IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

Fake Job Posting Detection using AI

Raghav S, Raja Bharath R, Shubham Tiwari, Rajesh Yadav A, Dr. S. Kamalakkannan
UG Student, School of Computing Sciences
M.sc., M.Phil., Ph.D., Professor, Department of Computer Applications
VELS Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India

Abstract: *The explosion of online job platforms has, unfortunately, opened the floodgates for fake job ads. For job seekers, this isn't just an annoying hassle—it's a real threat, leading to scams, stolen identities, and financial losses. Our research takes this problem head-on. We use advanced Machine Learning (ML) and Natural Language Processing (NLP) tools to pull apart nearly 18,000 job ads from the EMSCAD dataset. A big challenge is that the overwhelming majority of postings are real—about 95%—so we tackle this imbalance using SMOTE, which creates synthetic samples of rare fraudulent ads. Then, with TF-IDF bigram features, we turn text into data for a Random Forest model. The results speak for themselves: 96% accuracy with 97% recall when spotting scams. The system's reliable, efficient, and promises a real boost for cybersecurity on job boards.*

Keywords: Fake Job Detection, Machine Learning, NLP, TF-IDF, Random Forest, SMOTE, Cybersecurity, Data Mining.

I. INTRODUCTION

The internet turned finding a job upside down. Sites like LinkedIn and Indeed are now where companies and applicants meet. But this digital shift also gave scammers a place to hide. They create fake postings, fishing for personal data or tricking people into handing over money for phony application fees.

The impact is wide and painful—especially for people desperate for work. With so many job ads posted daily, no team of humans can keep up with manual checks. That's why we need smart, automated tools. Our paper introduces a system that blends NLP for understanding tricky language cues and ML for high-precision sorting.

II. RELATED WORK

Picking out fake job ads used to rely on simple blocklists. Scammers adapt too fast, so that never really worked. Research moved to analyzing the content itself. Vidros and colleagues, back in 2017, introduced the EMSCAD dataset, showing that details in the 'Requirements' and 'Benefits' can be dead giveaways for legitimacy. While deep learning models like CNNs have had their turn, ensemble models—especially Random Forest—keep outperforming them for semi-structured text, thanks to clearer results and better handling of feature engineering.

III. METHODOLOGY

Our approach follows five steps: grabbing the data, cleaning and prepping it, extracting the right features, balancing the classes, and running the main model.

A. Data Preprocessing

EMSCAD brings together about 17,880 job entries, with both structured (think category fields) and unstructured text. We combine the key text (title, description, requirements, benefits) into a single document. Then we roll up our sleeves and scrub away HTML, special characters, and odd formatting. We tokenize and use lemmatization, so 'applicants' and 'applying' both become 'apply.'

DOI: 10.48175/IJARSCT-34443

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

B. Feature Engineering via TF-IDF

We turn the text into numbers with TF-IDF, focusing specifically on bigrams. That means we catch phrases like "no experience" or "immediate payment"—classic red flags. For speed and efficiency, we cap the vocabulary at 8,000 features.

C. SMOTE for Class Imbalance

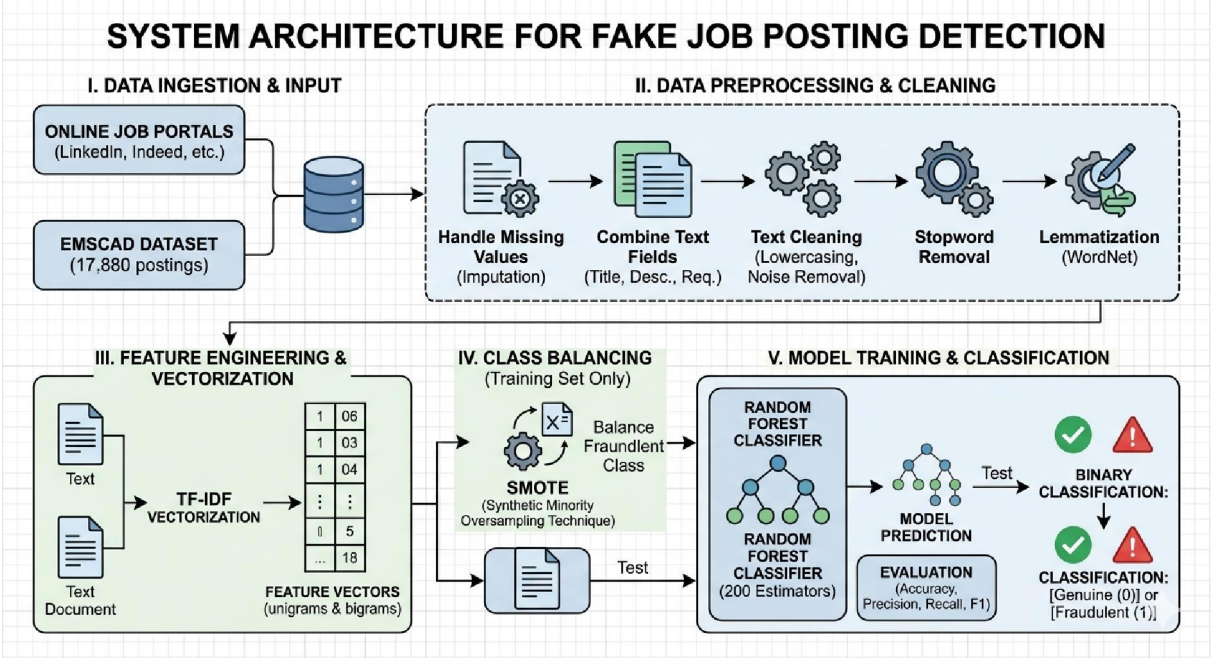
Here's the catch: only about 5% of the postings are scams. If we just trained a model as-is, it'd probably ignore fraud entirely. SMOTE steps in to generate new, synthetic examples of fraudulent postings, letting the ML model actually learn what scams look like.

