INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Vol 14, Issue 1, (2024) E-ISSN: 2222-6990

Machine Learning Approach in Predicting Fraudulent Job Advertisement

Atikah Hanisah Mohd Hanif, Nurazean Maarop, Norshaliza Kamaruddin and Ganthan Narayana Samy

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Jalan Sultan Putra, Kuala Lumpur, 54100, Malaysia

Corresponding Author Email: atikahhanisah@graduate.utm.my

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v14-i1/20532 DOI:10.6007/IJARBSS/v14-i1/20532

Published Date: 12 January 2024

Abstract

As the world population grows, the demand for workers increases, leading to a rise in online job advertisements to connect employers with potential employees on a national scale. However, this shift also brings the risk of falling victim to fraud. Reported commercial crimes in Malaysia saw a 15.3% increase in 2021, with fraud being the highest among them. Several studies have proposed Machine Learning models to classify genuine and fraudulent job advertisements, but the analysis of certain techniques remains limited. The paper aims to develop a predictive model for identifying fraudulent job advertisements using selected features from imbalanced and balanced datasets. The Employment Scam Aegean Dataset was utilized to build Machine Learning classification models using Logistic Regression, Support Vector Machine, Decision Tree, and Naïve Bayes algorithms. These models were combined with different vectorizers like Term Frequency-Inverse Document Frequency, Bag of Words, and Hash. The Decision Tree model with Bag of Words vectorizer on a balanced dataset outperformed other models, achieving an accuracy of 0.705, precision of 0.73, recall of 0.70, F1-score of 0.71, and Area Under Curve score of 0.68. This model shows promise in effectively identifying fraudulent job advertisements, safeguarding job seekers from scams in the online job market.

Keywords: Machine Learning, Predictive Models, Fraudulent Job Advertisements, Online Job Advertisements, Fraudulent Activities

Introduction

As of 2022, the world population has reached an all-time high of 8 billion, and the global internet penetration has steadily increased with the advancement of web and telecommunication technology. Approximately 4901 million people, as of 2021, are now connected to the internet, marking a significant shift in how individuals interact and seek opportunities (ITU, 2022). This exponential growth has led to the emergence of online job-

seeking platforms, enabling employers and recruiters to connect with potential employees on a national and international scale. However, this transformation has also introduced a concerning challenge – the proliferation of fraudulent job advertisements. This phenomenon presents both positive economic growth indicators and potential risks for job seekers. While the rise in job postings fosters economic opportunities, it also exposes applicants to the possibility of falling prey to fraudulent job scams (Vidros et al., 2016).

Fraudulent job advertisements pose significant threats to individuals seeking employment and the overall integrity of job advertising platforms. The impact of these scams is evident in crime statistics, with commercial crimes, especially fraud, showing a 15.3% increase in reported cases from 2020 to 2021 in Malaysia (DOSM, 2022). The consequences of such crimes extend beyond financial losses, affecting job seekers' reputation and personal security. Moreover, neighbouring countries in Southeast Asia, such as Indonesia and Singapore, also face similar challenges with an increasing number of job fraud victims (Idrus, 2022; SPF., 2022). In response, there is a pressing need to develop effective measures to identify and combat fraudulent job advertisements.

This research paper seeks to address the issue of fraudulent job advertisements by proposing a robust Machine Learning (ML) approach to predict and identify such postings accurately. The primary aim of this research is to develop a classification model that effectively distinguishes between genuine and fraudulent job advertisements by exploring various ML algorithms, dataset balancing techniques, and feature extraction methods. The findings from this research have the potential to contribute to the development of advanced fraud detection systems and foster a safer and more trustworthy job market environment for job seekers and employers alike.

Literature Review

Fraudulent in Employment

The literature review on fraudulent acts covers the definition of fraud, types of employment fraud, and strategies for fraud prevention. Fraud is described as the intentional perversion of truth to induce someone to part with something valuable or surrender a legal right (FBI, 2016). The Association of Certified Fraud Examiners (ACFE) further characterizes fraud as intentional behaviour aiming to deprive a person or organization of money or property through deception (ACFE, 2023). Online recruitment fraud is identified as a malicious act that jeopardizes privacy, inflicts economic damage, and undermines organizational credibility (Vidros et al., 2017).

