Proceedings of the Third International Conference on Augmented Intelligence and Sustainable Systems (ICAISS-2025)
DVD Part Number: CFP25CB2-DVD; ISBN: 979-8-3315-0723-7

Detection of fake online recruitment using machine learning approach

Jayanth Medapati
Department of Artificial Intelligence and Data Science,
Vardhaman College of Engineering
Hyderabad, India
jayanthmedapati18@gmail.com

Yashaswi Arradi
Department of Artificial Intelligence and Data Science,
Vardhaman College of Engineering
Hyderabad, India
arradiyashaswi@gmail.com

Ronan Kongala
Department of Artificial Intelligence and Data Science,
Vardhaman College of Engineering
Hyderabad, India
kongalaronan@gmail.com

ShanmugasundaramHariharan
Department of Artificial Intelligence and Data Science,
Vardhaman College of Engineering,
Hyderabad, India
mailtos.hariharan@gmail.com

J. Shanmugapriyan
Dept of Electrical and Electronics Engineering,
NSN College of Engineering and Technology
jspriyan@gmail.com

Karuppiah Natarajan
Dept of Electrical and Electronics Engineering,
Vardhaman College of Engineering
Hyderabad, India
natarajankaruppiah@gmail.com

***Abstract*—While many organizations these days prefer to post their job opportunities on the web so that job seekers can access them conveniently and easily, this practice might be an example of scam from the side of swindlers who offer job hunters tasks and services in exchange for money. Many people fall victims to this type of fraud and lose a considerable amount of money as a result. The proposed approach uses a variety of machine learning algorithms inclusive of supervised learning tools and natural language processing methods to analyze and sort job advertisements. By using both single classifiers and ensemble classifiers, the system assesses results and compares them, thus recognizing fraudulent job advertisements on the Internet. Model performance will be evaluated using metrics like accuracy, precision, recall, and F1-score. This study aims to demonstrate the potential of boosting techniques for achieving high accuracy in fake job posts prediction, potentially leading to improved outcomes Therefore, the value of the research in helping to create a more secure online job market can serve to establish a level of trust for job seekers and provides them with protection from the financial and emotional risks related to the misuse of deceptive job postings.**

***Index Terms*—Machine Learning; Natural Language Processing; Boosting.**

I. INTRODUCTION

In recent years, the rapid growth of online recruitment platforms has provided individuals with unprecedented access to job opportunities[7, 8]. However, this has also led to a rise in fraudulent job postings, which deceive job seekers and pose significant risks to their personal information and financial security [2, 6]. Detecting such fake job advertisements has become a pressing challenge, as manual screening processes are inefficient and prone to errors due to the massive volume of job postings generated daily.Traditional machine learning models often struggle with this task due to the class imbalance inherent in recruitment datasets, where the majority of postings are legitimate, and only a small fraction are fraudulent [5, 9]. This imbalance skews the models toward predicting the majority class (legitimate jobs), resulting in poor detection rates for fake postings [11].

To address these challenges, this project proposes a machine learning-based system for the detection of fake online recruitment. The proposed approach focuses on improving detection accuracy by utilizing Natural Language Processing (NLP) techniques for feature extraction and employing advanced oversampling methods, such as SMOTE and ADASYN, to balance the dataset. Additionally, we explore the performance of various machine learning models, including Random Forest, Adaboost, Gradient Boosting, and XGBoost, and apply hyperparameter tuning to optimize results [3, 9, 12].

A. *Random Forest*

Random Forest is an ensemble learning method that builds multiple decision trees during training [3]. It combines the predictions of all trees to make a final decision, which improves accuracy and reduces overfitting. It's particularly effective for classification tasks and handles large datasets well. In this project, Random Forest was used for its ability to capture complex patterns in job postings [13].

B. *Adaboost*

Adaboost (Adaptive Boosting) is an ensemble technique that improves model performance by combining several weak classifiers into a strong one. It focuses on the instances that previous models misclassified by assigning higher weights to them. For detecting fake job postings, Adaboost helps by refining predictions through iterative learning.

