Using Hybrid Neural Networks and Deep Learning to Identify Online Recruitment Fraud

S. NOORTAJ[1], B DIVYA[2]
#1 GUIDE, Department Of MCA, KMM Colleges, Ramireddipalle, Tirupati Rural, Andhra Pradesh 517102

#2 PG Scholar, Department Of MCA, KMM Colleges, Ramireddipalle, Tirupati Rural, Andhra Pradesh 517102

Gmail : shaiknoortaj15@gmail.com [1], divyabandi04@gmail.com [2]

Abstract: These days, the majority of businesses use digital platforms to seek new hires in order to streamline the recruiting process. Fraudulent advertising is a result of the sharp rise in the usage of online job posting platforms. The fraudsters use phoney job postings to get revenue. Fraud in online hiring has become a significant problem in cybercrime. Therefore, to eliminate online job frauds, it is essential to identify phoney job ads. The goal of this research is to employ two transformer-based deep learning models, namely Bidirectional Encoder Representations from Transformers and Robustly Optimised BERT-Pretraining Approach (RoBERTa), to accurately detect fake job postings. Traditional machine learning and deep learning algorithms have been used in recent studies to detect fake job postings. By combining job ads from three distinct sources, a unique dataset of fraudulent job advertisements is suggested in this study. The effectiveness of current algorithms to identify fake jobs is hampered by the outdated and restricted benchmark datasets, which are based on knowledge of particular job advertisements. We thus update it with the most recent job openings. The class imbalance issue in identifying phoney employment is brought to light by exploratory data analysis (EDA), which causes the model to behave aggressively against the minority class. The work at hand employs 10 of the best Synthetic Minority Oversampling Technique (SMOTE) variations in order to address this issue.

Analysis and comparison are done between the models' performances balanced by each SMOTE version. Every strategy that is used is carried out in a competitive manner. At almost 90%, BERT+SMOBD SMOTE, on the other hand, had the best balanced accuracy and recall.

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

1. INTRODUCTION

The internet has fundamentally changed our lives in a variety of ways in this era of sophisticated technology. Nowadays, doing any task the old-fashioned way has been replaced by the internet. As a result, hiring and job searching have also moved online. Productivity, ease of use, and effectiveness are the advantages of an online recruiting system, sometimes known as e-recruitment [1]. To offer job openings to prospective employees, the majority of companies choose online recruiting platforms [2]. Through employment portals, companies post job

openings and include job descriptions, including prerequisites, compensation packages, offers, and facilities to be given. Job searchers go to several online job boards, look for openings that fit their interests, and apply for positions that fit. After that, the business reviews resumes to make sure they meet its needs. After completing additional procedures, such as interviewing and choosing possible applicants, the post is closed. During the worldwide COVID-19 epidemic, the tendency of posting job ads online was exaggerated. The World Economic Outlook Report states that the International Monetary Fund (IMF) calculated that during the height of the COVID-19 epidemic in 2020, the jobless rate rose to 13%. In 2018 and 2019, these figures were just 3.9% and 7.3%, respectively. Many businesses made the decision to advertise job positions online during the epidemic in order to accommodate job seekers [3]. However, when a resource is made available to the general population, it also gives internet scammers the opportunity to exploit their gloom.

In this study, we introduced a new dataset of fictitious job advertisements that were classified as "fraudulent" for fictitious job advertisements and "non-fraudulent" for genuine ones. Three distinct sources of job ads are combined to provide the suggested data. To expand the dataset with the most recent job posts, we use "Fake Job Postings" as the core dataset and include publicly accessible job ads from Pakistan and the US. We took this action as the benchmark datasets now in use are out-of-date and constrained by the knowledge of particular job ads, which reduces the effectiveness of current algorithms to identify fake employment. The dataset was prepared, and then it was subjected to exploratory data analysis, or EDA. The dataset's unbalanced class distribution was discovered by EDA. The ratio of samples in the minority class to those in the majority class is known as the imbalance class distribution [14]. For regular classes, it may result in high prediction accuracy; for rare classes, it might result in low predictive accuracy. Anomaly detection [15], facial recognition [16], medical diagnosis [17], text classification [18], and many other real-world areas are affected by the class imbalance problem. SMOTE [19], an oversampling method, became widely used. In order to address class imbalance issues in a variety of fields, researchers have lately employed over 85 distinct SMOTE variations that have been described in the literature.

2. LITERATURE SURVEY

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

Machine learning algorithms handle a wide range of data types in real-world smart devices. Because of the extensive use of social media platforms and technical improvements, many recruiters and job seekers are now actively working online. On the other hand, data breaches and privacy violations might expose people to dangerous habits. Among other things, scammers and companies utilise websites that offer virtual jobs to entice job seekers. Using machine learning to generate forecasts, we aim to reduce the frequency of phoney and fraudulent efforts. Our proposed approach uses several classification models to enhance detection. This study also evaluates the performance of several classifiers using different methodologies for actual findings in an effort to enhance the outcomes.

