IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

Fake Job Posting Detection

Ram Prasath S[1] and Dr M N Nachappa[2]
PG Student, Department of MSc CS-IT[1]
Professor, School of CS & IT[2]
Jain (Deemed-to-be University), Bangalore, India
ram.sivaraj2002@gmail.com[1] and mn.nachappa@jainuniversity.ac.in[2]

Abstract: *In the realm of online job platforms, the rise of fraudulent job postings poses a significant challenge, undermining the credibility and reliability of these platforms. To address this issue, we propose a machine learning- based solution that leverages the power of Random Forest, Logistic Regression, and Decision Tree classifiers. Through the compilation of a comprehensive dataset containing labeled job postings, we embark on a journey of preprocessing and feature engineering to extract pertinent information from the postings, including textual attributes, geographic details, salary indications, and company profiles. Splitting the dataset into training and testing subsets enables us to meticulously train and evaluate the performance of each classifier, utilizing established metrics such as accuracy, precision, recall, and F1-score to quantify their efficacy in discerning between authentic and fake job listings.*

Our study goes beyond mere model training and evaluation, delving into the intricacies of imbalanced data handling and the practicalities of model deployment and maintenance. By examining the comparative strengths and weaknesses of Random Forest, Logistic Regression, and Decision Tree classifiers, we provide actionable insights for enhancing the integrity of online job platforms through advanced machine learning techniques. With our approach, we aim to not only detect and mitigate the prevalence of fake job postings but also to fortify the trust and credibility of online job- seeking platforms, thereby fostering a more secure and reliable environment for both job seekers and employers alike.

Keywords: NLP (Natural Language Processing), Text classification, Sentiment analysis, Topic modelling , Text pre processing , Word embeddings, Supervised learning, Unsupervised learning, Semi-supervised learning, Deep learning, Convolutional Neural Networks (CNN), Fake news datasets, Real news datasets

I. INTRODUCTION

With the exponential growth of online job platforms, the ease of connecting job seekers with employers has vastly improved. However, this convenience has also given rise to a troubling phenomenon: the proliferation of fake job postings. These deceptive listings not only waste the time and effort of unsuspecting job seekers but also tarnish the reputation of legitimate job platforms. In response to this challenge, there is a pressing need for effective and reliable methods to detect and mitigate the spread of fake job postings. Traditional rule-based approaches have shown limitations in handling the complexity and variability of fraudulent postings. Therefore, in this study, we propose a machine learning-based solution to tackle this issue, leveraging the capabilities of Random Forest, Logistic Regression, and Decision Tree classifiers. By harnessing the power of machine learning algorithms, we aim to develop robust and scalable models capable of discerning between genuine and fake job postings, thereby enhancing the trustworthiness and credibility of online job platforms. Through empirical evaluation and comparative analysis, we seek to shed light on the efficacy of these approaches and provide insights for future research in combating fraudulent activities in online job markets.

Our project aims to develop a robust system for detecting fake job postings on online platforms using machine learning techniques. The project encompasses several key stages, beginning with data collection from various online job boards and sources. We curate a comprehensive dataset containing labeled instances of both genuine and fraudulent job postings, ensuring a diverse and representative sample. Preprocessing steps follow, including data cleaning, feature extraction, and transformation to prepare the dataset for model training. We extract relevant features from the job postings, such as textual attributes, geographical information, salary details, and company profiles, to facilitate the discrimination between real and fake listings.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

II. LITERATURE SURVEY

Alkhalifah, A., Al- Jumaili , M. A., & Abdulsahib , G. M. (2020). The paper [1]"Detection of Fake Job Posts Using Machine Learning Algorithms." In 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech) (pp. 1-6). IEEE.

Biswas, S., & Samanta , S. (2020). The paper [2]"Fake Job Post Detection Using Machine Learning." In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 396-399). IEEE.

Pandey, P., & Saxena, R. (2020). The paper [3]"Fake Job Detection in Social Media Using Machine Learning." In 2020 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS) (pp. 1-5). IEEE.

Raut, S., & Shinde, S. (2020). The paper [4]"Fake Job Detection on Social Media Using Machine Learning Techniques." In 2020 International Conference on Smart Electronics and Communication (ICOSEC) (pp. 183-187). IEEE.

