WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Transformer Based Online Job Fraud Detection Using Albert And Smote

Sweatha M S[1], Dr. A Christiyana Arulselvi[2]

[1]Scholar, Department of MCA

[2]Associate Professor, Department of MCA, Dr. M.G.R Educational and Research Institute

[1]*sweathams27@gmail.com*

A B S T R A C T

Online job platforms have made the job search easier for people, but they have also enabled scammers to upload fake job advertisements. These deceptive listings can mislead job seekers and waste their time and resources. This project presents a system for detecting fraudulent job postings using a deep learning model called ALBERT (A Lite BERT) and a technique known as SMOTE, which tackles the challenge of unbalanced data where authentic jobs exceed counterfeit ones. We gathered a dataset from Kaggle and incorporated extra job postings to improve the dataset's thoroughness. The job descriptions are polished and analyzed, then delivered to the ALBERT model, which trains to identify patterns in fraudulent listings. SMOTE is used to balance the dataset so the model focuses on both actual positions and synthetic ones. This system helps job seekers safeguard against scams in online recruitment.

Keywords: Online job, Fraud Detection, Deep Learning Integration, ALBERT, SMOTE, Job Posting Classification

1. Introduction

Websites for online jobs help link job seekers with employers [1]. Nonetheless, these platforms can be misused by people who create fake job postings or pretend to be real employers to trick users [2], [3]. Identifying and thwarting these scams promptly is crucial for protecting job seekers and preserving confidence in these online platforms [2]. Conventional fraud detection approaches, such as rule-based systems and simple machine learning methods, have lost efficacy because of the changing patterns of fraudulent activities [5].

To tackle this issue, this initiative utilizes a sophisticated deep learning model called ALBERT (A Lite BERT), which can more effectively comprehend the context of job descriptions and detect fraudulent listings with improved precision. This research centers on identifying fraudulent online recruitment activities [2]. The dataset created for this aim was compiled from three sources, with the main source being the Fake Job Postings dataset found on Kaggle. To improve the dataset's relevance and completeness, more recent job postings were included as well. After collecting data, Exploratory Data Analysis (EDA) was performed to uncover patterns and trends [4]. Fake job postings were noted to represent a considerably smaller part of the dataset in comparison to real postings, resulting in a class imbalance problem [15], [19].

This disparity threatens the model's effectiveness, potentially leading it to favor genuine jobs and ignore fraudulent ones. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was utilized. SMOTE is a commonly used method for tackling class imbalance and has proven effective in areas like healthcare, fraud detection, and text classification [15], [19].

2. Literature Review

This section offers a summary of research concerning Transformer Based Online Job Fraud Detection and tackles the issue of class imbalance present in datasets utilized for these purposes [15]. Recent studies by researchers like Kumar and Garg (2020) and others have indicated a notable rise in deceitful practices on online job platforms. Different machine learning models have been utilized to identify such actions. Conventional classifiers such as decision trees and random forests, employed in investigations by Maheshwari et al. (2019), have shown restricted effectiveness in understanding the nuances of textual fraud because of the intricate language found in job advertisements [4], [6].

To address these constraints, natural language processing (NLP) methods have been more widely utilized. Scholars such as Shamsi et al. (2021) and Zhao et al. (2018) have investigated the application of models like BERT and its alternatives to examine job descriptions and retrieve textual attributes. These investigations concentrate on identifying patterns indicative of fraud and have demonstrated that deep learning models surpass conventional methods regarding accuracy and dependability [3]. ALBERT (A Lite BERT), presented by Lan et al. (2019), has attracted interest due to its effectiveness and efficiency in NLP applications. Research has shown that ALBERT performs at least as well as BERT or even surpasses it with fewer

parameters, which makes it ideal for specialized fields such as detecting recruitment fraud. Its ability for fine-tuning enables it to adjust efficiently to specific datasets.

Moreover, researchers such as Pathak and Srivastava (2020) have explored sentiment analysis as a technique to aid in fraud detection. These investigations indicate that negative or excessively amplified sentiment in job listings could signify deceptive motives. ALBERT's capacity to produce context-sensitive embeddings renders it beneficial for improving sentiment-driven models [2], [5].

