Analysis of Fraudulent Job Postings Using Machine Learning

Said Salloum
School of Science, Engineering, and Environment
University of Salford
Manchester, UK
salloum78@live.com

Khalaf Tahat
Media & Creative Industries, United Arab Emirates University and Yarmouk University, Jordan
k.tahat@uaeu.ac.ae

Ahmed Mansoori
Media & Creative Industries department, United Arab Emirates University, UAE
a.mansoor@uaeu.ac.ae

Raghad Alfaisal
Faculty of Computing and Meta-Technology, Universiti Pendidikan Sultan Idris, Tanjung Malim, Malaysia
raghad.alfaisal81@gmail.com

Dina Tahat
Al Ain University, Applied sociology department, UAE
dina.tahat@aau.ac.ae

***Abstract*—In the age of digital recruitment, the proliferation of fraudulent job postings poses significant challenges for job seekers and legitimate employers alike. These deceptive listings not only waste time and resources but also endanger personal data and propagate scams. Addressing this issue, we present a comprehensive machine learning methodology to accurately discern between genuine and counterfeit job opportunities. Leveraging a rich dataset procured from Kaggle, this paper details the deployment of a logistic regression classifier, judiciously trained on a fusion of textual and meta-features extracted from job advertisements. The classifier underwent rigorous evaluation, manifesting an impressive accuracy of 96.78% in segregating authentic posts from fraudulent ones. The implementation of Term Frequency-Inverse Document Frequency (TF-IDF) vectorization on textual data, alongside meta-features such as job description length, enabled the model to learn and predict with high precision. The implications of this research are substantial, offering a scalable and efficient tool for job platforms to safeguard their users and ensure the integrity of their listings.**

Keywords—Digital Recruitment, Fraudulent Job Postings, Machine Learning, Logistic Regression Classifier, TF-IDF Vectorization.

I. Introduction

The digital era has revolutionized the job market by facilitating the proliferation of online job portals, which serve as vital platforms for connecting employers with potential employees [1]–[3]. Despite the convenience and efficiency these platforms offer, they have also become a breeding ground for fraudulent job postings [4], [5]. These fake listings not only mislead job seekers [6], [7] but also pose potential risks to personal data security [8], [9].

Several studies have explored various methodologies for detecting deceptive online content, ranging from email phishing [10]–[12] to fake product reviews [13]. However, the specific domain of fraudulent job postings remains under-explored, presenting unique challenges due to the subtlety and complexity of the deception used.

In response to this, we employ machine learning techniques, which have shown promising results in text classification tasks [14], to differentiate between real and fake job advertisements. Our approach utilizes a logistic regression classifier, a model chosen for its balance of performance and interpretability [15] , and is trained on a combination of text data features and meta-features.

The contribution of this research lies in its focused application of logistic regression to the problem of fraudulent job postings, an area that has seen limited application of this method. Furthermore, by combining text analysis with meta-data, we provide a nuanced approach that captures a broader spectrum of indicators of fraud.

This paper is structured as follows: Section 2 reviews related work in the field. Section 3 details the methodology, including data preprocessing, feature engineering, and model training. Section 4 presents the results and a thorough evaluation of the model's performance. Section 5 discusses the implications, limitations, and avenues for future research. Finally, Section 6 concludes the paper with a summary of our findings and their significance for digital recruitment platforms.

II. Methodology

A. Dataset Description

The dataset utilized for this investigation was sourced from the Kaggle platform, a repository known for its comprehensive compilation of datasets for machine learning research [16]. It comprises a variety of job postings, each rich in attributes such as job descriptions, requirements, benefits, and crucially, a label indicating the posting's legitimacy. Such datasets have been pivotal in the development of automated systems for content validation.

B. Data Preprocessing

Consistent with standard practices in natural language processing (NLP), we concatenated the description, requirements, and benefits text from each job posting to form a singular, robust text feature [17], [18] . This decision aligns with the recommendations by [19] for text

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

aggregation to enhance model performance. In the presence of incomplete records, we imputed empty strings, a method supported by findings from [20] on maintaining consistency within machine learning feature sets.

