IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Detecting Online Recruitment Fraud through Deep Learning using Hybrid Neural Networks

[1]Ms.S.Sai Jaswanthi, [2]Dr.D.Bright Anand, [3] Dr.R.Karunia KrishnaPriya, [4]Mr. Pandreti Praveen., [5]Dr.T.Senthil

[1]PG Scholar,Dept of Computer Science and Engineering, Sreenivasa Institute of Technology and Management Studies, Chittoor, India.(517127)

[2,3,5]Associate Professor, Dept of Computer Science and Engineering, Sreenivasa Institute of Technology and Management Studies, Chittoor, India.

[4]Assistant Professor, Dept of Computer Science and Engineering ,Sreenivasa Institute of Technology and Management Studies, Chittoor, India.

Abstract: Fraudulent job ads have grown to be a major problem in online recruiting due to the quick expansion of online employment platforms. We suggest an improved job categorisation system that uses cutting-edge deep learning algorithms to identify fake job ads in order to address this issue. We use a Convolutional Neural Network (CNN2D) to enhance feature extraction and maximise classification performance, building on the usage of transformer-based models such as BERT and RoBERTa. By employing 2D convolutional layers, the CNN2D model efficiently and accurately detects fraudulent job ads by capturing intricate patterns in the dataset. Furthermore, the system's user-friendly interface, which makes use of the Flask framework, allows for easy job posting administration and real-time fraud detection. This enhancement increases the accuracy and expediency of the process of identifying fraudulent job postings.

Index terms - Fraudulent Job Postings, Convolutional Neural Network (Cnn2d), Feature Extraction, Real-Time Fraud Detection.

1. INTRODUCTION

Adverse Drug Reactions (ADRs) are unpleasant or dangerous side effects that arise from taking pharmaceuticals; they frequently call for medical attention, dose changes, or stopping the drug entirely. Because they lead to higher mortality,

longer hospital admissions, and a sharp increase in healthcare expenses, these responses represent a major danger to public health systems across the world. Early detection and prediction are crucial since many adverse drug reactions (ADRs) are not discovered during clinical trials and only become apparent after the medication has been released into the general market.

ADR incidence is impacted by a number of factors, including as healthcare quality, geographic region, and sex. For example, research indicates that pharmacokinetic and pharmacodynamic variations, as well as larger medication dosages per body weight, make women more vulnerable to adverse drug reactions (ADRs). ADR reporting and management may also be impacted by differences in medical norms and access to healthcare among nations. According to research, a sizable fraction of ADRs—roughly 71.6% in rich countries and 59.6% in developing ones—can be avoided. ADR-related mortality rates are constant across locations, underscoring the pressing need for efficient forecast systems.

In order to overcome these obstacles, this study investigates a deep learning-based method for predicting drug-drug interactions and the negative consequences that go along with them utilising Graph Neural Networks (GNNs) and Self-Supervised Learning. The model seeks to increase the precision of ADR identification prior to pharmaceuticals reaching patients by utilising structured representations of medications and their interactions, thereby improving drug safety and saving lives.

2. LITERATURE SURVEY

i) Online fake job advertisement recognition and classification using machine learning
https://dialnet.unirioja.es/servlet/articulo?codigo=8415586

In real-world smart systems, machine learning algorithms manage a wide variety of data types. Many recruiters and job seekers are now actively working online due to technological advancements and the widespread usage of social media platforms. On the other side, invasions of privacy and data may expose one to risky behaviours. Fraudsters and businesses use virtual work-supplying websites, among other things, to lure job seekers. Our goal is to decrease the number of fraudulent and phoney attempts by applying machine learning to provide projections. To improve detection, our suggested method makes use of many classification models. In order to improve the results, this study also compares the performance of many classifiers using various approaches for real results.

ii) Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms
https://revistageintec.net/old/wp-content/uploads/2022/02/1701.pdf