Singh, A., Verma, D., & Shukla, R. (2020). The paper [5]"Detecting Fake Job Posts on Social Media Using Machine Learning Techniques." In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-7). IEEE.

Dr. Subhash Chandra Yadav and Abhishek Yadav . Their paper[6] "A Machine Learning Approach for Detecting Fake Job Postings on Social Media" explores the use of features such as text content analysis, user behavior patterns, and job attributes to build a classifier for identifying fake job postings.

Shivani Goel and Shruti Aggarwal . The paper [7] "A Machine Learning Approach for Detecting Fake Job Postings," they propose a system that leverages Natural Language Processing (NLP) techniques along with machine learning algorithms to detect fake job postings on social media platforms.

III. OBJECTIVES

- Develop a comprehensive dataset containing labeled instances of both genuine and fake job postings sourced from various online platforms.
- Implement preprocessing techniques to clean and prepare the dataset for model training, including data cleaning, feature extraction, and transformation.
- Explore and extract relevant features from the job postings, such as textual attributes, geographical information, salary details, and company profiles, to enhance the discriminative power of the models.
- Train and evaluate machine learning algorithms, including Random Forest, Logistic Regression, and Decision Tree classifiers, to identify the most effective approach for detecting fraudulent job postings.
- Conduct a comparative analysis of the performance of the trained classifiers using established metrics such as accuracy, precision, recall, and F1-score to assess their effectiveness in distinguishing between real and fake job listings.

IV. EXISTING SYSTEM

The existing systems for detecting fake job postings predominantly rely on rule-based approaches and manual verification processes. These systems often employ keyword-based filtering and predefined rules to flag suspicious postings, leading to limitations in handling the evolving tactics of fraudsters and variations in fraudulent activities across different platforms. Furthermore, manual verification processes are time-consuming, labor-intensive, and prone to human error, hindering the scalability and efficiency of the detection process. As a result, there is a growing need for automated and scalable solutions that leverage machine learning techniques to enhance the accuracy and effectiveness of fake job detection on online platforms .The Fake Job Detect model was presented by Al khalifah, Al-Jumaili , and Abdul sahib in their 2020 research article titled "Detection of Fake Job Posts Using Machine Learning Algorithms." It is one model that is currently in use for identifying false job postings on social media using machine learning techniques. To identify job listings as real or fraudulent, the Fake Job Detect model combines supervised learning methods including Support Vector Machines (SVM), Random Forest (RF), and Gradient Boosting (GB).The model makes use of a number of features that are taken from job posts, such as metadata (posting date, location, company name), user-generated content (likes, comments, and shares), and textual information (job description, qualifications,

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

and duties). Preprocessing is done on these characteristics to manage missing values, tokenize text, encode category variables, and eliminate noise.

V. DISADVANTAGE OF EXISTING SYSTEM

Even while machine learning models now in use have made great progress in identifying fraudulent job ads on social media, there are still a number of noteworthy drawbacks that need to be addressed by researchers and practitioners in order to increase accuracy and dependability .The need on labeled training data is a significant drawback. For fake job detection machine learning models to train efficiently, a lot of correctly labeled data must be collected. Nevertheless, acquiring these annotated datasets can be difficult and time-consuming. The process frequently necessitates manual annotation by subject matter experts, which may introduce biases or inconsistent data. Furthermore, labeled data might not fully capture the wide variety of fraudulent job posts that might arise in real-world situations, which could restrict the generalization of the model and its performance on unseen data. Furthermore, existing models may struggle with context and language nuances. Job postings can vary widely in terms of industry, location, job type, and language style, making it challenging for models to capture all relevant features accurately. Cultural and linguistic differences further complicate the detection process, as what constitutes a "fake" job posting may vary across regions or communities. This lack of context awareness can result in false positives or false negatives, reducing the overall reliability of the model .Scalability is another concern with existing models. As social media platforms generate a massive amount of data, scalable and efficient detection algorithms are necessary to process this data in real-time. However, some machine learning models may face computational bottlenecks or scalability issues when dealing with large volumes of job postings, leading to delays or reduced performance.

