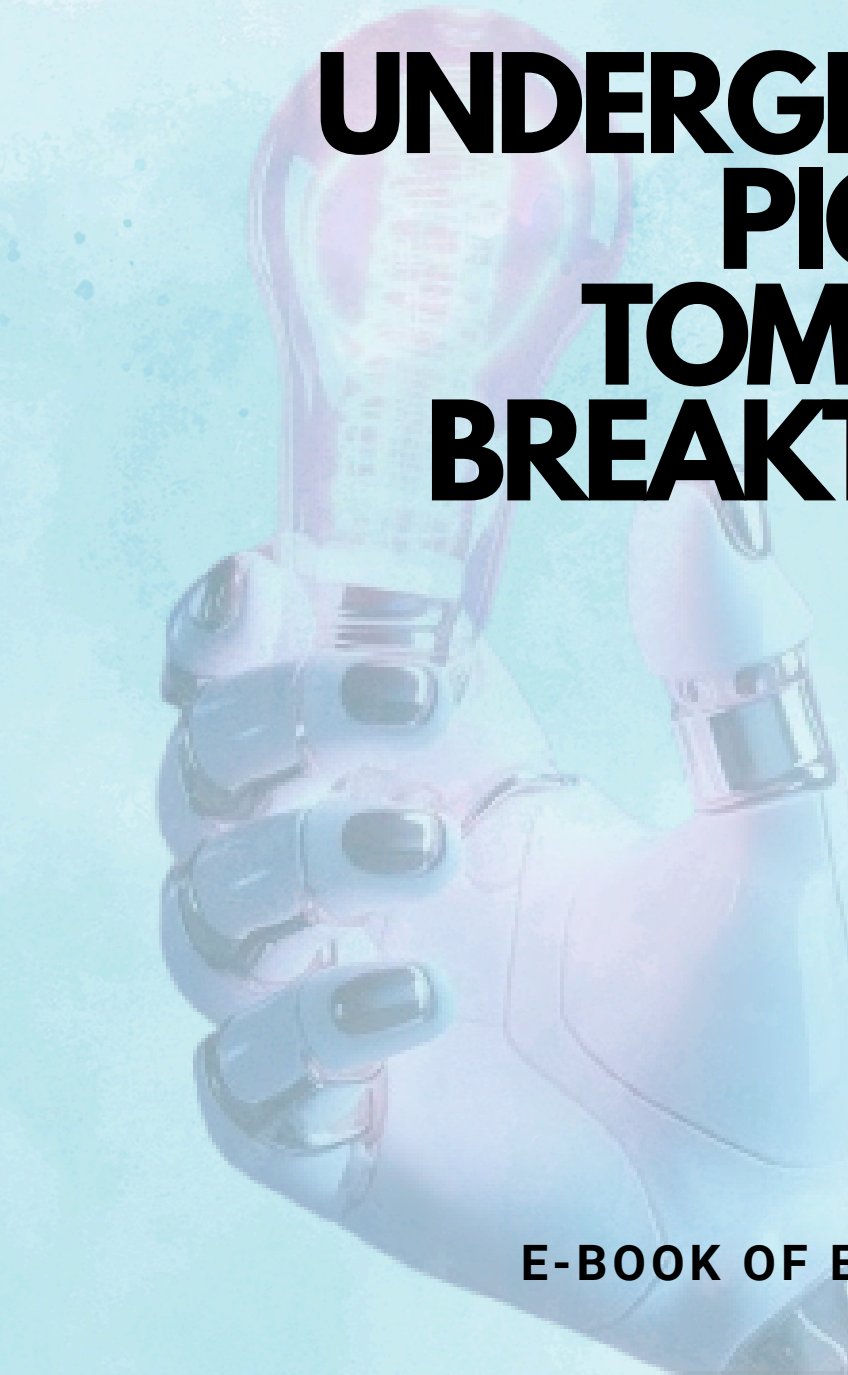
UNIVERSITI TEKNOLOGI MARA

I-URID 2025

INTERNATIONAL UNDERGRADUATE RESEARCH INNOVATION, INVENTION AND DESIGN

UNDERGRADUATES PIONEERING TOMORROW'S BREAKTHROUGH



E-BOOK OF EXTENDED ABSTRACTS

A PROTOTYPE OF FAKE JOB POSTING PREDICTION SYSTEM USING LOGISTIC REGRESSION

Alisa Sofia binti Mohd Nuhairi[1], Nor Azlina binti Aziz Fadzillah[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kampus Kuala Terengganu, Terengganu, Malaysia
[2] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Kampus Seremban, Negeri Sembilan, Malaysia