Online job postings at employment sites have increased as a result of the pandemic. On the other hand, some internet jobs are scams that steal sensitive and important data. These phoney jobs might be accurately recognised and

categorised from a pool of phoney and legitimate job posts using new deep learning and machine learning classification techniques. This research uses machine learning and deep learning techniques to detect bogus jobs. In order to make the classification system accurate and exact, this study recommends data cleaning and analysis. Data purification is essential to machine learning efforts because it influences the accuracy of machine learning and deep learning algorithms. Thus, the focus of this paper is on pre-processing and data purification. The classification and identification of fake jobs are precise and accurate. Therefore, applying machine learning and deep learning algorithms to cleaned and pre-processed data is essential for increased accuracy. To improve accuracy, deep learning neural networks are employed. Finally, the most precise and accurate classification method is found by comparing all of these models.

iii) Online Recruitment Fraud Detection using ANN
https://ieeexplore.ieee.org/abstract/document/9636978

Online job searchers may easily locate and apply for positions. Additionally, it helps recruiters find quality applicants, which improves the recruiting process. The prevalence of employment fraud is rising. Job advertisements might be real or fake. This paper presents an artificial neural network-based method for identifying job post fraud. The publicly available Employment Scam Aegean Dataset (EMSCAD) may be used for text preprocessing in order to train and evaluate the suggested model. Our model's accuracy, recall, and f-measure are 91.84%, 96.02%, and 93.88%, respectively. The findings demonstrate that the ANN-based model outperforms rival algorithms in identifying fraudulent employment.

iv) Classification of Genuinity in Job Posting Using Machine Learning
https://www.academia.edu/68038151/Classification_of_Genuinity_in_Job_Posting_Using_Machine_Learning

By using machine learning (ML) to forecast the likelihood that a job would be phoney, we assist candidates stay vigilant and make wise decisions, which in turn reduces the quantity of phoney job advertising on the internet. The model uses NLP to analyse job posting attitudes and trends, and the TF-IDF vectorizer is used to extract features. The balanced data will be precisely categorised using Random Forest and SMOTE. Even with large datasets, it performs well, improving model accuracy and preventing overfitting. Using information from job listings, the final algorithm will determine if the position is genuine or fraudulent.

v) A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques
https://ieeexplore.ieee.org/abstract/document/9331230

Posting new jobs has become commonplace due to social media and technology. Everyone will thus be worried about inaccurate job posting predictions. False job posing prediction has a number of issues, much like previous

categorisation initiatives. This study suggests using a number of data mining techniques and algorithms, including support vector machines, decision trees, and KNN, to identify fake job ads. We looked at 18,000 examples from the EMSCAD Employment Scam Aegean Dataset. This is where deep neural network classifiers excel. Three thick layers make up this deep neural network classifier. The trained classifier predicts bogus job posts (DNN) with 98% accuracy.

3. METHODOLOGY

i) Proposed Work:

The proposed system enhances traditional job classification methods by integrating a CNN2D-based model for advanced feature extraction. This deep learning approach allows for better detection of fraudulent job postings by identifying complex data patterns. The system is supported by a Flask-based web interface that simplifies user interaction, enabling real-time predictions and efficient upload of job posting files. This combination ensures high accuracy in classification and offers a user-friendly platform for administrators to monitor and manage job authenticity effectively.

ii) System Architecture:

The suggested approach uses Convolutional Neural Networks (CNN2D) to identify fake job postings on recruitment websites. While transformer-based models like BERT and RoBERTa have shown promise in detecting fake job postings, they often face challenges in effectively capturing complex patterns in the data. To address this, the system combines the strengths of these transformer models with the advanced capabilities of CNN2D, which is known for its ability to detect intricate patterns in datasets. The CNN2D model applies 2D convolutional layers to extract more nuanced features from job postings, thus improving the model's overall accuracy and efficiency in distinguishing between real and fake listings.