Two types of recruitment fraud are specified in a study by Cross & Grant-Smith (2021): labour trafficking, where individuals are forced into illegal labour arrangements, and theft of confidential information, involving unauthorized access to personal details. The International Labour Organization (ILO) defines forced or compulsory labour as work demanded under threat or penalty, not voluntarily offered by an individual (ILO, 2022). Fraudulent job recruitment often involves luring jobseekers into providing personal information, which is then resold to third parties, leading to spam emails and further fraud (Vidros et al., 2017).

The section also delves into the risk factors associated with employment fraud, providing insights from various studies. Goyal et al (2023) find that email addresses are commonly used to approach victims, with mentions of telecommuting jobs and COVID-19 being frequent among victims. Ravenelle et al (2022) identify early detection strategies for fraudulent job postings, such as requests for personal information, poor wording in the advertisement, jobs with significant benefits, and persistent communication through digital

media platforms. Carmen et al (2010) suggest identifying misspelled words, vague information, and incorrect details in job postings as indicators of fraudulent advertisements. Vidros et al (2016) discover that fake job advertisements often entice applicants with deceitful benefits like high salary, flexible working hours, teleworking, and career growth opportunities.

To combat fraudulent activities, various strategies have been proposed. Abdullah Asuhaimi et al (2017) highlight the importance of addressing false information and exploitation in advertisements but mention a lack of specific provisions for penalties in some countries. Daud (2021) discusses regulatory efforts such as self-regulation and internet co-regulation, involving acts like The Communication and Multimedia Act 1998, Anti Fake News Act 2018, and Emergence Ordinance No.2/2021. The ASEAN Ministers Responsible for Information (AMRI) propose a framework to reduce the negative impact of fake news, focusing on cooperation among member states and promoting awareness and understanding of the issue (ASEAN context, AMRI). By considering these risk factors and implementing preventive measures, governments and individuals can work together to combat employment fraud and reduce its impact.

Handling of Imbalanced Data

The literature review on imbalanced data classification highlights the impact of imbalanced class sizes on model classifiers. Imbalanced datasets, where one class has significantly fewer samples than the other, can lead to misclassifications and biased model performance. Various fields such as medical, insurance, banking, network, and information retrieval tasks are commonly affected by class imbalance (Santhi & Reddy, 2019). Examples of datasets with naturally skewed class distributions include fraud detection data, computer security data, and image recognition data (Johnson & Khoshgoftaar, 2019).

The handling of imbalanced data can be categorized into three groups: data-level methods, algorithm-level methods, and hybrid approaches (Santhi & Reddy, 2019; Niaz et al., 2022; Rekha et al., 2021; Johnson & Khoshgoftaar, 2019). Data-level methods involve modifying the dataset's samples to balance the distribution, achieved through oversampling (generating new samples for the minority class) and undersampling (removing samples from the majority class). Algorithm-level methods aim to increase the importance of the positive class without altering the dataset's distribution. Hybrid approaches combine data-level and algorithm-level methods for a more accurate model.

Several studies have explored imbalanced datasets in the context of different fraudulent cases, utilizing various resampling techniques and evaluation metrics (Rubaidi et al., 2022; Chen et al., 2021; Li et al., 2021; Mrozek et al., 2020; Bauder et al., 2018). Among the techniques used for handling imbalanced data were Random Undersampling (RUS), Random Oversampling (ROS), SMOTE, Borderline-SMOTE, Adaptive Synthetic Sampling (ADASYN), and cost-sensitive learning. Evaluation metrics commonly employed include precision, recall, F1-score, accuracy, and AUC.

After comparing the results of these studies, it was observed that RUS emerged as the most effective technique for handling imbalanced datasets in the context of fraud cases. The use of RUS addressed the class imbalance issue by reducing the instances from the majority class, improving classification performance, and reducing bias towards the majority class. Other techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning also showed promising results in specific scenarios.

