IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Fake Job Recruitment Detection Using a Machine Learning Approach

[1] Vendra mohana kalyani, [2] K Chinna Nagaraju,[3] V Anil Santosh

1 M.Tech Scholar, Department of CSE, International School Of Technology And Sciences For Women(A) NH-16 East Gonagudem Rajanagaram, AP, India,

2 Associate professor, Department of CSE, International School Of Technology And Sciences For Women (A) NH-16 East Gonagudem Rajanagaram, AP, India.

3 Associate Professor and HoD, Department of CSE, International School Of Technology And Sciences For Women(A) NH-16 East Gonagudem Rajanagaram, AP, lndia.

Abstract

This study presents a software application designed to detect fraudulent job advertisements through machine learning techniques. Leveraging a comprehensive dataset, the system employs multiple classifiers to analyze job listings and evaluate their authenticity, comparing individual classifiers with a joint classification approach. The results highlight that joint classification significantly improves fraud detection accuracy compared to single classifiers, effectively identifying fake job advertisements among genuine postings. The developed solution has the potential to enhance the security and reliability of online job platforms, protecting users from fraudulent recruitment schemes.

Keywords: Detecting fake job, machine learning model, authenticity, fraudulent recruitment schemes.

1. Introduction

Background

With the widespread use of online job boards, fraudulent job advertisements have become a prevalent issue, impacting millions of job seekers globally. Fake job listings mislead individuals, leading to personal data breaches, financial loss, and compromised personal safety. Traditional job platforms lack comprehensive mechanisms to detect such frauds effectively, creating an urgent need for automated detection systems.

Problem Statement

Detecting fake job postings is complex due to the subtle language manipulations scammers employ to make their listings appear legitimate. Fake recruitment detection requires sophisticated analysis capable of recognizing nuanced patterns that

distinguish scams from genuine job opportunities. Our project focuses on building a machine learning model to accurately classify job ads as fraudulent or legitimate, aiming to support platforms in reducing exposure to fake listings.

Objectives

This project seeks to:

Develop a machine learning model to classify job ads as either genuine or fraudulent.

Compare the effectiveness of individual classifiers versus a joint classification approach.

Implement a scalable, reliable solution that can be integrated into online job platforms to automatically filter out fake job advertisements.

2. Literature Review

Related Work

Fraud detection has been explored extensively across financial, e-commerce, and social networking platforms, with notable advances in machine learning applications. Recent research has explored spam detection in emails and product review systems using Natural Language Processing (NLP) and various classification algorithms. However, specific studies focusing on fake job recruitment detection remain limited, highlighting a need for targeted research in this field.

Challenges in Fake Job Detection

Fake job recruitment differs from other forms of fraud detection due to the complex nature of language manipulation in job ads. Unlike spam detection, where patterns are more overt, fake job ads often mimic genuine job postings, requiring the model to detect subtle indicators of fraud. Additionally, scammers continuously adapt their techniques, making it essential for detection systems to evolve accordingly.

Existing Approaches

Some online job platforms utilize basic keyword filtering and rule-based detection, which lack the adaptability of machine learning models. Studies in text classification for detecting fraudulent reviews and financial scams provide a foundation for developing robust machine learning approaches, suggesting that classification models, especially ensemble methods, could significantly improve the accuracy of detecting fake job ads.

3. Methodology

Data Collection

Our dataset consists of job advertisements labeled as either genuine or fraudulent, with attributes including job title, description, company information, contact details, and application procedures. We sourced data from online job boards and curated fake job postings identified through keyword searches and pattern matching.

Data Preprocessing

Data preprocessing involved:

Text Cleaning: Removing special characters, URLs, and stop words from text fields.

Normalization: Converting all text to lowercase for consistency.

Handling Missing Values: Filling or discarding incomplete records based on the attribute relevance.

Feature Engineering

Key features were extracted from each job advertisement, focusing on text analysis and specific indicators of fraud. Features included:

Textual Analysis: Common fraudulent patterns like "urgent hiring," "no experience needed," or "work from home."

Company Details: Legitimate companies often include clear contact information and professional language, while fake listings lack this clarity.

Job Description Length: Fake ads are often shorter and vaguer than genuine ads.

Classification Models

To classify the ads, we implemented several machine learning classifiers:

Naive Bayes: A probabilistic classifier useful for text classification.

Support Vector Machine (SVM): Effective in high-dimensional spaces, making it suitable for text data.

Random Forest: An ensemble learning method that increases robustness by using multiple decision trees.

Decision Tree: Provides interpretable results and identifies key decision points for fraud.

Single vs. Joint Classification

We compared individual classifiers with a joint classification approach, combining models to improve detection accuracy. Ensemble techniques such as stacking, voting, and boosting were explored to assess their effectiveness in identifying fraudulent listings.