The first step of the system involves gathering a comprehensive dataset, which includes both genuine and fraudulent job postings from various sources. This expanded dataset helps overcome the limitations of existing, outdated benchmark datasets and ensures the system remains relevant with the latest job posting trends. After gathering the data, the CNN2D model is employed to process and extract critical features, enabling the system to better capture complex patterns that may be indicative of fraudulent behavior. In combination with BERT and RoBERTa, this hybrid approach leverages both the language understanding capabilities of the transformer models and the feature extraction prowess of CNN2D.

To further enhance the model’s performance, the system tackles the common issue of class imbalance in the dataset by incorporating the Synthetic Minority Oversampling Technique (SMOTE). To train the system on a more balanced sample, many implementations of SMOTE generate phoney, fraudulent job ads. This improves the system's ability to detect

fraudulent listings, especially in cases where fraudulent jobs are fewer in number compared to legitimate postings.

The system also includes a user-friendly interface developed using the Flask framework, providing administrators with an intuitive way to manage job postings. Through this interface, administrators can upload job listings for real-time fraud detection. The system processes these listings instantly, classifying them as either legitimate or fraudulent, and providing actionable insights. Ensuring the security of online recruitment platforms, deep learning algorithms combined with an intuitive interface can detect fraudulent job ads with great efficiency and accuracy.

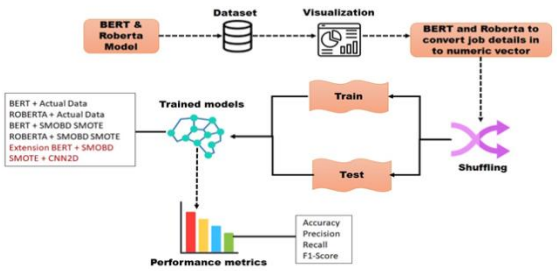

Fig.1. Proposed Architecture

iii) MODULES:

a. **Load BERT & RoBERTa Model**: This module involves importing pre-trained BERT and RoBERTa models from the Hugging Face library. These models serve as powerful text encoders for processing job descriptions and extracting contextual embeddings.

b. **Exploring the Dataset**: In this module, the dataset is examined for structure, content, and potential issues. Key statistics such as class distribution, missing values, and sample entries are analyzed to understand the data.

c. **Visualization**: This module utilizes visualization libraries to create graphs and charts, showcasing insights into the dataset. Visualizations help identify trends, relationships, and anomalies in job postings, enhancing understanding.

d. **BERT and RoBERTa for Vectorization**: Here, BERT and RoBERTa convert job details into numerical vectors. The models tokenize and encode the text, transforming it into dense representations suitable for classification tasks in the subsequent steps.

e. **Shuffling**: This module randomizes the order of data entries to ensure that the training process is unbiased and independent of the input order. Shuffling enhances model robustness and generalization.

f. **Split the data into train & test**: In this module, the dataset is divided into training and testing subsets, typically using an 80-20 split. This separation allows for effective training of the model and evaluation of its performance.

g. **Model generation:** Model building – BERT + Actual Data, ROBERTA + Actual Data, BERT + SMOBD SMOTE, ROBERTA + SMOBD SMOTE, Extension BERT + SMOBD SMOTE + CNN2D.

Performance evaluation metrics for each algorithm is calculated.

h. **Admin login**: In this module, admin can login into the application.

i. **Predict Fraud Job**: In this module user can upload the input data.

j. **Logout**: User can logout after the completion of all activities.

iv) ALGORITHMS:

i. **BERT + Actual Data:** This algorithm applies BERT to convert raw job postings into numerical vectors, leveraging its deep bidirectional context understanding for accurate language representation. By processing actual data without oversampling, BERT helps in capturing genuine patterns and features in job descriptions, aiding in identifying indicators of fraud effectively.

ii. **RoBERTa + Actual Data**: RoBERTa enhances BERT by utilizing optimized pre-training techniques and extensive training data to encode job details. It provides deeper contextual representations and improves detection accuracy on actual data. RoBERTa's enhanced language modeling helps distinguish subtle linguistic cues that may indicate fraudulent or deceptive job postings.