Machine Learning in Prediction of Fraudulent Cases

This subsection presents a literature review and comparison of various research studies conducted between 2018 and 2022 that focus on the application of ML in detecting and predicting fraudulent activities.

ML encompasses three main learning types: supervised learning, unsupervised learning, and reinforcement learning (Díaz-Ramírez, 2021). However, some studies expanded this categorization to include semi-supervised learning (Sarker, 2021; Maleki et al., 2020). Supervised learning involves models trained on labeled data for tasks like classification and regression. Unsupervised learning discovers patterns in unlabeled data, and semi-supervised learning combines labeled and unlabeled data. Reinforcement learning deals with agents that learn by improving their actions in an environment (Sarker, 2021; Maleki et al., 2020).

Numerous supervised learning algorithms have been applied across different domains. Some popular methods include Linear and Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), and Neural Networks (Maleki et al., 2020). Each of these methods has its strengths and limitations, making them suitable for various classification and regression tasks.

The evaluation of ML models is crucial for assessing their performance. Commonly used metrics include accuracy, error rate, sensitivity, specificity, precision, recall, F-measure, and Geometric-mean (M & M.N, 2015; Tharwat, 2018). The area under the curve (AUC) is often used for ROC curve comparison.

Several studies have employed ML to detect fraudulent activities across different contexts, including fake news detection, credit fraud detection, online recruitment fraud detection, and general prediction of fraud victimization (Agarwal et al., 2022; Gupta et al., 2022; Tran & Dang, 2021; Tabassum et al., 2021; Lokanan & Liu, 2021). Various supervised classification models, such as LR, DT, RF, KNN, and GradientBoosting, have been explored in these studies.

Each study utilized different performance metrics to evaluate the ML models, with precision being the most commonly used metric. Findings from these studies identified different best-performing models for predicting fraudulent activities, including Multinomial NB, LR, DT, RF, KNN, LightGBM, and GradientBoosting.

Analysis on the Study of the Employment Scam Aegean Dataset (EMSCAD)

Several studies have been conducted on fraudulent job advertisement prediction using the EMSCAD dataset. The EMSCAD dataset comprises 17,800 job listings published on the Workable platform from 2012 to 2014. Of these job postings, 866 were manually identified as fraudulent, with the determination based on various factors, including suspicious client behavior, falsified contact or company details, candidate grievances, and regular client analysis (Vidros et al., 2017). These listings encompass of 18 attributes which are title, location, department, salary_range, company_profile, description, requirements, benefits, telecommuting, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent and in_balanced_dataset. The attributes present are categorized into four types: string, HTML fragment, binary, and nominal attributes.

The dataset's target classification is the fraudulent attribute, where 't' denotes fake job advertisements and 'f' denotes real job advertisements. Around 866 of the 17,880 job advertisements in the dataset, or 4.8% of the total, have been determined to be fraudulent; the vast majority, or 17,014, or 95.2%, are real employment advertisements. The creation of

efficient predictive models for fraud detection is facilitated by this attribute's critical role in identifying legitimate and fraudulent job posts within the dataset.

Vidros et al (2017) focused on online recruitment fraud, using a balanced dataset and RF as the best-performing classifier with approximately 91% precision, recall, and F-measure. Lal et al (2019) proposed an ORF detection model based on ensemble learning, outperforming baseline classifiers with an average accuracy of 94% despite working with an imbalanced dataset. The model also exhibited commendable precision, recall, and F1-Score values of 94.9%, 95.6%, and 94%, respectively.

Mehboob & Malik (2021) utilized balanced dataset; introduced new features and found XGBoost to be the most effective classifier with 97.94% accuracy and recall also 98% precision and F1-score. Habiba et al (2021) compared various data mining techniques, with RF achieving the highest accuracy of 96.7% for traditional ML algorithms and Deep Neural Network (DNN) reaching 99% accuracy for deep learning. Nasser et al (2021) used an Artificial Neural Network (ANN)-based model with accuracy, precision, recall, and F-measure of 93.64%, 91.84%, 96.02%, and 93.88%, respectively. Lokku (2021) employed RF achieving 99% precision, recall, accuracy, and F1-score.