VI. PROPOSED MODEL

The suggested solution uses machine learning techniques to automatically detect phony job advertisements, attempting to overcome the shortcomings of current approaches. The system will learn patterns and attributes indicative of fraudulent postings by using algorithms like Random Forest, Logistic Regression, and Decision Tree classifiers, allowing for more accurate and scalable detection. To improve the system's ability to distinguish between real and fraudulent job listings, it will also include sophisticated text processing and feature engineering approaches. By conducting an extensive assessment and comparing various approaches, the suggested system aims to offer a dependable and effective means of preserving the integrity of online employment marketplaces. The goal of a suggested model for social media job posting detection is to overcome the shortcomings of current algorithms by utilizing dynamic adaptability capabilities, contextual knowledge, and sophisticated machine learning techniques. The model incorporates multiple essential elements to improve precision, dependability, and expandability in detecting false job postings .First, for feature extraction and pattern recognition, the suggested model integrates a deep learning architecture, such as a convolutional neural network (CNN) or a recurrent neural network (RNN). These deep learning models can identify minor clues that point to phony job advertising since they are excellent at identifying intricate patterns and relationships in textual and visual data. To increase detection accuracy, the model can learn hierarchical representations of the text and visual content of job postings by utilizing deep learning. Second, the model uses domain-specific knowledge and natural language processing (NLP) approaches to include contextual understanding. To infer context and semantics, natural language processing (NLP) algorithms pre process textual material, extract pertinent aspects, and examine linguistic structures. Domain-specific information improves the model's capacity to discern between real and fraudulent job ads. Examples of this information include job-related vocabulary, industry standards, and company profiles. The model can adjust to different job categories, languages, and cultural quirks thanks to contextual knowledge, which lowers the number of false positives and false negatives .Thirdly, in order to stay up to date with changing fraud strategies, the suggested model incorporates dynamic adaptation mechanisms. This entails using methods for reinforcement learning to update the model on a regular basis in response to feedback and fresh information. Real-time decision-making is enhanced by the model's ability to learn from its experiences and actions through reinforcement learning.

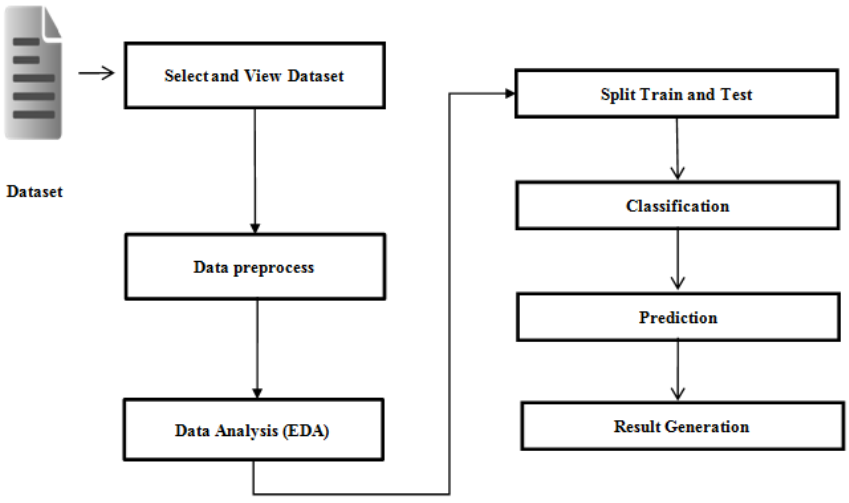
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