iii. **BERT + SMOBD SMOTE:** With BERT embeddings enhanced by SMOBD SMOTE, this approach addresses class imbalance in the dataset by oversampling minority class samples. This combination improves detection accuracy for underrepresented fraudulent postings, ensuring the model effectively captures the distinguishing features of rare fraudulent postings.

iv. **RoBERTa + SMOBD SMOTE:** By combining RoBERTa with SMOBD SMOTE, this approach leverages RoBERTa's robust feature extraction and SMOTE's balancing capabilities. This pairing improves model resilience against imbalance, enhancing the system's ability to detect nuanced indicators of fraudulent postings within an imbalanced dataset.

v. **Extension BERT + SMOBD SMOTE + CNN2D:** This extended model combines BERT embeddings with SMOBD SMOTE balancing and CNN2D for classification, enhancing feature extraction through spatial patterns. The CNN2D layer adds a deeper level of pattern recognition, improving the system's classification performance and yielding a higher accuracy for fraud detection.

4. EXPERIMENTAL RESULTS

Using the Basic Neural Network approach with features from BERT and ROBERTA, the model was able to classify job postings into real or fake, with a reasonable accuracy. However, when CNN2D was applied, the accuracy significantly improved. The CNN2D layers

helped the model capture more complex patterns in the data, resulting in better feature optimization and classification performance.

When BERT and ROBERTA were combined with the SMOTE SMOBD algorithm (for balancing the dataset), they performed exceptionally well in detecting fake job postings, with BERT showing higher accuracy compared to ROBERTA. The CNN2D extension further boosted accuracy, as it could better extract and process features, making it more effective than a basic neural network.

All datasets can be downloaded from below URL

https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction
https://www.kaggle.com/datasets/promptcloud/indeed-job-posting-dataset

https://www.kaggle.com/datasets/zusmani/pakistans-job-market

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$\text{Accuracy} = \frac{(\text{TN} + \text{TP})}{\text{T}}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Pr e cision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

$$\text{Recall} = \frac{\text{TP}}{(\text{FN} + \text{TP})}$$

mAP: One ranking quality statistic is Mean Average Precision (MAP). It takes into account the quantity of pertinent suggestions and where they are on the list. The arithmetic mean of the Average Precision (AP) at K for each user or query is used to compute MAP at K.

$$mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k$$

$$AP_k = \textit{the AP of class } k$$

$$n = \textit{the number of classes}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$\text{F1} = 2 \cdot \frac{(\text{Recall} \cdot \text{Pr e cision})}{(\text{Recall} + \text{Pr e cision})}$$