Nessa et al (2022) worked with imbalanced dataset and presented a Gated Recurrent Unit (GRU) model with a ROC-AUC score of 93.51%. Amaar et al. (2022) proposed an NLP and ML-based methodology with the Extra Trees Classifier (ETC) using TF-IDF and ADASYN sampling achieving 99.9% accuracy.

In summary, these studies primarily focus on finding the best performing ML algorithm for detecting fake job advertisements. Commonly used classifiers include RF, SVM, NB, and Neural Networks. The evaluation metrics include precision, recall, F1-measure, accuracy, AUC, sensitivity, and specificity. The research gap identified in the literature review suggests further exploration into the performance of the dataset using RUS with TF-IDF, BoW, and Hash as feature extraction techniques.

Table 1 provides the summary of results from various studies conducted using the EMSCAD data.

Table 1
Results of the ML Model Performance using EMSCAD Data

Study	Data	Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Vidros et al. (2017)	Balanced dataset of 900 job ads	RF	91.2%	91.4%	91.2%	91.4%	97%
Lal et al. (2019)	Imbalanced	Ensemble Learning	95.4%	94.9%	95.6%	94%	N/A
Mehboob & Malik (2021)	Balanced dataset of 940 job ads	XGBoost	97.94%	98%	97.9%	98%	N/A
Habiba et al. (2021)	N/A	RF	96.7%	93%	95%	93%	N/A
		DNN	98%	97%	97%	N/A	N/A
Nasser et al. (2021)	Down-sampling	ANN	93.64%	91.84%	96.02%	93.88%	N/A
Lokku (2021)	SMOTE sampling	RF	99%	99%	99%	99%	N/A
Nessa et al. (2022)	Imbalanced	GRU	N/A	N/A	N/A	N/A	93.51%
Amaar et al. (2022)	ADASYN sampling	ETC	99.9%	99.9%	99.9%	99.9%	N/A

Research Methodology

The research methodology involves the development and evaluation of a prediction model for classifying fraudulent job advertisements using ML techniques. The research design is based on a supervised learning approach, where four classification models—LR, SVM, DT, and NB—are utilized. The dataset used for training and testing the models is the EMSCAD data which is publicly available on the official website of the University of The Aegean. The research aims to identify the most effective model for accurately classifying fraudulent job advertisements and preventing individuals from falling victim to job scams.

Since the dataset is a readily available, hence no specific data collection method is required. The data preprocessing steps involve several key tasks to ensure data quality and suitability for ML modeling. These tasks include removing irrelevant and duplicate data, handling missing values, and cleaning nominal and text data. For nominal data cleaning, label encoding is employed to handle categorical variables. Text cleaning techniques are applied to process textual data, which includes removing HTML tags and URLs, normalizing text to lowercase, removing punctuation, stopwords, numerical values, and stemming to reduce words to their root form.

In the feature engineering phase, the textual data is transformed into a numerical representation to be used in ML algorithms. Three different vectorization techniques are explored for this purpose: BoW, TF-IDF, and Hash vectorizer. Each technique is evaluated to determine its effectiveness in capturing relevant information from the text. Additionally, due to the imbalanced class distribution in the dataset, the research addresses this issue by using the RUS technique for resampling. Both the original imbalanced dataset and the balanced dataset obtained through resampling are used to fit the classification models. The

performance of the models is evaluated using various performance measures such as confusion matrix, accuracy, precision, recall, F1-score, and AUC.

The models—LR, SVM, DT, and NB are trained separately using both the balanced and imbalanced datasets. The Python tool is employed to implement the entire project, including data preprocessing, feature engineering, model training, and performance evaluation. The evaluation process aims to identify the best performing model for predicting fraudulent job advertisements based on its accuracy and other performance metrics. The research rigorously assesses the models' performance to ensure the selected model is robust and capable of accurately classifying fraudulent job advertisements.