C. *Gradient Boosting*

Gradient Boosting is another ensemble technique that builds models sequentially, where each new model corrects the errors of the previous ones [5,15]. It optimizes the loss function (error) using gradient descent, resulting in highly accurate models. In this project, Gradient Boosting was used to enhance classification accuracy by iteratively improving the model [4].

D. *XGBoost*

XGBoost (Extreme Gradient Boosting) is an advanced form of gradient boosting that includes regularization to prevent overfitting and improve performance. It's known for its efficiency and speed in handling large datasets. XGBoost was

employed in this project for its superior ability to handle complex classification problems with imbalanced data.

E. *SMOTE*

Synthetic Minority Over-SamplingTechnique(SMOTE) is a data balancing technique used to address class imbalance by generating synthetic samples for the minority class. This create fake instances based on existing ones, which helps the model learn better from minority class examples (fake job postings). In this project, SMOTE was applied to balance the dataset and improve detection accuracy [1].

F. *ADASYN*

Adaptive Synthetic Sampling(ADASYN) Similar to SMOTE, ADASYN generates synthetic samples for the minority class but focuses more on the difficult-to-classify instances. It adjusts the number of synthetic samples created based on the distribution of minority examples. ADASYN was used in this project to further enhance the model's ability to detect fake job postings by focusing on harder-to-learn cases [2].

II. Related Work

The detection of fake online job postings has become a significant area of research due to the increasing prevalence of online recruitment platforms and the associated risks of fraudulent advertisements. Machine learning techniques, particularly ensemble methods, have been extensively explored to address this challenge by enhancing classification accuracy and reducing error rates.The ensemble approach is a robust method for improving system precision by combining multiple machine learning algorithms. Random Forest (RF), a classification-based ensemble method, constructs multiple tree-like classifiers using subsets of the data. Each tree votes for the most appropriate class, with the final decision based on majority voting. This method enhances model robustness and generalization, making it effective in detecting anomalies, including fraudulent job postings [3], [13].

Boosting techniques, such as AdaBoost and Gradient Boosting, have also demonstrated remarkable success in handling classification problems. AdaBoost combines weak learners by iteratively adjusting the weights of misclassified instances, resulting in a stronger overall model [4]. Gradient Boosting, on the other hand, sequentially optimizes a loss function by correcting the errors of prior learners through gradient descent. These methods are particularly effective in tasks requiring high precision, such as spam filtration and fraud detection [14]. To address the issue of imbalanced datasets, researchers have utilized data balancing techniques such as the Synthetic Minority Over-Sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN). SMOTE generates synthetic samples by interpolating between existing minority class instances, while ADASYN focuses on creating more samples in challenging regions of the feature space. These methods improve the detection rates of minority classes, such as fake job postings, without compromising the performance on legitimate data [1, 2, 9].

Natural Language Processing (NLP) techniques have also played a critical role in feature extraction for fake job detection. Methods like Term Frequency-Inverse Document Frequency (TF-IDF) and word embedding enable the extraction of meaningful textual features, which are essential for identifying fraudulent behavior in job descriptions [6], [12]. Integrating these features with ensemble classifiers has led to significant improvements in precision, recall, and F1-scores.

Studies in related domains, such as fake news detection, provide additional insights. Fake news detection models analyze how false information is written, distributed, and connected to users. These models utilize social and contextual features to build learning frameworks, which can be adapted for fake job detection. By leveraging techniques from fake news research, researchers aim to improve the identification of fraudulent job advertisements [6], [10].The existing body of work highlights the effectiveness of combining advanced ensemble methods, data balancing techniques, and NLP-based feature extraction to detect fake job postings. These approaches form the foundation for developing robust and scalable systems that safeguard job seekers and improve the integrity of online recruitment platforms [5, 8, 11].