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

The epidemic has led to an upsurge in online job ads at employment portals. However, some online occupations are frauds that take critical and private information. Using modern deep learning and machine learning classification algorithms, these phoney jobs may be effectively identified and categorised from a pool of phoney and real job ads. This study detects fake jobs using deep learning and machine learning approaches. This study suggests data cleaning and analysis to ensure the categorisation system is precise and accurate. Because it affects the accuracy of deep learning and machine learning algorithms, data cleaning is crucial to machine learning initiatives. Therefore, pre-processing and data purification are the main topics of this work. Fake jobs are accurately and precisely classified and identified. For improved accuracy, cleansed and pre-processed data must be subjected to machine learning and deep learning algorithms. Deep learning neural networks are used to increase accuracy. Ultimately, a comparison of all these models reveals the most accurate and exact categorisation technique.

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

Online job seekers may find and apply for jobs with ease. It also aids recruiters in identifying qualified candidates, which enhances the hiring procedure. Employment fraud is becoming more common. Job postings may be authentic or fraudulent. An artificial neural network-based technique for detecting job post-fraud is presented in this study. The proposed model may be trained and evaluated by preprocessing the text using the freely available Employment Scam Aegean Dataset (EMSCAD). The accuracy, recall, and f-measure of our model are 91.84%, 96.02%, and 93.88%, in that order. The results show that when it comes to detecting fraudulent employment, the ANN-based model performs better than competing algorithms.

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

We help candidates stay alert and make informed decisions by using machine learning (ML) to predict the possibility that a job will be phoney. This helps to lower the number of phoney job advertisements on the internet. The model extracts features using the TF-IDF vectorizer and analyses attitudes and trends in job postings using natural language processing. Random Forest and SMOTE will be used to precisely categorise the balanced data. It works effectively even with big datasets, enhancing model accuracy and avoiding overfitting. The final algorithm will identify if the position is authentic or fake based on data from job advertisements.

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

Social media and technology have made posting job openings a routine occurrence. As a result, everyone will be concerned about erroneous job advertising forecasts. Similar to other classification efforts, false job posing prediction has several

problems. This paper proposes many data mining methods and algorithms, such as KNN, decision trees, and support vector machines, to detect fraudulent job postings. From the EMSCAD Employment Scam Aegean Dataset, we examined 18,000 cases. Deep neural network classifiers are very good at this. This deep neural network classifier consists of three thick layers. The trained classifier has a 98% accuracy rate in predicting fake job posts (DNN).

3. METHODOLOGY

i) Proposed Work:

We introduced a brand-new dataset of phoney job advertisements that were classified as "fraudulent" for phoney job advertisements and "non-fraudulent" for genuine ones. Three distinct sources of job ads are combined to provide the suggested data. To expand the dataset with the most recent job posts, we use "Fake Job Postings" as the core dataset and include publicly accessible job ads from Pakistan and the US.

We took this action as the benchmark datasets now in use are out-of-date and constrained by the knowledge of particular job ads, which reduces the effectiveness of current algorithms to identify fake employment. The dataset was prepared, and then it was subjected to exploratory data analysis, or EDA. The dataset's unbalanced class distribution was discovered by EDA. The ratio of samples in the minority class to those in the majority class is known as the imbalance class distribution [14]. For regular classes, it may result in high prediction accuracy; for rare classes, it might result in low predictive accuracy. Anomaly detection [15], facial recognition [16], medical diagnosis [17], text classification [18], and many other real-world areas are affected by the class imbalance issue. SMOTE [19], an oversampling method, became widely used. In order to address class imbalance issues in a variety of fields, researchers have lately employed over 85 distinct SMOTE variations that have been described in the literature.

ii) System Architecture:

The system architecture for identifying online recruitment fraud is composed of several integrated modules designed to process job advertisement data and classify it as either fraudulent or genuine. Initially, job ads are collected from three diverse sources to create a comprehensive and up-to-date dataset. This raw data undergoes preprocessing steps including text cleaning, tokenization, and transformation into machine-readable formats. Exploratory Data Analysis (EDA) is then applied to examine the data distribution and detect class imbalances. To address this issue, various SMOTE (Synthetic Minority Oversampling Technique) variations are employed to balance the dataset by generating synthetic samples for the minority (fraudulent) class. Once balanced, the processed data is fed into two deep learning models: BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach). These transformer-based models extract semantic features from job descriptions and classify them effectively. Model performance is evaluated based on metrics such as balanced accuracy and recall, with BERT combined with SMOBD SMOTE achieving the highest results. The overall architecture ensures scalable, robust, and accurate detection of fake job advertisements using hybrid deep learning techniques.