C. Feature Engineering

Following the feature preparation, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was employed, as advocated by [11], [21], [22], transforming the corpus of textual content into a numerical array indicative of term significance in the document and corpus at large. This approach has been validated in its effectiveness for text classification tasks [23], [24]. We also computed meta-features, including the length of job descriptions, to enrich our feature set, as meta-features have been shown to enhance classification models.

D. Model Training and Evaluation

In accordance with prevalent practices, a logistic regression classifier, known for its efficacy in binary classification problems [25], [26], was trained on the dataset. Model performance was gauged through established classification metrics, such as accuracy, precision, recall, and the F1-score, which together offer a holistic view of the model's predictive capabilities. The Receiver Operating Characteristic (ROC) curve, a tool for assessing the true-positive to false-positive ratio, was also plotted, in line with the recommendations by ROC for evaluating classifier threshold settings.

III. Findings and Discussion

In evaluating the performance of our logistic regression model, we obtained a high accuracy of 96.78% on the test dataset, which indicates a strong overall ability to classify job postings correctly. This high accuracy, however, must be contextualized with the understanding that our dataset is imbalanced, with a greater number of real job postings compared to fraudulent ones.

The precision metric reached the upper bound at 1.00, suggesting that every job posting predicted as fraudulent by our model was indeed a fraudulent posting. This level of precision implies that the model is highly reliable when it flags a job posting as potentially fraudulent.

Conversely, the recall was notably lower at 0.36, which indicates that the model failed to identify a significant number of fraudulent postings. This lower recall points to the model's conservative nature in classifying postings as fraudulent, potentially allowing some fraudulent postings to be classified as real.

The F1-score, which balances the precision and recall, was calculated to be 0.53. The modest F1-score underscores the disparity between the precision and recall and highlights an area for improvement. The aim would be to enhance the model's sensitivity to fraudulent postings without compromising its precision.

Upon analyzing the Receiver Operating Characteristic (ROC) curve (see Fig. 1), we observed that the model exhibits an excellent area under the curve (AUC) of 0.95. The ROC curve, which plots the true positive rate against the false positive rate, demonstrates the model's ability to discriminate between the classes at various threshold settings.

The confusion matrix further solidifies our findings (See Fig. 2), where we see that out of the 181 actual fraudulent postings, the model correctly identified 66 as fraudulent (true positives) and misclassified 115 as real (false negatives). Importantly, there were no real postings misclassified as fraudulent (false positives), which aligns with the precision score.