Proposed Conceptual Model/Framework

The Proposed solution for this paper is visually represented as shown in Figure 1.

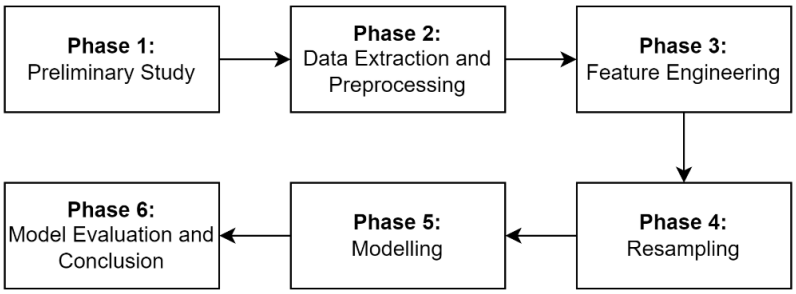

Figure 1. Work Process of Proposed Solution for Predicting Fraudulent Job Advertisements

The conceptual model illustrates the step-by-step flow of the work process. It starts with the preliminary study where problem background and literature review were conducted to identify the gap. In this phase, research questions and objectives were developed, and a literature review was conducted to explore existing publications on fraudulent activities, ML, and imbalanced data, providing insights to resampling techniques, ML approaches, and evaluation methods.

In the data extraction and pre-processing phase, data was extracted from The Aegean University website, with a specific emphasis on the EMSCAD data. Missing values were addressed by removing columns with a missing value percentage of 70% or more, replacing missing values with empty spaces, and performing extensive cleaning of 'text' data, including tasks like HTML and URL removal, lowercase normalization, punctuation removal, stop words removal, numerical removal, and stemming. The cleaned and processed data then moves on to the feature engineering phase, where three vectorization techniques—BoW, TF-IDF, and Hash vectorizer—are applied to represent the text data.

Then, the research addresses the issue of class imbalance through RUS, which creates a balanced dataset. This approach created a balanced EMSCAD dataset while mitigating bias in the data. The balanced dataset is then used to train and test the four classification models—LR, SVM, DT, and NB—during the model training and evaluation phase. These models were designed to predict fraudulent job advertisements based on the features extracted from the dataset. The performance of each model is assessed using various performance measures.
Finally, model evaluation was conducted using confusion matrices, and model performance was measured using various metrics, including accuracy, precision, recall, F1-score, and AUC. The performance of all four models was compared using these metrics, and the research concludes by identifying the best performing model for predicting fraudulent job advertisements based on the evaluation results.

Results and Discussions

Model Performance on Imbalanced Dataset

Table 2 provides the results of the ML model performance the imbalanced dataset based on different performance metrics.

Table 2
Results of the ML Model Performance the Imbalanced Dataset

Vectorizer	ML Algorithm	Evaluation Matrix				
		Accuracy	Precision	Recall	F1-Score	AUC
TF-IDF	LR	0.970	0.00	0.00	0.00	0.70
	SVM	0.970	0.00	0.00	0.00	0.68
	DT	0.952	0.25	0.32	0.28	0.66
	NB	0.875	0.05	0.17	0.07	0.54
BoW	LR	0.970	0.00	0.00	0.00	0.68
	SVM	0.970	0.00	0.00	0.00	0.66
	DT	***0.948***	***0.23***	***0.37***	***0.31***	***0.66***
	NB	0.856	0.03	0.12	0.05	0.54
Hash	LR	0.970	0.00	0.00	0.00	0.69
	SVM	0.970	0.00	0.00	0.00	0.65
	DT	0.954	0.26	0.30	0.28	0.61
	NB	0.911	0.03	0.07	0.05	0.61

The DT model with BoW technique emerged as the top performer in multiple metrics, including accuracy, precision (class 1), recall (class 1), F1-score (class 1), and AUC-score. This model demonstrated promising performance in accurately classifying fraudulent job advertisements.