3. METHODOLOGY

To accurately identify the successfully address online recruitment fraud, a tailored dataset was created by gathering and merging job listings from Kaggle public sources. These sources offered varied and unorganized job listing data, which was meticulously cleaned and processed to ready it for model training. This encompassed managing absent or null values, standardizing the case and format of text content, eliminating duplicate entries, and unifying inconsistent column headers for smooth integration. The textual elements, especially job descriptions, were processed with ALBERT's tokenizer to guarantee a profound contextual comprehension of language meanings. A train-validation division was utilized to guarantee the model's performance assessment occurred in realistic and impartial circumstances. In the data analysis stage, a significant class imbalance was noted, with authentic job postings greatly exceeding fraudulent ones. This created a possibility of model bias, causing the classifier to favor predicting the majority class. To address this issue, synthetic resampling techniques were utilized—namely, enhanced variants of SMOTE like SMOTE-NC (for managing both categorical and numerical attributes) and SMOBD (Borderline-SMOTE to concentrate on challenging instances). These methods synthetically created samples of the minority class, thus enhancing the training dataset and allowing the model to more efficiently learn patterns associated with fraud. Techniques for feature selection were utilized to preserve only the most pertinent attributes—like job title, description, and company profile—while eliminating noise or less valuable features. The polished dataset was subsequently employed to enhance ALBERT (A Lite BERT), a transformer-driven language model tailored for NLP applications. The HuggingFace Transformers library was utilized to adapt ALBERT for binary classification, differentiating between authentic and fraudulent job postings. Hyperparameter tuning was carried out iteratively by experimenting with various learning rates, batch sizes, and epoch counts to enhance the model's performance. The training procedure was carried out on a PyTorch framework with early stopping techniques implemented to terminate training when the validation

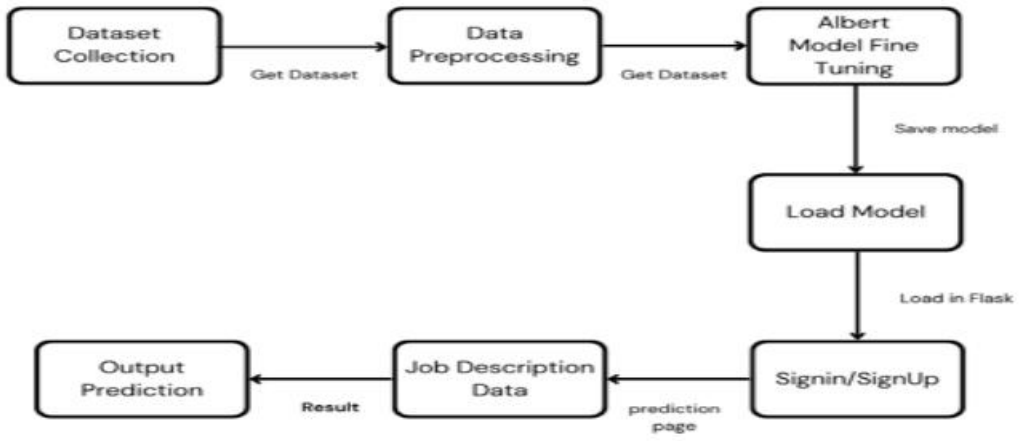

Fig. 1 – System Architecture.

4. Functional Description

This stage starts with configuring the environment by installing necessary libraries, including Transformers, PyTorch, scikit-learn, imbalanced-learn, pandas, and Flask. To improve computational efficiency, particularly in training, the environment is set up to leverage GPU acceleration. After the setup is finalized, the dataset goes through thorough preprocessing, which involves addressing missing values, standardizing text formats, eliminating duplicates, and ensuring column headers are aligned for uniformity. Job descriptions are subsequently tokenized utilizing ALBERT's tokenizer to convert the text into a format appropriate for deep learning input. Tackling class imbalance is an essential aspect of the process, attained through the use of synthetic resampling methods like SMOTE and its enhanced variations such as SMOTE-NC and SMOBD. These techniques aid in producing artificial samples for the minority class, thus enhancing the model's capacity to generalize without favoring the majority class. After data preparation, the ALBERT model undergoes fine-tuning with a supervised learning method on the balanced and processed dataset. Model effectiveness is evaluated using multiple metrics such as precision, recall, F1-score, and balanced accuracy, guaranteeing an equitable assessment of both classes. Hyperparameter tuning is conducted to find the best values for learning rate, batch size, and training epochs, resulting in improved prediction accuracy. After the model shows consistent performance, it is launched via an easy-to-use web interface created with Flask. This online tool enables users to enter job descriptions and get immediate predictions about the authenticity of the job listing, indicating if it is legitimate or a scam. A MySQL database is incorporated to record user queries and prediction outcomes, facilitating long-term usage and feedback evaluation. Furthermore, logging systems are established to track model outputs and gather user input, facilitating ongoing enhancement and improved fraud detection as time progresses..