*norazlina@uitm.edu.my

ABSTRACT

Online job platforms make job searching efficient by offering broader opportunities. However, scammers use these platforms to deceive job seekers by creating fake job postings. Fraudsters use complex tactics that make it hard to differentiate between legitimate and fraudulent posts. It can be harmful for job seekers as it risks their privacy and also financial. The objectives of this system is to develop a prototype of fake job posting prediction using logistic regression and to evaluate the performance of the logistic regression algorithm. This Fake Job Posting Prediction System uses a Logistic Regression trained on a Employment Scam Aegean Dataset(EMSCAD) of job postings. Key preprocessing steps include handling missing values, combining important text fields, encoding categorical variables, and applying TF-IDF vectorization. SMOTE was used to balance the dataset, and GridSearchCV was applied for hyperparameter tuning. The model achieved an accuracy of 97.83%, recall of 79.49%, precision of 76.54% and an F1-score of 77.99% using a 90/10 train-test split with SMOTE and tuning applied. It successfully predicts fake job postings and provides keyword-based explanations to support user understanding. The proposed system successfully predicts fake job postings using a transparent and user-friendly interface.

Keywords: Fake Job Posting, Logistic Regression, TF-IDF Vectorization, SMOTE, GridSearchCV

INTRODUCTION

Nowadays, everything is done online, from schooling to buying. Technology has advanced significantly (Reddy, 2023). In order to reach a wider audience, businesses are now using web technologies such as employment portals (Wang et al., 2024). Online job platforms like LinkedIn and JobStreet allow employers to post open positions and connect with candidates efficiently. The amount of job information available on the internet is rapidly increasing. Employers and job seekers now primarily rely on online job portals and websites to find suitable employment possibilities. (Wang et al., 2024). Job seekers have numerous options available to them. Applying for jobs online is convenient and saves time. People received the benefits, however there are a few people out there that post fraudulent job openings and attempt to swindle job searchers (Reddy, 2023).

The increasing number of scam online has become serious issues and impacting the trust and safety in the job market. Job seekers face significant challenges in distinguishing between genuine and fraudulent job postings, which can expose them to being victim to scams, resulting in the loss of their personal information, financial resources, or in some cases, even their existing employment. Employment scams are one of the main issues with online recruitment fraud (ORF) that have received a lot of attention lately. These days, a lot of businesses have decided to post job positions online so that candidates can find them easily. However, this can be one of the scammers' goals since they offer positions to applicants in return for their money. It is possible to harm a reputable company's reputation by publishing false job openings (Boka, 2024). Scammers frequently use false company websites, clone bank websites, and clone official-looking papers to attract job seekers. Job scammers typically use email instead of direct contact (Habiba et al., 2021). Not all people are fully aware of what is correct and what is not. Some people fall victim to fraudulent legitimate businesses that provide employment opportunities (Reddy, 2023).

People are unaware that the jobs that are posted may be genuine or fraudulent (Srivedi, 2024). These scammers usually target platforms like Linkedln to prove themselves as recruitment agencies. One can quickly find a job that fits their qualifications and desired field by searching the internet (Srivedi, 2024). Usually fraudster will try to represent their company profile or websites to the job seekers as realistic as possible (Habiba et al., 2021). The fraudsters also use complex tactics to lure people into the scam, making it impossible for them to tell the difference between real and false employment adverts (Akram et al., 2024) .The consequences of this is individuals may not be interested in new job advertisements due to concerns about personal, academic, and professional information security (Habiba et al., 2021). It becomes increasingly difficult to keep data and privacy safe online. The number of people who have been victims of fake job postings is constantly increasing. Companies and recruiters attract job seekers in a variety of ways, most of which come from digital job-providing websites (Ranparia et al., 2020).