Additionally, other models, such as LR and SVM with various vectorization techniques, also showed good accuracy and performance, but the DT model with BoW outperformed them in most metrics. The DT model with Hashing technique displayed high precision (class 1), indicating its ability to accurately identify instances of fraudulent job advertisements. The LR model with TF-IDF technique achieved the highest AUC-score, demonstrating its capability to distinguish between fraudulent and non-fraudulent job advertisements effectively.

Model Performance on Balanced Dataset

Table 3 presents the results of the ML model performance on the balanced dataset based on different performance metrics.

Table 3
Results of the ML Model Performance the Balanced Dataset

Vectorizer	ML Algorithm	Evaluation Matrix				
		Accuracy	Precision	Recall	F1-Score	AUC
TF-IDF	LR	0.664	0.69	0.64	0.67	0.71
	SVM	0.647	0.68	0.62	0.65	0.66
	DT	0.651	0.65	0.71	0.68	0.65
	NB	0.548	0.58	0.52	0.54	0.57
BoW	LR	0.635	0.66	0.63	0.65	0.64
	SVM	0.631	0.66	0.61	0.63	0.68
	DT	***0.705***	***0.73***	***0.70***	***0.71***	***0.68***
	NB	0.589	0.62	0.55	0.58	0.57
Hash	LR	0.664	0.70	0.63	0.66	0.71
	SVM	0.635	0.68	0.56	0.62	0.66
	DT	0.647	0.67	0.70	0.68	0.65
	NB	0.564	0.60	0.48	0.54	0.59

For the balanced dataset, the LR model with TF-IDF technique emerged as the top performer in accuracy, demonstrating a high level of correctness in classifying both fraudulent and non-fraudulent job advertisements. The DT model with Hashing technique outperformed other models in precision (class 1), showcasing its effectiveness in accurately identifying fraudulent job advertisements.

Moreover, the DT model with TF-IDF technique achieved the highest recall (class 1) score, indicating its ability to capture a higher proportion of fraudulent job advertisements. This model also attained the highest F1-score for class 1, representing a well-balanced performance between precision and recall. In terms of AUC-score, both the LR model with TF-IDF technique and the LR model with Hashing technique demonstrated the highest discrimination power in distinguishing between fraudulent and non-fraudulent job advertisements in the balanced dataset.

Comparison of Results between Imbalanced and Balanced Dataset

Comparing the DT model with BoW results between the imbalanced and balanced datasets, few differences were observed. Figure 2 visualized the comparison of the model performances.

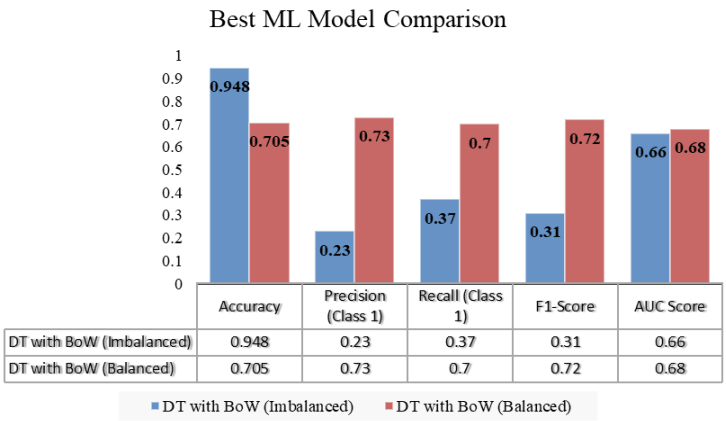

Figure 2. Comparison of Model Performance

The model achieved higher accuracy in the imbalanced dataset, indicating better overall class label predictions. However, in the balanced dataset, the model showed higher precision, recall, F1-score, and AUC-score, indicating its effectiveness in identifying fraudulent job advertisements while minimizing false positives.

This paper's results demonstrate the effectiveness of ML models in predicting fraudulent job advertisements. The DT model with BoW technique consistently emerged as the best performer in multiple metrics for both imbalanced and balanced datasets. This model's ability to accurately classify fraudulent job ads indicates its potential in providing a reliable and efficient fraud detection solution for job advertisement platforms, recruitment agencies, and job seekers.