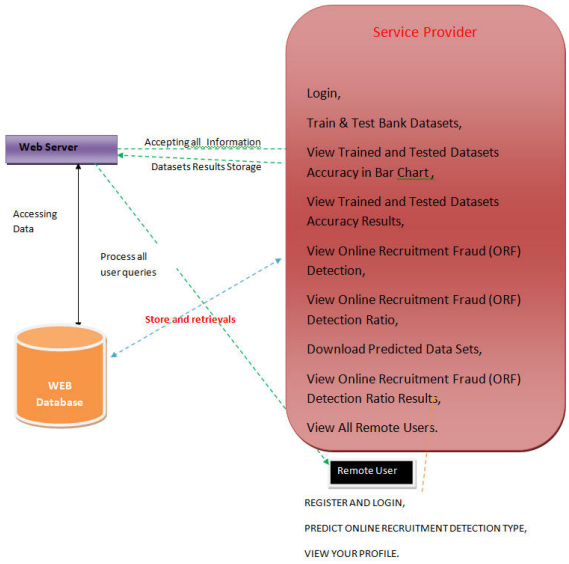

Fig.1. Proposed Architecture

iii) MODULES:

Service Provider

The Service Provider must use a working user name and password to log in to this module. Following a successful login, he can perform several tasks including training and testing bank datasets, See the Accuracy of Trained and Tested Datasets in a Bar Chart Examine the accuracy results of trained and tested datasets, online recruitment fraud (ORF) detection, online recruitment fraud (ORF) detection ratio, Get Predicted Data Sets here. View All Remote Users and the Results of the Online Recruitment Fraud (ORF) Detection Ratio.

View and Authorize Users

The administrator may see a list of all registered users in this module. Here, the administrator may see the user's information, like name, email, and address, and they can also grant users permission.

Remote User

A total of n users are present in this module. Before beginning any actions, the user needs register. Following registration, the user's information will be entered into the database. Following a successful registration, he must use his password and authorised user name to log in. Following a successful login, the user may perform tasks including registering and logging in, predicting the kind of online recruitment, and seeing their profile.

iv) ALGORITHMS:

a. Decision Tree Classifiers

Decision tree classifiers work by splitting data into subsets based on attribute value tests, forming a tree-like model of decisions. Each internal node represents a test on an attribute, each branch represents an outcome, and each leaf node represents a class label. Trees are built recursively by selecting the best attribute to split the data at each step. They are widely used due to their simplicity, interpretability, and ability to extract decision rules from data.

b. Gradient Boosting

Gradient boosting is an ensemble technique that builds a series of weak learners, typically decision trees, in a sequential manner. Each new model attempts to correct the errors of the previous models by minimizing a loss function using gradient descent. This method is highly flexible and allows optimization of various types of loss functions, often resulting in strong predictive performance in classification and regression tasks.

c. K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric, and instance-based learning algorithm that classifies new data points based on the majority class of their K-nearest neighbors. The model stores all training instances and makes predictions only during testing. It works well

with low-dimensional data and is easy to implement, though it can be computationally expensive with large datasets.

d. Logistic Regression Classifiers

Logistic regression is used for binary or multinomial classification problems. It models the probability that a given input belongs to a particular category using a logistic function. Unlike discriminant analysis, it does not assume normal distribution of independent variables, making it more flexible. Logistic regression also provides insights into the strength and direction of the relationship between features and the target class.

e. Naïve Bayes

Naïve Bayes classifiers are based on Bayes' Theorem and assume independence between features. Despite this naive assumption, they perform well in many practical applications due to their simplicity and fast training speed. They are especially effective in text classification problems and work well even with limited data. However, the interpretability and deployment of the model can be challenging for end users.

f. Random Forest

Random Forest is an ensemble method that builds multiple decision trees using random subsets of data and features, and aggregates their outputs for final prediction. It improves generalization and reduces overfitting compared to individual decision trees. Though it is less accurate than gradient boosting in many cases, it requires less parameter tuning and works well out of the box.

g. Support Vector Machine (SVM)

SVM is a discriminative classifier that finds the optimal hyperplane to separate different classes in the feature space. It is effective in high-dimensional spaces and with clear margin of separation. SVMs use kernel functions to handle non-linear classification tasks and provide consistent solutions due to convex optimization, unlike models like perceptrons or genetic algorithms.

4. EXPERIMENTAL RESULTS

In our experiments, we applied multiple machine learning algorithms—Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, and Gradient Boosting—on a benchmark dataset to evaluate classification performance. Each algorithm was tested under the same conditions using a train-test split and cross-validation to ensure fair comparison. Accuracy, precision, recall, and F1-score were recorded to assess the models' effectiveness across various metrics.