4. Experimental Setup

Environment

The project was implemented in Python, using libraries such as Pandas, Scikit-Learn, and NLTK for data processing, model building, and evaluation. Experiments were conducted on a system with a minimum of 8GB RAM and an Intel i5 processor, sufficient to handle the dataset size and processing requirements.

Evaluation Metrics

The effectiveness of each model was evaluated using:

Accuracy: Proportion of correctly classified ads.

Precision and Recall: To measure true positives against all positives and false negatives.

F1-Score: A balance between precision and recall.

AUC-ROC: Provides insights into the model's true positive rate versus false positive rate.

5. Results

Model Performance

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Naive Bayes	85%	0.83	0.82	0.82	0.87
SVM	89%	0.88	0.87	0.88	0.91
Random Forest	92%	0.90	0.91	0.91	0.93
Joint Classifier	94%	0.93	0.92	0.93	0.96

Single vs. Joint Classification Results

The joint classification approach outperformed individual classifiers, with an accuracy of 94% and an AUC-ROC of 0.96. These results demonstrate that combining classifiers enhances the model's ability to detect fake job advertisements.

Case Study Analysis

Several cases were analyzed to understand specific instances where the model accurately identified fraud. The model successfully flagged ads with vague job descriptions, unrealistic pay, and non-standard contact information.

6. Discussion

Interpretation of Results

The high accuracy achieved by the joint classification method highlights the advantage of combining classifiers to improve detection accuracy. Models like Random Forest and SVM demonstrated strong individual performance but benefited further from ensemble techniques.

Strengths of Joint Classification

The ensemble model leverages the strengths of different classifiers, making it more resilient to the diverse ways in which scammers structure fake ads.

Limitations

While the joint classification approach performed well, challenges included handling ambiguous job ads and ads that closely mimicked legitimate language.

7. Conclusion

This project successfully demonstrates the potential of machine learning in combating fake job recruitment schemes, which pose a growing threat to job seekers globally. Through extensive experimentation, we found that while individual classifiers, such as Random Forest and SVM, offer high accuracy, a joint classification approach significantly improves the detection rate, achieving a higher degree of reliability and robustness. This finding aligns with trends in machine learning that highlight ensemble methods as a powerful tool for fraud detection tasks.

The joint classifier model not only achieved the highest accuracy but also demonstrated resilience to the evolving tactics of scammers, who frequently adjust language and format to bypass traditional keyword or rule-based detection systems. By leveraging a combination of classifiers, our model could recognize complex patterns and nuances across diverse job listings, which may not be apparent when using a single classifier. This capability is particularly beneficial for detecting fraud in unstructured

text, where fraudulent indicators may vary widely.

Broader Implications

The model developed in this study has several practical applications. It could be integrated into online job platforms, recruitment portals, and career websites as an automated filter to screen postings before they reach job seekers. This would not only protect users from scams but also enhance the credibility and trustworthiness of these platforms, thereby improving user experience and engagement. Additionally, human resource departments could employ similar models internally to screen unsolicited job offers or potential candidates who may attempt to leverage false credentials.

Insights Gained

Several insights emerged from this project:

Feature Selection and Engineering: Language patterns and specific keywords were strong indicators of fraud. However, combining these with contextual features such as company reputation, job description structure, and application requirements significantly enhanced detection accuracy.

Adaptability of Ensemble Methods: The success of the joint classification approach underlines the importance of using adaptable and resilient models in fraud detection. In environments where scammers continuously change tactics, the adaptability of machine learning methods is essential.

Importance of Diverse Datasets: A diverse dataset was instrumental in training the model to handle various types of fake ads. Expanding datasets with additional fraud scenarios, industry variations, and evolving scam tactics could further refine model accuracy.

Limitations and Areas for Further Research

While the joint classification model achieved high accuracy, certain limitations remain. The model's performance could be affected by changes in language patterns that were not represented in the training data, as scammers adapt their methods over time. To address this, future research could focus on developing models with real-time learning capabilities, enabling the system to adapt dynamically to new fraud patterns. Additionally, expanding the dataset to cover a broader range of industries and job levels would help generalize the model's performance across various recruitment contexts.

Practical Application and Future Directions

The results underscore the viability of implementing machine learning models for large-scale fraud detection in online job advertisements. To make this system more accessible, a user-friendly interface could be developed, allowing job platforms to integrate the model easily. Furthermore, incorporating advanced Natural Language Processing (NLP) techniques, such as deep learning-based language models, could improve the model's ability to detect subtler forms of fraud.

Future work could also explore the integration of social media and company reviews data as additional verification layers, which would help validate job listings by cross-referencing details. This

multi-layered approach could yield even greater accuracy and further reduce the chances of job seekers falling victim to scams.

Final Thoughts

This project has highlighted the efficacy of machine learning in tackling modern recruitment fraud, offering a scalable and effective solution for a problem that affects millions of job seekers. By advancing automated fraud detection systems, we are not only improving the safety of online job markets but also contributing to a more trustworthy and accessible job-seeking environment. This work serves as a foundation for future research and development in the field of recruitment fraud detection, paving the way for more sophisticated, real-time, and adaptive fraud detection solutions.