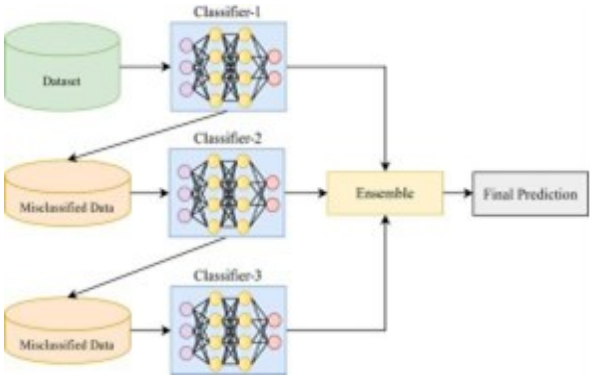

Fig.1 Boosting algorithm

An improved machine learning approach involves using an iterative process where subsequent algorithms address and rectify the errors of earlier iterations. For instance, the second algorithm identifies and corrects the inaccuracies of the initial decision tree, while the third algorithm refines the errors made by both the first and second trees. This iterative refinement continues through multiple levels, enhancing the overall model's accuracy and robustness. The final prediction is derived from a collective decision, integrating the outputs of all individual trees to ensure a comprehensive and precise outcome. This method exemplifies the strength of ensemble models in handling complex data patterns and improving classification accuracy.Therefore, this research is focused on the above algorithms in detecting fake job posts. By investigating these two algorithms, we can find a feasible and appropriate algorithm to find fraudulent job postings in social media.

III. Proposed Work

A. *Data set*

The Public Employment Scam Aegon Dataset is a comprehensive collection of data that aims to uncover fraudulent activities within public employment schemes.It's got

all sorts of information like employment records, personal details of employees, salary info, job positions, and how long people have been employed for. It might even have some extra stuff like timestamps, locations, and communication logs. This dataset is designed to help researchers and analysts find patterns and irregularities that could signal fraud, like fake employees, phony work records, and misused funds. It's a really detailed and organized dataset, which makes it super useful for developing machine learning models and other tools to catch and prevent employment scams in public organizations. This dataset is structured in a way that makes it easy to do all sorts of analysis. For example, you can cross- reference employment records with personal info to make sure everything checks out. You can also look at salary details to spot any weird inconsistencies or outliers. And the job position and duration data can help you figure out if someone's employment history is legit. With advanced analytics and machine learning algorithms, you can automate the detection of suspicious activities and generate reports that highlight potential fraud. So, this Public Employment Scam Aegon Dataset is a crucial tool for government agencies, auditors, and researchers who want to make sure public employment systems are honest and prevent any financial misconduct. Fake job postings pose a significant risk to job seekers, leading to financial loss and identity theft. The task of detecting these fraudulent postings is complicated by the inherent imbalance in the dataset, where legitimate job postings vastly outnumber the fake ones.

B. *Implementation Details:*

Traditional machine learning models frequently encounter challenges in dealing with class imbalances, especially in scenarios where minority classes—such as fraudulent job postings—are significantly underrepresented. This imbalance not only skews predictions toward the majority class but also reduces the effectiveness of detecting the minority class. Addressing this issue requires innovative techniques capable of balancing datasets and enhancing model performance.To address these limitations, we propose a robust system that incorporates advanced oversampling techniques, specifically Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Over-SamplingTechnique (SMOTE). These techniques are designed to create a more balanced dataset by synthesizing new data points for the underrepresented class which enables the model to learn effectively from legitimate and fraudulent postings significantly improving the detection of fake postings.

1) *Data Preprocessing and Feature Extraction*

The system initiates with comprehensive preprocessing steps to ensure the dataset is clean and ready for analysis. This includes the removal of duplicates, handling of missing values, and standardization of text formats to maintain consistency. Advanced feature engineering methods are employed to extract valuable information from job postings. Attributes such as job title, company name, job description, location, and salary details are converted into numerical representations using techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddingare crucial for enabling the model to identify subtle patterns etc.