The comparison between the imbalanced and balanced datasets showed that the model's performance was influenced by the dataset's balance. While the imbalanced dataset yielded higher accuracy, the balanced dataset led to better precision, recall, F1-score, and AUC-score. This highlights the importance of dataset balance in achieving optimal model performance for fraud detection tasks.

Conclusions

In conclusion, this paper successfully achieved its objectives by developing a robust prediction model and offering valuable insights into fraud detection. The findings hold significant implications for job advertisement platforms, recruitment agencies, and job seekers, as they contribute to the ongoing efforts in combating fraudulent job advertisements and ensuring job search safety.

The main contribution of this research lies in the development and evaluation of an effective prediction model for fraudulent job advertisement classification. The DT model with BoW vectorizer on a balanced dataset emerged as the most effective solution, displaying superior performance in accurately identifying fraudulent job ads. These findings not only benefit job advertisement fraud detection but also serve as a valuable resource for researchers and practitioners working on fraud detection and classification tasks in diverse domains.

Looking ahead, future research can further enrich this field by exploring advanced text processing techniques, alternative resampling methods, and ensemble methods to enhance the prediction model's accuracy and generalization capabilities. Additionally, developing context-specific prediction models tailored to different job markets, languages, and geographic regions could lead to more targeted and accurate fraud detection approaches.

Acknowledgments

We would like to thank Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, and Majlis Amanah Rakyat Malaysia.

References

Abdullah Asuhaimi, F., Pauzai, A. N., Yusob, L. M., & Asari, K.-N. (2017). *Rules on advertisement in Malaysia. World Applied Sciences Journal*, 35(9), 1723–1729.

ACFE. (2023). *Fraud 101: What is Fraud? Association of Certified Fraud Examiner*.

Agarwal, P., Reddivari, S., & Reddivari, K. (2022). *Fake news detection: An investigation based on machine learning. Proceedings - 2022 IEEE 23rd International Conference on Information Reuse and Integration for Data Science*, IRI 2022, 61–62.

Amaar, A., Aljedaani, W., Rustam, F., Ullah, S., Rupapara, V., & Ludi, S. (2022*). Detection of fake job postings by utilizing machine learning and natural language processing approaches. Neural Processing Letters*, 54(3), 2219–2247.

Bauder, R. A., Khoshgoftaar, T. M., & Hasanin, T. (2018). *Data sampling approaches with severely imbalanced big data for medicare fraud detection. Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2018-Novem, 137–142.

Carmen, Glover; Janet, N. (2010). How to Use the Internet to Get Your Next Job.

Chen, Y. R., Leu, J. S., Huang, S. A., Wang, J. T., & Takada, J. I. (2021). *Predicting default risk on peer-to-peer lending imbalanced datasets. IEEE Access*, 9, 73103–73109.

Cross, C., & Grant-Smith, D. (2021). *Recruitment Fraud: Increased opportunities for exploitation in times of uncertainty*? 40(4), 9–14.

Daud, M. (2021). *Freedom of misinformation and the relevance of co-regulation in malaysia: a cross-jurisdictional analysis. IIUM Law Journal*, 29(2), 27–54.

DOSM. (2022). *Big Data Analytics Job Market Insights and My Job Profile: Job Vacancies Landscape in Malaysia, Third and Fourth Quarter of 2021 Job [Media Statement]* (pp. 1–4).

FBI. (2016). Crime in United States: Offense Definitions.

Goyal, N., Mamidi, R., Sachdeva, N., & Kumaraguru, P. (2023). *Warning: It's a scam!! Towards understanding the employment scams using knowledge graphs. ACM International Conference Proceeding Series*, 303–304.

Gupta, V., Mathur, R. S., Bansal, T., & Goyal, A. (2022). *Fake news detection using machine learning. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, May, 84–89.

Habiba, S. U., Islam, M. K., & Tasnim, F. (2021). *A comparative study on fake job post prediction using different data mining techniques. International Conference on Robotics, Electrical and Signal Processing Techniques*, 543–546.