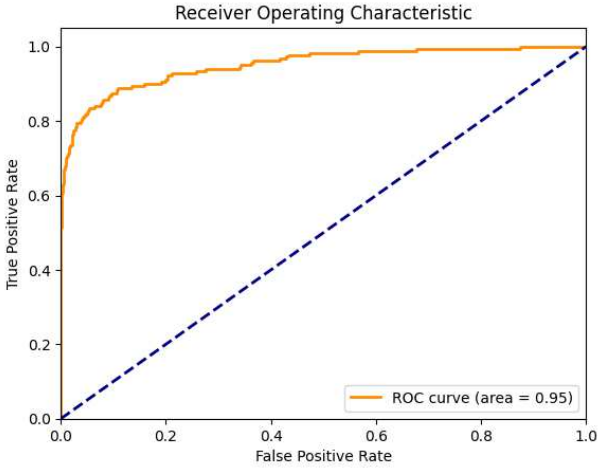

Fig. 1. ROC Curve

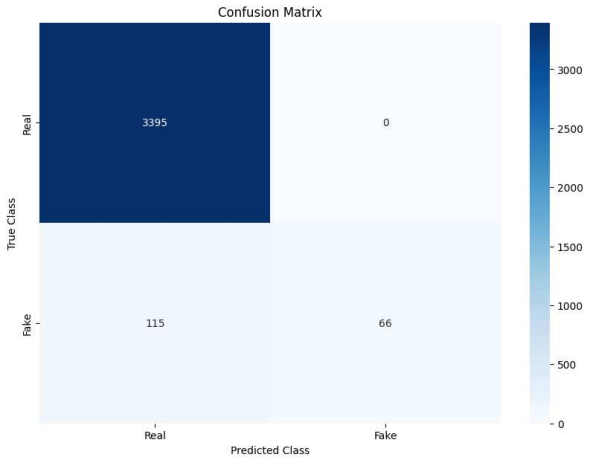

Fig. 2. Confusion Matrix

IV. Coclusion

This investigation underscores the significant potential of machine learning techniques in the realm of digital recruitment, particularly for the detection of fraudulent job postings. By implementing a logistic regression classifier trained on a substantial Kaggle dataset, the study achieved a model with high precision, indicating a strong capability to correctly identify fraudulent postings when they are detected. The practical implications of such a model are considerable; it offers a tool for job platforms to significantly diminish the

presence of fraudulent listings, thereby protecting job seekers and upholding the platforms' reputations.

However, the study is not without its limitations. Most notably, the model's recall was relatively low, which means that while the predictions made by the model are highly reliable, it does not catch all fraudulent postings. This limitation suggests that many fraudulent postings may still evade detection, which could undermine the trustworthiness and safety of online job markets.

In terms of future work, several avenues appear promising. Enhancing the model's recall without substantially lowering precision could be achieved through the application of more sophisticated machine learning algorithms, such as ensemble methods or deep learning. These approaches could provide a more nuanced analysis of the textual data and potentially uncover patterns that a logistic regression model might miss.

Additionally, the feature set could be expanded to include more meta-features or even utilize alternative text representation techniques beyond TF-IDF, such as word embeddings or transformer-based models like BERT, which have demonstrated impressive results in NLP tasks.

Another area for future research could involve the balancing of the dataset. Given the imbalance observed with a greater number of real postings compared to fraudulent ones, techniques such as synthetic data generation or resampling methods could be employed to provide a more balanced training ground for the model.

In conclusion, while the current model exhibits high precision, it is the balanced interplay between precision and recall that will truly enhance the model's utility. Continued research and model refinement are essential for developing a robust solution that can reliably filter out fraudulent job postings, thereby fostering a safer job-seeking environment on the internet.

References

[1] S. P. Parker, G. G., Van Alstyne, M. W., & Choudary, "Platform revolution: How networked markets are transforming the economy and how to make them work for you," *WW Nort. Co.*, 2016.

[2] M. C. Habes M., Alghizzawi M., Salloum S.A., M. Habes, M. Alghizzawi, S. A. Salloum, and C. Mhamdi, "Effects of Facebook Personal News Sharing on Building Social Capital in Jordanian Universities.," *Al-Emran M., Shaalan K., Hassanien A. Recent Adv. Intell. Syst. Smart Appl. Stud. Syst. Decis. Control. vol 295. Springer, Cham*, vol. 295, no. June 2020, pp. 653–670, 2021, doi: 10.1007/978-3-030-47411-9_35.

[3] A. Alyammahi, M. Alshurideh, B. A. Kurdi, and S. A. Salloum, *The Impacts of Communication Ethics on Workplace Decision Making and Productivity*, vol. 1261 AISC. 2021.

[4] A. A. A. Mehrez, M. Alshurideh, B. A. Kurdi, and S. A. Salloum, *Internal Factors Affect Knowledge Management and Firm Performance: A Systematic Review*, vol. 1261 AISC. 2021.

[5] F. Al Suwaidi, M. Alshurideh, B. Al Kurdi, and S. A. Salloum, *The Impact of Innovation Management in SMEs Performance: A Systematic Review*, vol. 1261 AISC. 2021.

[6] K. M. Hunt, "Gaming the system: Fake online reviews v. consumer law," *Comput. law Secur. Rev.*, vol. 31, no. 1, pp. 3–25, 2015.