2) *Handling Class Imbalance*

Class imbalance is a critical challenge in detecting fake job postings, as fraudulent cases are often significantly outnumbered by legitimate ones. SMOTE addresses this issue by generating synthetic samples between existing minority class points, ensuring better representation. ADASYN further refines this process by focusing on regions with sparse minority class samples, emphasizing difficult-to-learn cases. Together, these techniques create a balanced dataset, allowing machine learning models to achieve higher sensitivity and precision in detecting fraudulent postings.

3) *Model Training and Optimization*

With the dataset balanced, the system trains machine learning models using ensemble learning techniques such as Random Forest, Gradient Boosting Machines (GBM), and XGBoost. These algorithms are well-suited for handling complex feature interactions, making them ideal for tasks requiring high accuracy and robustness. Hyperparameter tuning is applied to optimize model performance, ensuring that the models achieve superior results across metrics such as accuracy, precision, recall, and F1-score. Additionally, cross-validation techniques validate the models' generalizability, ensuring they perform consistently across diverse datasets.

4) *Impact and Insights*

The implementation of the proposed system significantly improved the detection of fraudulent job postings by addressing critical challenges like class imbalance and enhancing model performance through advanced techniques such as SMOTE and ADASYN. These oversampling methods balanced the dataset, enabling ensemble models like Random Forest, Gradient Boosting, and XGBoost to achieve higher accuracy, recall, and F1-scores. SMOTE particularly stood out by generating reliable synthetic samples that enhanced model training stability. The system's integration into job platforms reduced the number of fake postings, providing a safer environment for job seekers while identifying industries like logistics and online platforms as more vulnerable to fraud. By combining robust classification techniques with actionable insights, the system not only enhances trust in recruitment platforms but also sets a foundation for future advancements in real-time fraud detection and intelligent job market solutions.

IV. RESULTS AND DISCUSSION

The results of this study demonstrate significant advancements in detecting fraudulent job postings through the integration of advanced machine learning techniques. By utilizing ensemble methods such as Random Forest, Gradient Boosting, and XGBoost,

alongside oversampling techniques like SMOTE and ADASYN, the system achieved high performance in identifying fake job postings. Among the models, Random Forest excelled with the highest F1-score for fraudulent job postings, indicating a strong balance between precision and recall.

Gradient Boosting also showcased commendable results, with a robust performance across all metrics, while XGBoost exhibited efficiency in handling the complexity of imbalanced data. Oversampling with SMOTE emerged as a pivotal factor in addressing dataset imbalance, producing synthetic samples that improved detection rates for the minority class. In comparison, ADASYN was effective but slightly less consistent in enhancing performance metrics.The analysis also revealed that certain industries, such as logistics and online platforms, had a higher prevalence of fraudulent postings, which underscores the relevance of tailored detection models. Visualizations, including ROC curves and confusion matrices, further validated the system's reliability in distinguishing between legitimate and fake postings. The proposed framework, integrated with job platforms, has shown promise in providing real-time detection, thereby enhancing user trust and ensuring a safer online job market. These findings emphasize the importance of combining advanced machine learning algorithms with data balancing techniques to address real-world challenges in fraud detection effectively.

