International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 2 Volume No.9 March - April – 2025
DOI:10.46647/ijetms.2025.v09i02.019 ISSN: 2581-4621

DEEP LEARNING SOLUTIONS FOR DETECTING FRAUD IN ONLINE RECRUITMENT

G.Sravani Prasad[1], A.Emmanuel Raju[2]
[1]M.Tech Student, Dept of CSE, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh
[2]Asst.Professor, Dept of CSE, Dr.K.V.Subba Reddy Institute of Technology, Kurnool, Andhra Pradesh

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-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 Existing benchmark datasets are outdated and limited due to knowledge of specific job postings, which limits the existing models' capability in detecting fraudulent jobs.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%.

Keywords: Fake Job, Profile, EDA, SMOTE

Introduction

Online recruitment platforms have revolutionized the hiring process by providing organizations with access to a vast pool of potential candidates. However, the rapid digitization of recruitment has also led to an increase in fraudulent activities, including fake job postings, identity theft, resume fraud, and scam employers. Traditional fraud detection methods often rely on rule- based or statistical approaches, which may struggle to adapt to evolving fraudulent tactics.

Deep learning, a subset of artificial intelligence (AI), has emerged as a powerful tool for detecting fraud in online recruitment. By leveraging neural networks and advanced machine learning techniques, deep learning models can identify patterns, anomalies, and suspicious behaviors that may indicate fraudulent activity. These models can analyze vast amounts of structured and unstructured data, including job descriptions, applicant profiles, social media activity, and historical fraud patterns, to improve fraud detection accuracy.

This paper explores the role of deep learning in detecting fraud in online recruitment, discussing various model architectures, data sources, and evaluation metrics used for fraud detection. Additionally, it highlights the challenges and future directions in developing robust deep learning

solutions for securing online hiring platforms.
The primary objective of this study is to develop and evaluate deep learning-based solutions for detecting fraud in online recruitment platforms. Specifically, the objectives include:
Identifying Fraudulent Activities – Recognize various types of fraud, such as fake job postings, fraudulent recruiters, identity theft, and manipulated resumes, within online recruitment systems.
Developing Deep Learning Models – Design and implement deep learning models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), and transformer-based models, to detect fraudulent activities based on structured and unstructured recruitment data.

Feature Extraction and Data Processing – Utilize advanced natural language processing (NLP) techniques, such as word embeddings, sentiment analysis, and entity recognition, to extract meaningful features
from job descriptions, resumes, and user interactions.

Enhancing Fraud Detection Accuracy – Improve detection performance by integrating multiple data sources, including job listings, applicant behaviours, recruiter histories, and social media data, using deep learning techniques.
Evaluating Model Performance – Assess the effectiveness of different deep learning models using appropriate evaluation metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC).