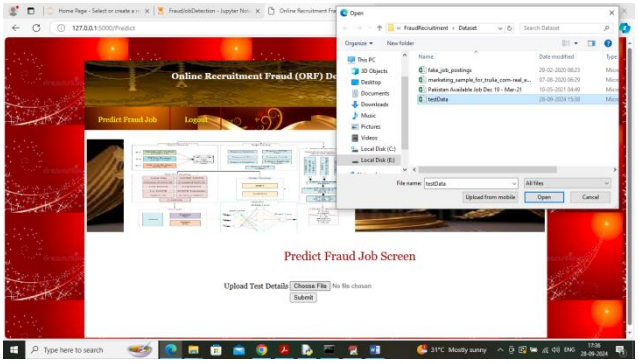

Fig.2. upload test file

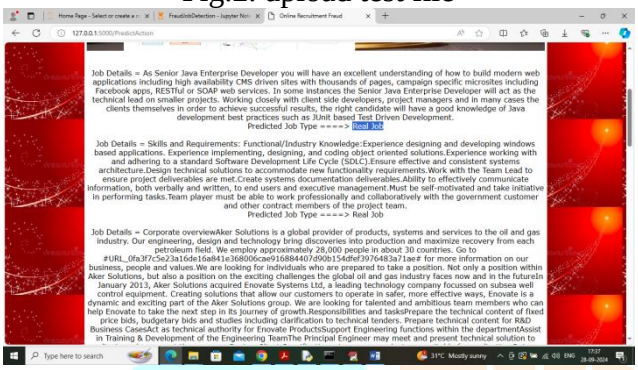

Fig.3. dataset analysis

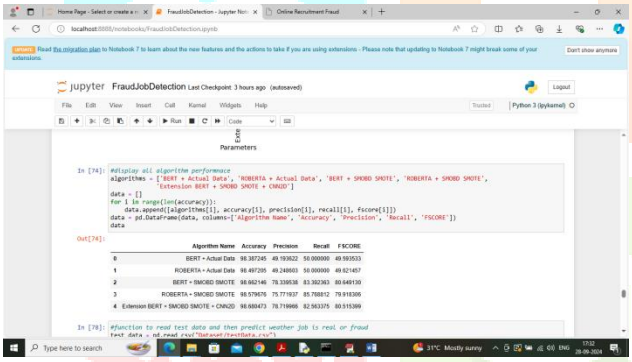

Fig.4. algorithm performance in tabular format

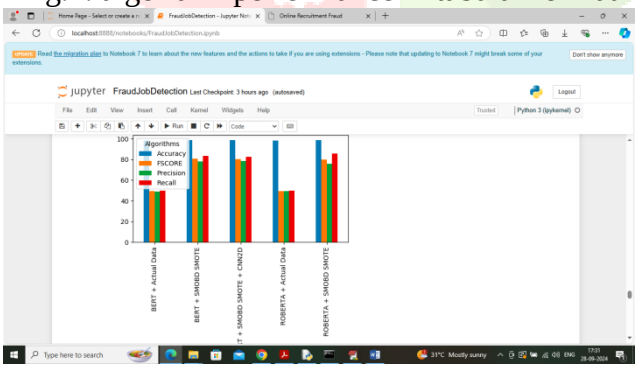

Fig.5. algorithm performance graph

5. CONCLUSION

The extension concept of using CNN2D with BERT and ROBERTA for classifying job postings has proven to significantly enhance the model's performance. While the basic neural network approach provided a solid foundation, the introduction of CNN2D allowed for better feature extraction and optimization, leading to improved accuracy. By combining BERT and ROBERTA with the SMOTE SMOBD algorithm, the model was able to effectively address data imbalance and detect fake job postings with higher precision. Overall, this extension concept provides a robust solution for accurately identifying fraudulent job postings, outperforming traditional approaches in terms of both accuracy and feature extraction capabilities.

6. FUTURE SCOPE

The research intends to identify online recruiting fraud using ensemble techniques and sophisticated feature extraction algorithms. Combining attention systems with recurrent neural networks (RNNs) might help the model to better collect contextual data in job ads. On sma, transfer learning from pre-trained models can maximise performance.

REFERENCES

[1] 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.

[2] A.Adhikari, A.Ram, R.Tang, and J.Lin, ‘‘DocBERT: BERT for document classification,’’ 2019, arXiv:1904.08398.

[3] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ‘‘Online recruitment fraud detection using ANN,’’ in Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT), Sep. 2021, pp. 13–17.

[4] C.Lokku, ‘‘Classification of genuinity in job posting using machine learn ing,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575, Dec. 2021.

[5] S. U. Habiba, Md. K. Islam, and F. Tasnim, ‘‘A comparative study on fake job post prediction using different data mining techniques,’’ in Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST), Dhaka, Bangladesh, Jan. 2021, pp. 543–546.

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

[7] 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, 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. Alharby, ‘‘An intelligent model for online recruitment fraud detection,’’ J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, ‘‘ORFDetector: Ensemble learning based online recruitment fraud detec tion,’’ in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India, Aug. 2019, pp. 1–5.