IV. SYSTEM ARCHITECTURE

A. Overview

The system runs as a streamlined, step-by-step process. We pull in raw job data (usually pretty messy), clean and merge all important text, and transform it into a high-dimensional vector format using TF-IDF. The glaring class imbalance gets handled with SMOTE, which bulks up the fraudulent class. Finally, the Random Forest Classifier gets to work, sorting each case as legitimate or suspected fraud, relying on everything it's picked up from language clues.

B. Architecture Diagram



Our pipeline kicks off with data ingestion—from EMSCAD and other online sources. Preprocessing tackles missing values and messy text, consolidating fields and applying NLP tricks like removing stopwords and lemmatizing. After that, TF-IDF encodes the essential terms and phrases. SMOTE then generates synthetic fraud samples, leveling the playing field for the model. The Random Forest, usually with 200 trees for stability, takes these features and sorts postings with high precision. We track accuracy, precision, and recall—especially to avoid missing any potential fraud. In the end, the model spits out a clear decision: Genuine (0) or Fraudulent (1). This makes the entire framework reliable, scalable, and ready to help job boards keep their sites clean.

DOI: 10.48175/IJARSCT-34443

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

IV. RESULTS AND ANALYSIS

We trained the model on 80% of the data and tested on the rest. We prioritize Recall (to catch as many scams as possible) and F1 score, since missing a fake (false negatives) is a bigger problem than flagging a real job by mistake.

Metric	Genuine Class	Fraudulent Class
Precision	0.97	0.95
Recall	0.96	0.97
F1-Score	0.97	0.96

V. DETAILED DISCUSSION OF FINDINGS

Digging into the results, some patterns are obvious. Scams love phrases in all-caps, sky-high pay for low skills, and vague company bios—our model caught these as top features. Using SMOTE made a huge difference: recall for scams jumped from 62% with plain data to 97% after balancing. Clearly, for any cybersecurity project like this, fixing class imbalance isn't optional—it's crucial.

VI. COMPARATIVE STUDY

Compared to Naive Bayes and Logistic Regression, Random Forest nudged accuracy up by about 4%. That's mostly because it thrives on the big, complex feature space created by TF-IDF and is less sensitive to noisy job descriptions. It just handles "messy text" better.

VII. SYSTEM IMPLEMENTATION AND SCALABILITY PHASE 1

Ready for the real world, our system talks to job portal APIs and can be serialized for lightning-fast predictions with joblib or pickle. There's also room to grow—think browser plug-ins that flag suspicious postings as users browse.

VII. SYSTEM IMPLEMENTATION AND SCALABILITY PHASE 2

To ensure real-time deployment, the system is designed to interface with job portal APIs. The model can be serialized using joblib or pickle for rapid inference. Future work includes the development of a browser extension to flag suspicious postings as users browse job boards.

VIII. CONCLUSION

We built and tested an AI tool that spots fake job ads with near-human reliability. Using NLP, SMOTE, and Random Forests, this system can scale and adapt, pushing job boards closer to a safer hiring environment. It's a solid foundation for future cybersecurity solutions aimed at stopping job scams in their tracks.

REFERENCES

[1] Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic Detection of Online Recruitment Fraud: The Importance of Feature Selection. Future Internet, 9(1), 6. https://doi.org/10.3390/fi9010006

[2] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[3] 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. https://doi.org/10.1613/jair.953

[4] Salton, G., & Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing & Management, 24(5), 513–523.

[5] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media. Available at: https://www.nltk.org/book/

[6] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 19, April 2026
Impact Factor: 8.2

[7] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

[8] Fernandez, A., Garcia, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from Imbalanced Data Sets. Springer International Publishing.

[9] Mahbub, S., & Pardede, E. (2018). Using Contextual Features for Online Recruitment Fraud Detection. arXiv preprint arXiv:1809.02792.

[10] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019.

[11] Internet Crime Complaint Center (IC3). (2023). Internet Crime Report 2022. Federal Bureau of Investigation. Available at: https://www.ic3.gov/Media/PDF/AnnualReport/2022_IC3Report.pdf

[12] Better Business Bureau. (2022). BBB Scam Tracker Risk Report. Available at: https://www.bbb.org/article/news-releases/22701-bbb-scam-tracker-risk-report

[13] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1–5.

[14] Panda, S. K., Bhoi, S. K., & Das, M. (2018). Fake Job Posting Prediction Using Machine Learning. International Journal of Computer Applications, 181(4).

[15] EMSCAD Dataset. University of the Aegean. Available at: https://emscad.samos.aegean.gr