Idrus, P. G. (2022). *Indonesia to increase supervision to stop citizens from being trafficked to Cambodia—BenarNews. BenarNews.*

International Labour Organization (ILO). (2022). Global Estimates of Modern Slavery: Forced Labour and Forced Marriage.

ITU. (2022). Number of internet users worldwide from 2005 to 2021 (in millions) [Graph].

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1–54.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). *A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Computing Surveys*, 52(4), 1–36.

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019). *ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection. 2019 12th International Conference on Contemporary Computing*, IC3 2019.

Lokanan, M., & Liu, S. (2021). *Predicting fraud victimization using classical machine learning. Entropy*, 23(300), 1–19.

Lokku, C. (2021). *Classification of Genuinity in job posting using machine learning. International Journal for Research in Applied Science and Engineering Technology*, 9(12), 1569–1575.

Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., & Forghani, R. (2020). *Overview of machine learning Part 1: Fundamentals and classic approaches. Neuroimaging Clinics of North America*, 30(4), 17–32.

Mehboob, A., & Malik, M. S. I. (2021). *Smart fraud detection framework for job recruitments. Arabian Journal for Science and Engineering*, 46(4), 3067–3078.

Mrozek, P., Panneerselvam, J., & Bagdasar, O. (2020). *Efficient resampling for fraud detection during anonymised credit card transactions with unbalanced datasets. Proceedings - 2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing, UCC 2020*, 426–433.

Nasser, I. M., Alzaanin, A. H., & Maghari, A. Y. (2021). *Online Recruitment Fraud Detection using ANN. Proceedings - 2021 Palestinian International Conference on Information and Communication Technology, PICICT 2021*, 13–17.

Nessa, I., Zabin, B., Faruk, K. O., Rahman, A., Nahar, K., Iqbal, S., Hossain, M. S., Mehedi, M. H. K., & Rasel, A. A. (2022). *Recruitment Scam Detection Using Gated Recurrent Unit. 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC) 2022*, 445–449.

Niaz, N. U., Shahariar, K. M. N., & Patwary, M. J. A. (2022). *Class imbalance problems in machine learning: A review of methods and future challenges. ACM International Conference Proceeding Series*, 485–490.

Ravenelle, A. J., Janko, E., & Kowalski, K. C. (2022). *Good jobs, scam jobs: Detecting, normalizing, and internalizing online job scams during the COVID-19 pandemic. New Media and Society*, 24(7), 1591–1610.

Rekha, G., Tyagi, A. K., Sreenath, N., & Mishra, S. (2021). *Class Imbalanced Data: Open Issues and Future Research Directions. 2021 International Conference on Computer Communication and Informatics, ICCCI 2021*.

Rubaidi, Z. S., Ammar, B. Ben, & Aouicha, M. Ben. (2022). *Fraud detection using large-scale imbalance dataset. International Journal on Artificial Intelligence Tools*, 31(8), 1–23.

Santhi, K., & Rama Mohan Reddy, A. (2019). *A systematic methodology on class imbalanced problems involved in the classification of real-world datasets. International Journal of Recent Technology and Engineering*, 8(3), 7071–7081.

Sarker, I. H. (2021). *Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Computer Science*, 2(3), 1–21.

SPF. (2022). Leading types of scams in Singapore in 2021, by number of cases [Graph] (p. [Online]).

Tabassum, H., Ghosh, G., Atika, A., & Chakrabarty, A. (2021). *Detecting Online Recruitment Fraud Using Machine Learning. 2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 472–477.

Tharwat, A. (2018). *Classification assessment methods. Applied Computing and Informatics*, 17(1), 168–192.

Tran, T. C., & Dang, T. K. (2021). *Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection. Proceedings of the 2021 15th International Conference on Ubiquitous Information Management and Communication, IMCOM 2021*.

Vidros, S., Kolias, C., & Kambourakis, G. (2016). *Online recruitment services: Another playground for fraudsters. Computer Fraud & Security*, 2016(3), 8–13.

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). *Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset*. Future Internet, 9(1), 1–19.