Job seekers can face serious risks, such as losing their privacy, money, or even their current jobs, due to online recruitment fraud and employment scams. While online recruitment systems have brought convenience and benefits to both job seekers and recruiters, they can also cause harm if not implemented carefully. This issue has been highlighted in studies, showing how vulnerable job seekers are in today′s digital job market (Akram et al., 2024). Furthermore, fraudsters undermine the integrity of well-known organizations by damaging their reputation in the employment market (Akram et al., 2024). By rethinking the approaches to overcome this issue, the proposed solution can helps them to prevent financial losses, pay an application fee to register, or pay for the hiring process (Swathi et al., 2023).

The product developed from this project is the prototype of fake job posting prediction system using logistic regression that serves as the practical solution for the problems mentioned. The system allows user to input job posting details such as employment type, required experience, required education, job function and other job details which are then analyzed through logistic regression model trained on real dataset, Employment Scam Aegean Dataset(EMSCAD). Important features such as job title, description, requirements and categorical information are combined and processed using TF-IDF and encoding techniques. The system not only provide the prediction results such as fake job posting or real job posting, but also highlights the influential keywords that contributed to the result of the prediction. By having these features, it offers transparency and increases user trust in the output. This system is designed to assist job seekers by acting as a decision support tool, helping them to differentiate between genuine job opportunities and potentially harmful job scams.

This project will identify fraudulent job postings using selected prediction model. Prediction model is designed to analyze and classify job listings based on various features that distinguish legitimate opportunities from scams. One of the example of prediction model is Logistic Regression, also known as the logistic model or logit model. It is to evaluates the relationship between several independent factors to categorize dependent variable (Khandagale et al., 2022). Based on this approach the objectives is defined as follow. The first objective is to develop a prototype of fake job posting prediction using logistic regression and the second objective is to evaluate the performance of the logistic regression algorithm.

RESULTS AND DISCUSSION

Table 1 shows the comprasion table results for each splitting set under three comparisons which are vefore smote & tuning, after smote and after smote & tuning.

Table 1: Comparison Table Results for each Splitting Set with 3 Comparisons

Splitting Set	Comparison	Accuracy	Precision	Recall	F1-Score
90/10	Before SMOTE & Tuning	96.65	100	30.77	47.06
90/10	After SMOTE	96.35	58.56	83.33	68.78
90/10	After SMOTE & Tuning	97.83	76.54	79.49	77.99
80/20	Before SMOTE & Tuning	96.68	98.04	32.05	48.31
80/20	After SMOTE	96.43	59.62	81.41	68.83
80/20	After SMOTE & Tuning	97.67	73.97	80.13	76.92
70/10	Before SMOTE & Tuning	96.44	98.53	28.15	43.79
70/10	After SMOTE	96.59	61.66	81.09	70.05
70/10	After SMOTE & Tuning	97.50	73.88	76.05	74.95

This resulted in a total of nine distinct experiments, each evaluated using four standard metrics which are Accuracy, Precision, Recall, and F1 Score. Starting with the 90/10 split, the model achieved an accuracy of 96.65% before any form of balancing or tuning was applied. Interestingly, the precision in this case was perfect 100.00%, indicating that all predictions labeled as fake were indeed correct. However, this came at the cost of very poor recall 30.77%, which means that many actual fake job postings were missed. The overall F1 score was only 47.06%, highlighting that the model, though highly precise, failed to generalize well to the minority class due to severe class imbalance. When SMOTE was applied, the model′s recall significantly improved to 83.33%, but precision dropped to 58.56%, yielding a more balanced F1 score of 68.78%. After tuning the hyperparameters, the model achieved its best performance under the 90/10 setting, with an improved accuracy of 97.83%, precision of 76.54%, recall of 79.49%, and an F1 score of 77.99%, a considerable improvement in both detection quality and consistency.

Under the 80/20 split, a similar trend was observed. Initially, the model showed a high accuracy of 96.68% and precision of 98.04%, but recall was only 32.05%, resulting in a low F1 score of 48.31%. This confirmed once again that while the model was very cautious in predicting fake postings, hence the high precision, it failed to predict a significant portion of them. After SMOTE was applied, recall dramatically increased to 81.41%, and although precision dropped to 59.62%, the F1 score improved to 68.83%, offering better balance. With additional hyperparameter tuning, the model reached its best condition with accuracy of 97.67%, precision of 73.97%, recall of 80.13%, and F1 score of 76.92%, showing slightly better recall and stability than the 90/10 version.