[11] G. Kovács, ‘‘An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets,’’ Appl. Soft Comput., vol. 83, Oct. 2019, Art. no. 105662.

[12] S. Gazzah and N. E. B. Amara, ‘‘New oversampling approaches based on polynomial fitting for imbalanced data sets,’’ in Proc. 8th IAPR Int. Workshop Document Anal. Syst., Nara, Japan, Sep. 2008, pp. 677–684.

[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ‘‘Detecting scam in online job vacancy using behavioral features extraction,’’ in Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.

[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ‘‘Handling imbalanced datasets: A review,’’ GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006.

[15] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, ‘‘Toward credible evaluation of anomaly-based intrusion-detection methods,’’ IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.

[16] Y.-H. Liu and Y.-T. Chen, ‘‘Total margin based adaptive fuzzy support vector machines for multiview face recognition,’’ in Proc. IEEE Int. Conf. Syst., Man Cybern., Waikoloa, HI, USA, Oct. 2005, pp. 1704–1711.

[17] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, ‘‘Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,’’ Neural Netw., vol. 21, nos. 2–3, pp. 427–436, Mar. 2008. 109406

[18] Y.Li,G.Sun,andY.Zhu,‘‘Dataimbalanceproblemintextclassification,’’ in Proc. 3rdInt. Symp.Inf. Process., Luxor, Egypt, Oct.2010,pp. 301–305.

[19] N.V.Chawla, K.W.Bowyer, L.O.Hall, and W.P.Kegelmeyer, ‘‘SMOTE: Synthetic minority over-sampling technique,’’ J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[20] S. Barua, M. M. Islam, and K. Murase, ‘‘ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning,’’ in Proc. Pacific–Asia Conf. Knowl. Disc. Data Min. II, Gold Coast, QLD, Australia, Apr. 2013, pp. 317–328.

[21] J. Stefanowski and S. Wilk, ‘‘Selective pre-processing of imbalanced data for improving classification performance,’’ in Proc. 10th Int. Conf. Data Warehousing Knowl. Disc. (DaWaK), Turin, Italy, Sep. 2008, pp. 283–292.

[22] A. Gosain and S. Sardana, ‘‘Handling class imbalance problem using oversampling techniques: A review,’’ in Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI), Delhi, India, Sep. 2017, pp. 79–85.

[23] F. Akhbardeh, C. O. Alm, M. Zampieri, and T. Desell, ‘‘Handling extreme class imbalance in technical logbook datasets,’’ in Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 4034–4045.

[24] J. Ah-Pine and E.-P. Soriano-Morales, ‘‘A study of synthetic oversampling for Twitter imbalanced sentiment analysis,’’ in Proc. Workshop Interact. Between Data Min. Nat. Lang. Process. (DMNLP), Riva del Garda, Italy, Sep. 2016, pp. 17–24.

[25] J. David, J. Cui, and F. Rahimi, ‘‘Classification of imbalanced dataset using BERT embeddings,’’ Dalhousie Univ., Halifax, Canada, Jan. 2020. Accessed: Jan. 2024. [Online]. Available: https://fatemerhmi.github.io/files/Classification_of_imbalanced_dataset_using_BERT_embedding.pdf

[26] Kanika, J. Singla, A. Kashif Bashir, Y. Nam, N. UI Hasan, and U. Tariq, ‘‘Handling class imbalance in online transaction fraud detection,’’ Comput., Mater. Continua, vol. 70, no. 2, pp. 2861–2877, 2022. [27] S. Bansal. (2020). [Real or Fake] Fake Jobposting Prediction. [Online]. Available: https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

[28] Indeed Job Posting Dataset. Accessed: Feb. 16, 2023. [Online]. Avail able: https://www.kaggle.com/datasets/promptcloud/indeed-job-posting-dataset

[29] Pakistan’s Job Market. Accessed: Feb. 16, 2023. [Online]. Available: https://www.kaggle.com/datasets/zusmani/pakistans-job-market