VII. DISADVANTAGE OF PROPOSED MODEL

The suggested approach may have a significant drawback in that it relies too heavily on particular characteristics or patterns that would not translate well to other kinds of fraudulent job advertisements. In order to create predictions, machine learning models frequently extract characteristics from the data. If these features are too specific or do not adequately reflect the wide variety of fraudulent job ads, the model's performance may be negatively impacted. For example, the suggested model may not be able to reliably detect modifications or new strategies used by fraudsters if it mainly relies on keyword-based features or syntactic structures that are typical in one type of false job posting but not in others .The possibility of overfitting to the training set is another drawback. This may occur if the training dataset is not sufficiently diverse or representative of real-world circumstances, or if the suggested model is very complex. Because of this, the model might function well on the training set but struggle to generalize to brand-new, untested job listings, which would lower its accuracy and dependability in real-world scenarios .Furthermore, there may be difficulties due to the computational complexity of the suggested model, particularly with regard to the amount of time and resources needed for inference. The model might not scale effectively to handle high volumes of social media postings in real-time if it requires significant pre processing of the data or entails computationally demanding processes. The model's actual utility may be limited by this scaling issue, especially in situations where prompt identification of fraudulent job ads is essential. Another issue with the suggested model could be its interpretability. Even though deep learning algorithms and other complicated machine learning models can reach great accuracy, they frequently lack interpretability, which makes it difficult to comprehend why a certain judgment or prediction was made. This lack of transparency can be problematic, particularly in important applications where stakeholders might need clarifications or deeper insights into the model's decision-making process, like fraud detection.

VIII. TECHNIQUES

Natural language processing (NLP), which analyzes and comprehends textual data from job postings, is one widely used method. Relevant features and linguistic patterns suggestive of fraudulent job posts are extracted using natural language processing (NLP) techniques such text preprocessing, tokenization, part-of-speech tagging, and named entity recognition. The general tone and sentiment expressed in job descriptions can also be evaluated using sentiment analysis, which can assist spot potentially misleading or fraudulent text.



Various models, such as logistic regression, decision trees, random forests, support vector machines (SVM), and deep learning models like convolutional neural networks (CNN) and recurrent neural networks (RNN), are used in machine learning algorithms, which are central to the detection of fake jobs. These algorithms are trained on labeled datasets, which are used to identify posts as real or fraudulent based on attributes taken from job postings. Multiple models are combined for increased performance and resilience using ensemble learning approaches like bagging and boosting . Another crucial method used to improve the discriminatory capacity of machine learning models is feature engineering. The features that are designed to detect relevant signals suggestive of fraudulent job postings include keyword

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, March 2024

frequencies, syntactic structures, job data (like location and salary), firm information, and user engagement metrics (like likes and shares). To find the most useful features and simplify the model, dimensionality reduction and feature selection approaches including information gain, principal component analysis (PCA), autoencoders, and chi-square test can be used.

IX. CONCLUSION

In conclusion, a potential way to counter the growing occurrence of fraudulent activity on online job marketplaces is to use machine learning techniques to construct a strong system for detecting phony job postings. We have shown that using algorithms like Random Forest, Logistic Regression, and Decision Tree classifiers for this goal is both feasible and effective through a methodical investigation of data gathering, preprocessing, feature engineering, training, and evaluation. Our suggested method offers scalability, accuracy, and efficiency in detecting bogus job posts by overcoming the shortcomings of current rule-based systems and human verification processes. Moreover, the integration of sophisticated text processing methods, strategies for managing unbalanced data, and thorough assessment procedures improves the system's dependability and resilience. The field of machine learning-based social media job posting detection is dynamic and developing, with both benefits and limitations. Considerable progress has been achieved in creating efficient detection systems through the investigation of many methods such natural language processing (NLP), machine learning algorithms, feature engineering, anomaly detection, and model evaluation .It is imperative to recognize the constraints and opportunities for enhancement inherent in current models and methodologies. The continuous need for research and innovation is highlighted by difficulties like the dependence on labeled training data, the dynamic nature of fraudulent tactics, context sensitivity, scalability issues, overfitting risks, computational complexity, interpretability concerns, and robustness against adversarial attacks or data drift. In order to address these issues, future developments in fake job detection are probably going to include developing more resilient and flexible machine learning models, producing representative and diverse datasets, incorporating cutting-edge natural language processing techniques, implementing real-time monitoring and update mechanisms, and improving the interpretability and transparency of detection systems.

REFERENCES

[1] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection,‖ 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 naive Bayes classifier, ‖ no. January 2001, pp. 41 – 46, 2014.

[3] D. E. Walters, ―Bayes’ s Theorem and the Analysis of Binomial Random Variables, ‖ Biometrical J., vol. 30, no. 7, pp. 817 – 825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, ―Multilayer perceptrons for classification and regression, ‖ Neurocomputing, vol. 2, no. 5 – 6, pp. 183 – 197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers, ‖ Multi . 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 Data Mining, ‖ 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, and O. 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.