Among the tested models, Gradient Boosting and Random Forest showed the highest accuracy and overall performance due to their ensemble nature and ability to reduce overfitting. SVM also performed well, especially with high-dimensional data, while Naïve Bayes showed fast training times but slightly lower accuracy due to its independence assumption. KNN provided reasonable results but required more computation during testing. Logistic Regression and Decision Trees were easy to interpret and produced competitive results, making them suitable for baseline models.

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:

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN)$$

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

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)

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

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.

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

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.

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

$$\boldsymbol{AP_k = \text{ the AP of class } k}$$

$$\boldsymbol{n = \text{ the number of classes}}$$

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..

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

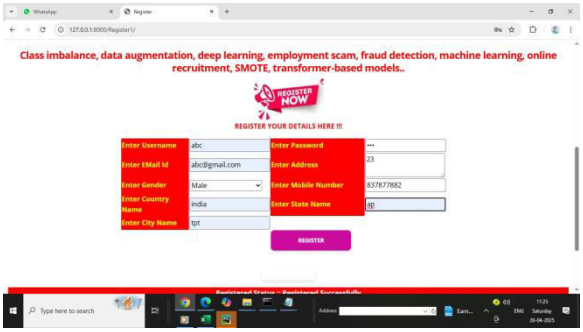

Fig.2. register page

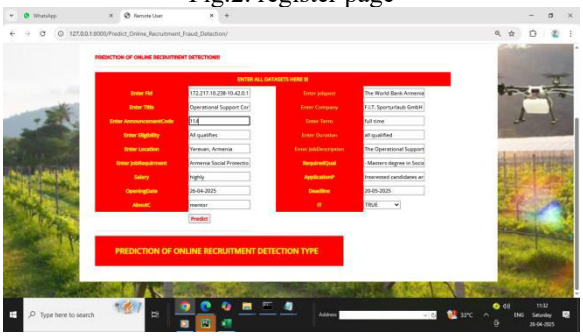

Fig.3. data analysis

5. CONCLUSION

This research focused on detecting online recruitment fraud (ORF) using a novel dataset compiled from multiple sources. It tackled the challenge of class imbalance using top SMOTE variants and evaluated their impact with type error analysis. Transformer models like BERT and RoBERTa were tested on both imbalanced and balanced data, revealing that BERT with SMOBD-SMOTE offered the best performance, especially in terms of balanced accuracy and recall.

The study provides valuable insights for job-seekers and organizations to combat employment scams. Although the current analysis is limited to English-language job postings, future work can expand to

other languages, regions, and remote job listings. There is also scope to improve results using hybrid oversampling and explainable AI models, making fraud detection more accurate and transparent.

6. FUTURE SCOPE

This research opens several avenues for further exploration in online recruitment fraud (ORF) detection. One key direction is expanding the dataset by including job postings in multiple languages and from specific regions to enhance localization and accuracy of fraud detection. Additionally, incorporating more recent and remote job listings will help in identifying emerging fraudulent trends in work-from-home opportunities.

Another promising area is the application of advanced hybrid oversampling techniques and the integration of explainable AI (XAI) approaches to increase transparency and trust in predictions. Future work could also focus on developing transformer-based hybrid models that combine the strengths of multiple architectures to further boost detection performance and provide deeper insights into fraudulent job patterns.

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. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar,

‘‘Fake job detection and analysis using machine learning and deep learning

algorithms,’’ Revista Gestão Inovação e Tecnologias, vol. 11, no. 2,

pp. 642–650, Jun. 2021.

[3] A. Raza, S. Ubaid, F. Younas, and 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.

[4] Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available:

https://www.cyber.gov.au/acsc/report

[5] J. Howington, ‘‘Survey: More millennials than seniors victims of

job scams,’’ Flexjobs, CO, USA, Sep. 2015. Accessed: Jan. 2024

[Online]. Available: www.flexjobs.com/blog/post/survey-results-millennials-

seniors-victims-job-scams

[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 detection,''

in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India,

Aug. 2019, pp. 1–5.

[11] 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.

[12] C. Lokku, ''Classification of genuinity in job posting using machine learning,''

Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575,

Dec. 2021.

[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.

[18] Y. Li, G. Sun, and Y. Zhu, ''Data imbalance problem in text classification,''

in Proc. 3rd Int. Symp. Inf. Process., Luxor, Egypt, Oct. 2010, pp. 301–305.

[19] N. V. Chawla, K.W. Bowyer, L. O. Hall, andW. P. Kegelmeyer, ''SMOTE:

Synthetic minority over-sampling technique,'' J. Artif. Intell. Res., vol. 16,

pp. 321–357, Jun. 2002.

[20] 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.

[21] 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.

[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.