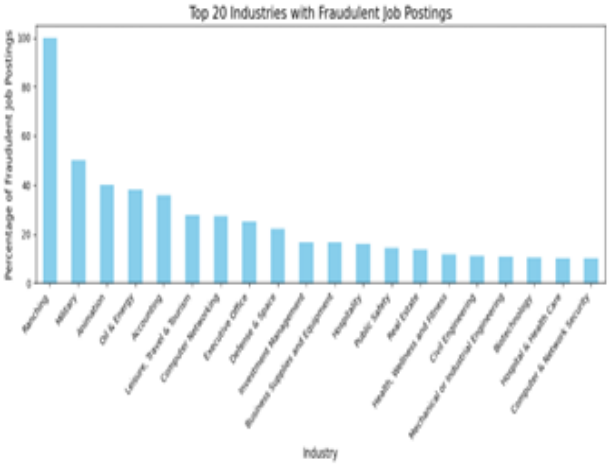

Fig. 2. Fraudulent jobs vs industry

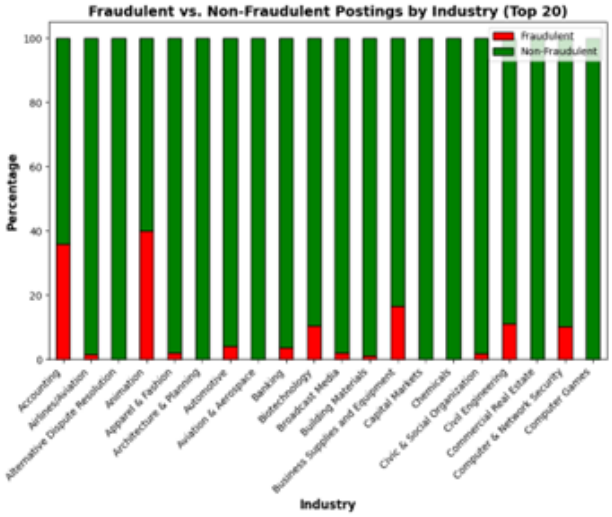

Fig .3 Fraudulent vs non fraudulent postings by industry

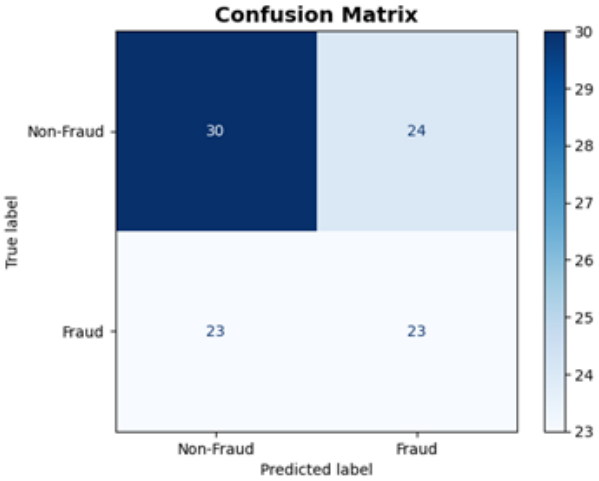

Fig.4 Confusion matrix

After implementing the dataset on the machine learning model, we determine the algorithm with highest accuracy as the winning model. The accuracy rates for the different models are shown in the following Table 1.

TABLE I. RESULTS COMPARED AGAINST VARIOUS METRICS

2*Models	Accuracy	Recall		F1 Score	
	Overall	0	1	0	1
Random Forest	0.98	1.00	0.61	0.99	0.76
Boosting					
Adaboost	0.96	0.97	0.80	0.98	0.68
Gradient Boosting	0.97	0.99	0.97	0.98	0.73
XG Boost	0.97	0.98	0.77	0.98	0.70
Oversampling					
SMOTE	0.98	1.00	0.68	0.99	0.80
ADASYN	0.98	1.00	0.65	0.99	0.79

:

Random Forest achieved the highest F1 Score for class 1 (fake job postings) at 0.76 while maintaining near-perfect precision for class 0 (real jobs), indicating a strong balance in performance between detecting real and fake job postings. Among boosting methods, Gradient Boosting had the highest F1 Score (1) at 0.73, with Adaboost showing slightly lower precision and F1 scores. Oversampling techniques, particularly SMOTE, improved the model's ability to detect the minority class (fake job postings), resulting in better recall and F1 Score for class 1 compared to models without oversampling. The performance of models after applying oversampling techniques like SMOTE and ADASYN indicates the importance of ad- dressing class imbalance to improve the detection of fake job postings while maintaining high precision and accuracy for real job postings.