5. Result

The suggested system successfully distinguished between genuine and fraudulent job postings by employing the ALBERT model combined with enhanced SMOTE methods. Of all the configurations evaluated, the pairing of ALBERT and SMOBD produced the most favorable results, achieving approximately 90% balanced accuracy and strong recall, thereby minimizing the risk of overlooking fraudulent job postings**.** This configuration outperformed conventional machine learning models and even other transformer models such as BERT, particularly with imbalanced data. The model was integrated into a live web application, allowing users to input job descriptions and receive fast and precise results. The web application featured a seamless interface and utilized MySQL for secure data storage. The findings indicated that employing transformer models with adequate data balancing and sanitization significantly enhances the identification of fraudulent job listings on the internet.

6. Conclusion

In summary, this project presents a robust AI-driven method for identifying Online Recrruitment Fraud . It provides not only an efficient model but also a comprehensive implementation that can be utilized on practical platforms Future efforts could concentrate on model explainability, multilingual capabilities, and broadening the dataset to enhance adaptability and confidence in varied surroundings

References

1. P. Kaur, "E-recruitment: A conceptual study", *International Journal of Applied Research*,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 Inovaçãoe Tecnologias, vol. 11, no. 2, pp. 642–650, Jun. 2021
3. Raza, A., Ubaid, S., Younas, F., Akhtar, F., “Fake e-job posting prediction based on advanced machine learning,” Procedia Computer Science, 2020.
4. Australian Cyber Security Centre, “Online Fraud,” Jun. 2022.
5. J. Howington, “Survey: More millennials than seniors victims of job scams,” FlexJobs, Sep.2015
6. 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, pp. 6, Mar. 2017.
7. B. Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection", Journal of Information Security, vol. 10, no. 3, pp. 155–176, 2019.D. Wang, C. Liu, et al., "ICFHR 2020 Competition on Offline Recognition and Spotting of Handwritten Mathematical Expressions (OffRaSHME)," in 17th Int. Conf. on Frontiers in Handwriting Recognition, IEEE, 2020.
8. S. Lal et al., "ORFDetector: Ensemble learning-based online recruitment fraud detection", Proceedings of the 12th International Conference on Contemporary Computing (IC3), pp. 1–5, Aug. 2019
9. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, "Online recruitment fraud detection using ANN", Proceedings of the Palestinian International Conference on Information and Communication Technology (PICICT), pp. 13–17, Sep. 2021.
10. C. Lokku, "Classification of genuinity in job posting using machine learning", International Journal of Research in Applied Science and Engineering Technology, vol. 9, no. 12, pp. 1569–1575, Dec. 2021.
11. Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification", Proceedings of the 3rd International Symposium on Information Processing, pp. 301–305, Oct. 2010
12. S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review", GESTS International Transactions on Computer Science and Engineering, vol. 30, no. 1, pp. 25–36, 2006
13. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002
14. Z. Lan et al., "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations", Proceedings of ICLR, 2020
15. N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique", Journal of AI Research, 2002.
16. Pathak, R. & Srivastava, S., "Sentiment-based job fraud detection", Journal of AI Research, 2020.
17. Shamsi, A., et al., "Detecting job fraud using BERT-based contextual embeddings", Journal of Information Security, 2021.
18. Zhao, Y., et al., "Job advertisement fraud detection using deep learning", ACM Digital Threats, 2018.
19. S. Kotsiantis et al., "Handling imbalanced datasets: A review", GESTS Int’l Trans. CS & Engg, 2006.