[7] M. Alshurideh, B. H. Al Kurdi, H. M. Alzoubi, and S. Salloum, *The Effect of Information Technology on Business and Marketing Intelligence Systems*. Springer, 2023.

[8] R. Naudé, M., Adebayo, K. J., & Nanda, "A machine learning approach to detecting fraudulent job types," *AI Soc.*, vol. 38, no. 2, pp. 1013–1024, 2023.

[9] S. Salloum, K. Tahat, D. Tahat, A. Mansoori, and R. Alfaisal, "Delving Into the Security & Privacy of the Metaverse Matrix," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2023, pp. 1–5.

[10] S. Salloum, T. Gaber, S. Vadera, and K. Sharan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," *IEEE Access*, 2022.

[11] S. Salloum, "Enhancing Cybersecurity: Machine Learning and Natural Language Processing for Arabic Phishing Email Detection," University of Salford, 2024.

[12] S. Salloum, T. Gaber, S. Vadera, and K. Shaalan, "Phishing Email Detection Using Natural Language Processing Techniques: A Literature Survey," *Procedia Comput. Sci.*, vol. 189, pp. 19–28, 2021.

[13] B. J. Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, "Creating and detecting fake reviews of online products," *J. Retail. Consum. Serv.*, 2022.

[14] and A. T. Taha, Kamal, Paul D. Yoo, Chan Yeun, "Text Classification: A Review, Empirical, and Experimental Evaluation," *arXiv Prepr. arXiv*, p. 2401.12982, 2024.

[15] and R. K. Cemernek, David, Shafaq Siddiqi, "Effects of Class Imbalance Countermeasures on Interpretability," *IEEE Access*, 2024.

[16] "Real / Fake Job Posting Prediction," *Kaggle*, 2020. https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction.

[17] M. Bahja, "Natural Language Processing Applications in Business," in *E-Business [Working Title]*, IntechOpen, 2020.

[18] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: an overview," *J. King Saud Univ. Inf. Sci.*, 2019.

[19] T. A. F. Green, "Using NLP to resolve mismatches between jobseekers and positions in recruitment," 2023.

[20] and D. Z. Abowd, John M., Joelle Abramowitz, Margaret C. Levenstein, Kristin McCue, Dhiren Patki, Trivellore Raghunathan, Ann Michelle Rodgers, Matthew D. Shapiro, Nada Wasi, "Finding needles in haystacks: Multiple-imputation record linkage using machine learning," *Work. Pap.*, 2021, [Online]. Available: https://hdl.handle.net/10419/273032.

[21] S. A. Salloum, M. Al-Emran, S. Abdallah, and K. Shaalan, "Analyzing the Arab Gulf Newspapers Using Text Mining Techniques," in *International Conference on Advanced Intelligent Systems and Informatics*, 2017, pp. 396–405, doi: 10.1007/978-3-319-64861-3_37.

[22] S. A. Salloum, M. Al-Emran, and K. Shaalan, "Mining Text in News Channels: A Case Study from Facebook," *Int. J. Inf. Technol. Lang. Stud.*, vol. 1, no. 1, pp. 1–9, 2017.

[23] S. A. Salloum, M. Al-Emran, and K. Shaalan, "A Survey of Lexical Functional Grammar in the Arabic Context," *Int. J. Com. Net. Tech*, vol. 4, no. 3, 2016.

[24] K. Tahat, A. Mansoori, D. N. Tahat, M. Habes, and S. Salloum, "Leveraging Soft Power: A Study of Emirati Online Journalism Through Arabic Topic Modeling," in *International Conference on Business and Technology*, 2023, pp. 13–20.

[25] S. De Cock, M., Dowsley, R., Horst, C., Katti, R., Nascimento, A. C., Poon, W. S., & Truex, "Efficient and private scoring of decision trees, support vector machines and logistic regression models based on pre-computation," *IEEE Trans. Dependable Secur. Comput.*, vol. 16, no. 2, pp. 217–230, 2017.

[26] K. Tahat *et al.*, "Uncovering the Share Fake News on Social Media During Crisis," in *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2023, pp. 1–6.