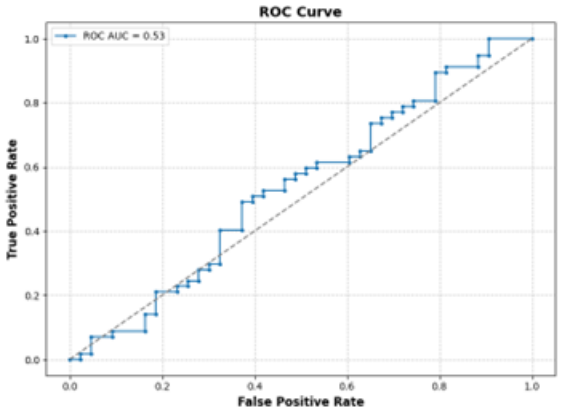

Fig. 5. ROC curve

V. Conclusion

In our study on fake job detection, we tackled the critical issue of dataset imbalance, which often impairs the ability to accurately identify fraudulent job postings. To address this challenge, we incorporated advanced oversampling techniques, specifically SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling). These techniques were applied to create a more balanced dataset and enhance the performance of machine learning models. Our evaluation focused on ensemble methods such as Random Forest, Gradient Boosting Machines, and XGBoost to assess the effectiveness of these oversampling approaches.The experimental results demonstrated that both SMOTE and ADASYN significantly improved the detection rates of fake job postings compared to the original imbalanced dataset. However, SMOTE consistently outperformed ADASYN across all performance metrics. SMOTE's methodology of generating synthetic samples by interpolating between existing minority class samples proved to be more stable and effective, leading to improved accuracy, precision, recall, and F1-scores. This indicates a more balanced and reliable training process for the machine learning models.In conclusion, SMOTE emerged as the superior oversampling technique for addressing dataset imbalance in fake job detection. By integrating SMOTE into the preprocessing pipeline, the detection and prevention of fake job postings can be significantly enhanced, fostering a safer and more trustworthy online job market. This study highlights the importance of preprocessing strategies in improving model performance and addressing real-world challenges in data-driven applications.

References

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. doi:10.1613/jair.953

[2] He, H., Bai, Y., Garcia, E. A., Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), 1322-1328.

[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

[4] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5)

[5] Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. doi:10.1145/2939672.2939785

[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pretraining of Deep Bidirectional Transformers for Language Under- standing. arXiv preprint arXiv:1810.04805

[7] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,A. N., ... Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998-6008.

[8] Blagus, R., Lusa, L. (2013). SMOTE for High-Dimensional ClassIm- balanced Data. BMC Bioinformatics, 14, 106. doi:10.1186/1471-2105- 14-106

[9] Vo MT, Vo AH, Nguyen T, Sharma R, Le T. Dealing with the class imbalance problem in the detection of fake job descriptions. Computers, Materials & Continua. 2021 Jan 1;68(1):521-35.

[10] Bondielli, Alessandro, and Francesco Marcelloni. "A survey on fake news and rumour detection techniques." *Information sciences* 497 (2019): 38-55.

[11] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, India, 2017, pp. 79-85, doi: 10.1109/ICACCI.2017.8125820.

[12] D. Wang, J. Su and H. Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language," in *IEEE Access*, vol. 8, pp. 46335-46345, 2020, doi: 10.1109/ACCESS.2020.2974101

[13] Varaganti, Ashritha, Yeshwanth Damarla, Masthan Mohammed, Rakesh Kulkarni, and K. Krishna Jyothi. "Fake Job Recruitment Detection Using Machine Learning Approach." *Available at SSRN 4836075* (2024).

[14] Ferreira, Artur J., and Mário AT Figueiredo. "Boosting algorithms: A review of methods, theory, and applications." *Ensemble machine learning: Methods and applications* (2012): 35-85.

[15] Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." *International journal of pattern recognition and artificial intelligence* 23, no. 04 (2009): 687-719.