8. Future Work

Future improvements could include:

Natural Language Processing: Advanced NLP techniques could improve detection of nuanced fraud patterns.

Real-Time Detection: Integrating real-time analysis could enhance applicability.

Larger Datasets: Expanding the dataset would improve model generalizability across more industries.

9. References

.AlghamdiandF.Alharby,—AnIntelligentModelforOnlineRecruitmentFraudDetection,"J.Inf. Secur., vol. 10,no. 03,pp. 155–176,2019, doi:10.4236/jis.2019.103009.

[2] I. Rish, —An Empirical Study of the Naïve Bayes Classifier An empirical study of the naiveBayesclassifier,‖no. January2001, pp. 41–46, 2014.

[3] D.E.Walters,—Bayes'sTheoremandtheAnalysisofBinomialRandomVariables,‖BiometricalJ., vol. 30, no. 7, pp. 817–825,1988, doi: 10.1002/bimj.4710300710.

[4] F.Murtagh,—Multilayerperceptronsforclassificationandregression,‖Neurocomputing,vol.2,no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P.CunninghamandS.J.Delany,—K-NearestNeighbourClassifiers,‖Mult.Classif.Syst.,no.May,pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6] H. Sharma and S. Kumar, —A Survey on Decision Tree Algorithms of Classification in DataMining,‖ Int. J.Sci. Res.,vol. 5,no. 4,pp. 2094–2097,2016, doi:10.21275/v5i4.nov162954.

[7] E.G.Dada,J.S.Bassi,H.Chiroma,S.M. Abdulhamid,A.O.Adetunmbi,andO.E.Ajibuwa,"Machine learning for email spam filtering: review, approaches and open research problems,‖Heliyon,vol. 5, no. 6, 2019, doi: 10.1016/j.heliyon.2019.e01802.

[8] L. Breiman, —ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001,doi:10.1017/CBO9781107415324.004.

[9] B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, —Bagging classifiers for fightingpoisoningattacksinadversarialclassificationtasks,"Lect.NotesComput.Sci.(includingSubser.Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp.

350–359, 2011, doi:10.1007/978-3-642-21557-5_37.

[10] A. Natekin and A. Knoll, ―Gradient boosting machines, a tutorial,‖ Front. Neurorobot., vol.7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[11] N. Hussain, H. T. Mirza, G. Rasool, I. Hussain, and M. Kaleem, ―Spam review detectiontechniques:Asystematicliteraturere view,‖Appl.Sci.,vol.9,no.5,pp.1–26,2019,doi:10.3390/app9050987.

[12] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, ―Fake News Detection on Social Media,‖ACMSIGKDD Explor. Newsl., vol. 19,no. 1, pp. 22–36, 2017, doi:10.1145/3137597.3137600.

[13] ShivamBansal(2020,February).[RealorFake]FakeJobPostingPrediction,Version1. Retrieved March 29,2020 fromhttps://www.kaggle.com/shivamb/real-or-fake-fakejobposting-prediction

[14] H. M and S. M.N, ―A Review on Evaluation Metrics for Data Classification Evaluations,‖Int.J.DataMin.Knowl.Manag.Process,vol.5,no.2,pp.01–11,2015,doi:10.5121/ijdkp.2015.5201.

[15]S.M.Vieira,U.Kaymak,andJ.M.C.Sousa,―Cohen'skappacoefficientasaperformancemeasure for feature selection," 2010 IEEE World Congr. Comput. Intell. WCCI 2010, no. May2016,2010, doi: 10.1109/FUZZY.2010.5584447

Biography of authors:

Vendra mohana kalyani was a M.Tech Scholar in Department of CSE inInternational School Of Technology And Sciences For Women(A) NH-16 East Gonagudem Rajanagaram, AP, India. Her intersted research area is machine learning and artificial intelligence (AI) typically focuses on advanced computational techniques.

K chinna Nagaraju was an Associate Professor (PhD) in Department of CSE inInternational School Of Technology And Sciences For Women(A) NH-16 East Gonagudem Rajanagaram, AP, India. His current research work is is machine learning and artificial intelligence (AI) typically focuses on advanced computational techniques that enable machines to learn from data, identify patterns, and make decisions without being explicitly programmed.

V Anil Santosh was an Associate Professor (PhD) and Head of the Department of Department of CSE in International School Of Technology And Sciences For Women(A) NH-16 East Gonagudem Rajanagaram, AP, India. His current research work is a variety of AI subfields, including deep learning, neural networks, natural language processing, and reinforcement learning. Their work may involve developing new algorithms, applying AI to solve real-world problems (like forecasting, automation, or image recognition), and exploring ethical concerns related to AI deployment. Many such authors combine academic research with industry applications, publishing papers, books, or articles aimed at both technical and non-technical audiences.