The 70/30 split, having the largest proportion of test data, posed a greater challenge. Initially, the model scored accuracy of 96.44% and an extremely high precision of 98.53%, but with a very poor recall of 28.15%, the F1 score was limited to 43.79%. This imbalance is typical in imbalanced datasets without proper resampling. When SMOTE was introduced, recall surged to 81.09%, precision balanced at 61.66%, and F1 score reached 70.05%, suggesting a more practical model for real use. Finally, after tuning, the model achieved accuracy of 97.50%, precision of 73.88%, recall of 76.05%, and F1 score of 74.95%, showing robust and consistent results comparable to those of the 80/20 configuration.

Finally, the comparison revealed that using SMOTE significantly improved the model's recall, allowing it to spot more fake job posts that were previously overlooked. The most effective results were achieved when SMOTE was combined with hyperparameter tuning, providing a good balance between precision and recall. Among all examined train-test splits, the 90/10 configuration with SMOTE and tuning performed the best, with an F1 score of 77.99% and an overall accuracy of 97.83%.

CONCLUSION

This project successfully achieved all the objectives outlined at the beginning of the study. The first objective, to develop a prototype of fake job posting prediction using logistic regression, was fulfilled by building prediction model implemented in Python. The model is capable of predicting whether a job posting is real or fake and displays explainable output by highlighting influential keywords in the prediction. The second objective, to to evaluate the performance of the logistic regression algorithm, was successfully addressed through a series of experiments using three different train-test splits: 90/10, 80/20, and 70/30. Evaluation metrics such as accuracy, precision, recall, and F1-score were applied. The best result was achieved using the 90/10 split with SMOTE and hyperparameter tuning, reaching 97.83% accuracy and an

F1-score of 77.99%. This confirmed that the model performs well, especially in predicting fake job postings while minimizing misclassification of real ones. This project has successfully designed, implemented, and evaluated a fake job posting prediction system using logistic regression. It provides a strong foundation to protect job seekers from scams.

REFERENCES

Akram, N., Irfan, R., Al-Shamayleh, A. S., Kousar, A., Qaddos, A., Imran, M., & Akhunzada, A. (2024). Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches. *IEEE Access*, *4*, 1–22. https://doi.org/10.1109/ACCESS.2024.3435670

Boka, M. (2024). Predicting Fake Job Posts Using Machine Learning Models. *SSRN Electronic Journal*, *April*. https://doi.org/10.2139/ssrn.4807145

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. https://doi.org/10.1109/ICREST51555.2021.9331230

K. Sridevi, G. Likitha, P. Chandana, & Shrutika Shamarthi. (2024). Real or Fake Job Posting Detection. *International Research Journal on Advanced Engineering and Management (IRJAEM)*, *2*(8 August 2024), 2499–2504. https://doi.org/10.47392/irjaem.2024.0361

Khandagale, P., Utekar, A., Dhonde, A., & Karve, P. S. S. (2022). Fake Job Detection Using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology*, *10*(4), 1822–1827. https://doi.org/10.22214/ijraset.2022.41641

Ranparia, D., Kumari, S., & Sahani, A. (2020). Fake Job Prediction using Sequential Network. *2020 IEEE 15th International Conference on Industrial and Information Systems, ICIIS 2020 - Proceedings*, *978*, 339–343. https://doi.org/10.1109/ICIIS51140.2020.9342738

Reddy, S. M., Mohammed Ali, S., Battula, K. M., Charan, P. R. N. L., & Rashmi, M. (2023). Web App for Predicting Fake Job Posts Using Ensemble Classifiers. *2023 4th International Conference for Emerging Technology, INCET 2023*, 1–5. https://doi.org/10.1109/INCET57972.2023.10170567

Swathi, M., Mounika, K., & Professor, A. (2022). Prediction of Fake Job Post Using Machine Learning. *International Research Journal of Modernization in Engineering Technology and Science*, *04*(10), 1691–1696. https://doi.org/https://www.doi.org/10.56726/IRJMETS30925

Wang, R., Chen, Q., Wang, Y., Xiong, L., & Shen, B. (2024). JobViz: Skill-driven visual exploration of job advertisements. *Visual Informatics*, *8*(3), 18–28. https://doi.org/10.1016/j.visinf.2024.07.001