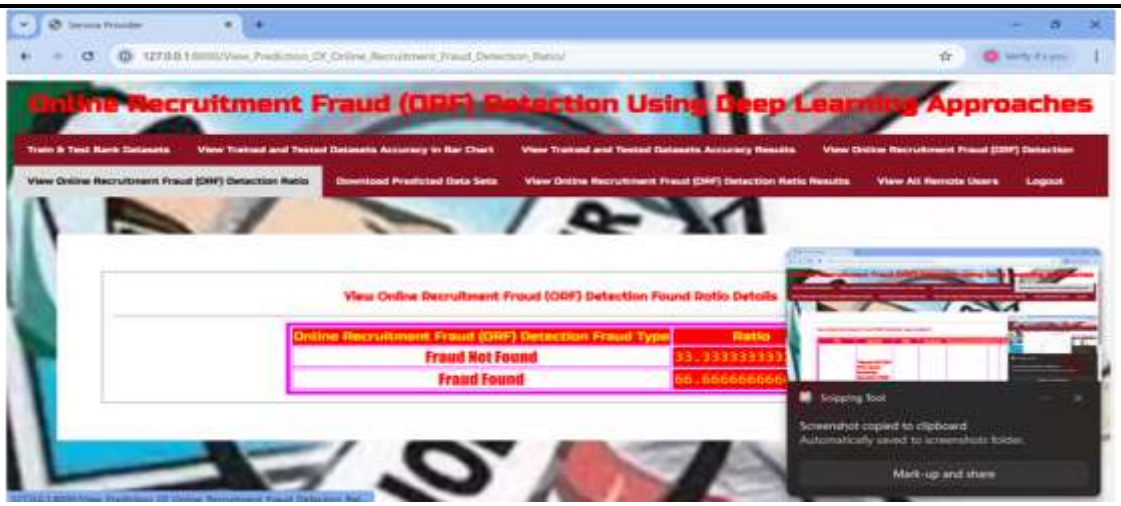
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Screen : Accuracy in Bar Charts

Screen : Accuracy in Line Charts

Screen : Predictied Datsets

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



Screen : Ratio Analysis

10. CONCLUSION AND FUTURE SCOPE

This project 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. Due to the class imbalance issue, only accuracy as an evaluation metric failed to provide an accurate representation of the overall performance. Because high predictive accuracy for the majority class can be misleading, as it may overshadow the minority class, leading to incomplete assessment. Thus, this study prioritized enhancing balanced accuracy and recall as evaluation metrics. All implemented approaches exhibited commendable performance. However, based on the type error and classification results, it was observed that BERT, in conjunction with the SMOBD SMOTE technique, demonstrated exceptional performance on our data and achieved optimal outcomes.

10.2. FUTURE SCOPE

Development of more advanced deep learning models like transformers (e.g., BERT, GPT) to detect fraudulent job postings with higher accuracy. Integration of multimodal learning, combining text, images, and metadata for improved fraud detection. Implementation of real-time fraud detection to prevent scam job postings before they reach job seekers. Use of automated AI-driven moderation systems for job portals to verify job listings efficiently.

11. REFERENCES

1. Raut, Nitisha, “Facial Emotion Recognition Using Machine Learning” (2018). Master’s Projects. 632. https://doi.org/10.31979/etd.w5fs-s8wd
2. Hemanth P,Adarsh ,Aswani C.B, Ajith P, Veena A Kumar, EMO PLAYER: Emotion Based Music Player, International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 4, April 2018, pp. 4822-87.
3. Music Recommendation System: Sound Tree, Dcengo Unchained: Sla KAYA, BSc.; Duygu KABAKCI, BSc.; Insu KATIRCIOLU, BSc. and Koray KOCAKAYA BSc. Assistant : Dilek Ã–nal Supervisors: Prof. Dr. smail Hakk Toroslu, Prof. Dr. Veysi ler Sponsor Company: ARGEDOR
4. Tim Spittle, lucyd, GitHub, , April 16, 2020. Accessed on: [Online], Available at: https://github.com/timspit/lucyd

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

5. Abdul, J. Chen, H.-Y. Liao, and S.-H. Chang, An Emotion-Aware Personalized Music Recommendation System Using a Convolutional Neural Networks Approach, Applied Sciences, vol. 8, no. 7, p. 1103, Jul. 2018.
6. Manas Sambare, FER2013 Dataset, Kaggle, July 19, 2020. Accessed on: September 9, 2020. [Online], Avalilable at: https://www.kaggle.com/msambare/fer2013
7. MahmoudiMA, MMA Facial Expression Dataset, Kaggle, June 6, 2020. Accessed on: September 15, 2020. [Online], Available at: https://www.kaggle.com/mahmoudima/mma-facial-expression
8. Dr. Shaik Asif Hussain and Ahlam Salim Abdallah Al Balushi, A real time face emotion classification and recognition using deep learning model, 2020 Journal. of Phys.: Conf. Ser. 1432 012087
9. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops, San Francisco, CA, USA, 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.
10. Puri, Raghav & Gupta, Archit & Sikri, Manas & Tiwari, Mohit & Pathak, Nitish & Goel, Shivendra. (2020). Emotion Detection using Image Processing in Python.