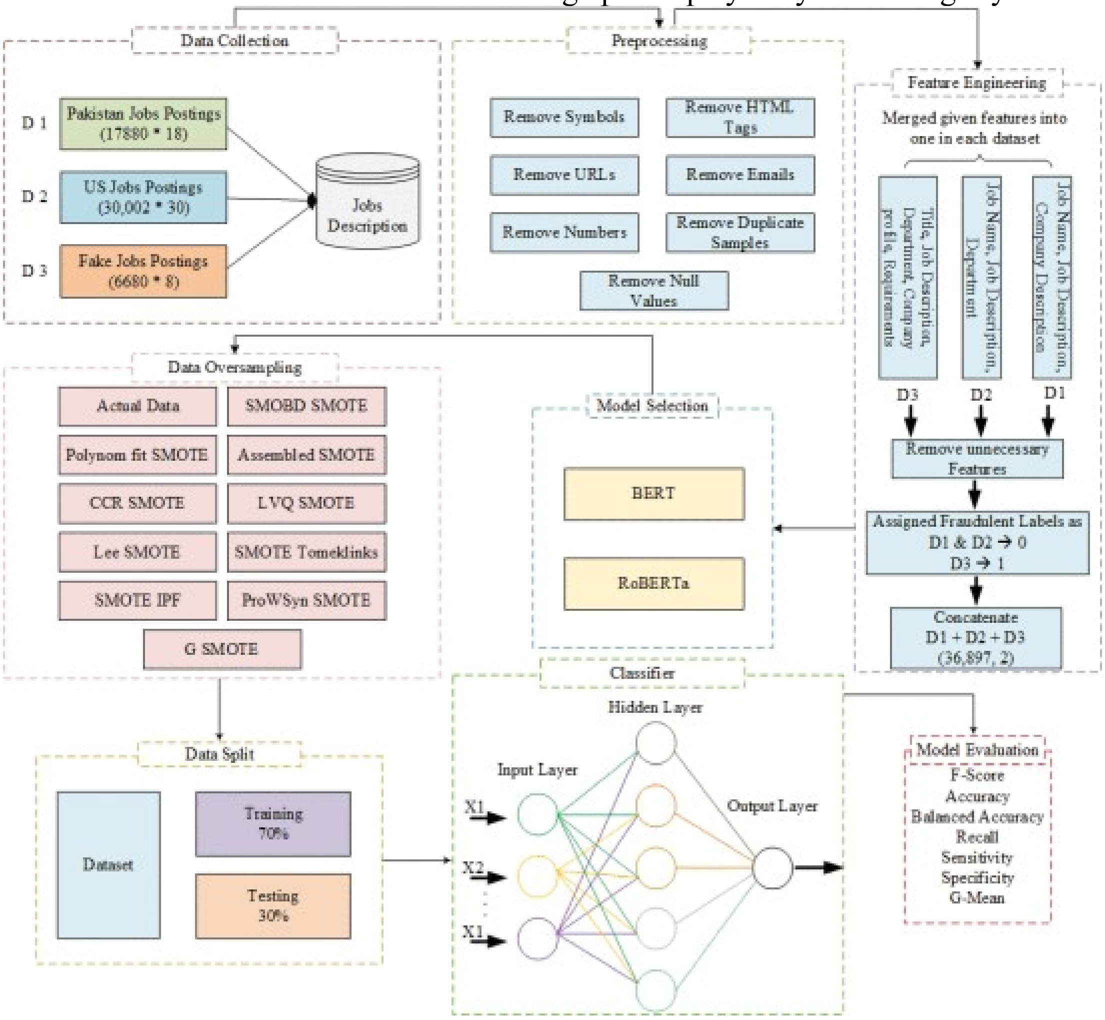
Developing a Scalable and Adaptive Framework – Propose a scalable, real-time fraud detection framework that can adapt to evolving fraudulent tactics using continuous learning and model retraining.

Ensuring Ethical and Fair AI Implementation – Address ethical concerns, such as bias in AI models, transparency, and data privacy, to ensure responsible and fair decision-making in fraud detection systems.

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 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.

Implementation

This project is implements like application using python and the Server process is maintained using the SOCKET & SERVERSOCKET and the Design part is played by Cascading Style Sheet.



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. 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.

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.

For a more comprehensive examination, it is possible to conduct a similar analysis on job postings published in languages other than the current dataset. Given the rising popularity of online recruitment, the dataset can be enriched by incorporating the latest job postings. This study encompasses job postings from diverse regions worldwide. A comparable analysis can be conducted specifically for job postings within a particular region to ascertain the rate of fraudulent postings, thereby serving the public's best interests. Furthermore, the inclusion of job postings that pertain to remote work opportunities through online platforms can be deemed crucial for the creation of a novel dataset, given the significant prevalence of fraudulent activities associated with online jobs from home.

In the present research, a range of SMOTE variants were employed to address class distribution imbalance. To attain even more precise results, the utilization of hybrid oversampling techniques can be considered. For future research, explainable AI and novel transformer based hybrid models need to be explored.

References

[1] V. Stilwell, *There are now more than five million job openings in america*, 2015. [Online]. Available: https://www.bloomberg.com/news/articles/2015- 02-10/job-openings-in-u-s-rose-by-181- 000-in-december-to-5-03-million.

[2] J. Howington. (2015). "Survey: More millennials than seniors victims of job scams," [Online]. Available: https://www. flexjobs. com/blog/post/survey- results-millennials-seniors-victims-job- scams/.

[3] Z. A. Palak, *Digital bangladesh - a story of transformation*, 2019. [Online]. Available: https://www.daily-sun.com/post/407497/Digital-Bangladesh--a- Story-of-Transformation.

[4] 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, 2017.

[5] S. Morgan and M. Park, "Cybercrime report 2017," 2017. [Online]. Available: https://www.herjavecgroup.com/cybercrime-report-2017/.

[6] "Personal fraud," 2014-2015. [Online]. Available: https://www.abs.gov.au/ statistics/people/crime- and-justice/personal-fraud/latest-release.

[7] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, "Orfde- tector: Ensemble learning based online recruitment fraud detection," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 2019,
pp. 1–5.

[8] B. Alghamdi and F. Al-Harby, "An intelligent model for online recruitment fraud detection,"
Journal of Information Security, vol. 10, pp. 155–176, 2019.

[9] T. Suryanto and R. R. A. Saleh, "Employee recruitment fraud prevention with the implementation of decision support system," *Journal of Physics: Confer- ence Series*, vol.

1028, p. 012 055, Jun. 2018.

[10] S. Vidros, C. Kolias, and G. Kambourakis, "Online recruitment services: An- other playground for fraudsters," *Computer Fraud Security*, vol. 2016, pp. 8– 13, Mar. 2016.

[11] S. Mahbub and E. Pardede, "Using contextual features for online recruitment fraud detection," Aug. 2018.

[12] S. G. Thiago and M. C. Walmir, "A review of machine learning approaches to spam filtering,"
Expert Systems with Applications, vol. 36, no. 7, pp. 10 206–
10 222, 2009, ISSN: 0957-4174.

[13] J. R. Scanlon and M. S. Gerber, "Automatic detection of cyber-recruitment by violent extremists,"
Security Informatics, vol. 3, no. 1, p. 5, 2014.

[14] K. Shivagangadhar, H. Sagar, S. Sohan, and C. Vanipriya, "Fraud detection in online reviews using machine learning techniques," *International Journal of Computational Engineering Research (IJCER)*, vol. 5, no. 5, pp. 52–56, 2015.

[15] D. Sculley and G. M. Wachman, "Relaxed online svms for spam filtering," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Re- search and Development in Information Retrieval*, ser. SIGIR '07, Amsterdam, The Netherlands: Association for Computing Machinery, 2007, pp. 415–422.

[16] Newtium, *Steps of predictive modelling*. [Online]. Available: https://newtiummedia. blob.core.windows.net/images/Steps-to-Predictive-Modelling.jpg (visited on 10/04/2020).

[17] I. S. Candanedo, E. H. Nieves, S. R. Gonzâlez, M. T. S. Martın, and A. G. Briones,
Machine learning predictive model for industry 4.0, Springer, 2018.

[18] A. Kafi, M. S. A. Alam, S. B. Hossain, S. B. Awal, and H. Arif, "Feature- based mobile phone rating using sentiment analysis and machine learning approaches," in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, IEEE, 2019, pp. 1–6.

[19] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas,
A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning soft- ware: Experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[20] O. Pauly, "Random forests for medical applications," Ph.D. dissertation, Tech- nische Universität München, 2012.

[21] G. James, "Majority vote classifiers: Theory and applications," 1998.

[22] *Welcome to lightgbm's documentation!* 2018. [Online]. Available: https : / /li

[23] *Lightgbm (light gradient boosting machine)*, 2020. [Online]. Available: https:
//www.geeksforgeeks.org/lightgbm-light-gradient-boosting-machine/ (visited on 12/25/2020).

[24] T. Parr and J. Howard, *How to explain gradient boosting*. [Online]. Available: https://explained.ai/gradient-boosting/index.html (visited on 12/25/2020).

[25] S. Touzani, J. Granderson, and S. Fernandes, "Gradient boosting machine for modeling the energy consumption of commercial buildings," *Energy and Buildings*, vol. 158, pp. 1533–1543, 2018.

[26] O. González-Recio, J. Jiménez-Montero, and R. Alenda, "The gradient boost- ing algorithm and random boosting for genome-assisted evaluation in large data sets," *Journal of dairy science*, vol. 96, no. 1, pp. 614–624, 2013.

[27] *Boosting algorithm: Adaboost*, 2017. [Online]. Available: https://towardsdatascience. com/boosting- algorithm- adaboost- b6737a9ee60c?gi=f8043aeb37ba (visited

on 12/25/2020).

[28] K. Vidhya and R. Shanmugalakshmi, “Improved diabetic data analytic model for complication prediction,” *Int. J. Eng. Adv. Technol*, 2019.