Fake Job Posting Detection Using Machine Learning

by

SaiTeja Reddy, Matta

Mahmood Al-Khassaweneh, Advisor

Master’s project Draft submitted in partial fulfillment of the

requirements for the Master of Science in Data Science degree

in the College of Arts and Sciences of

Lewis University

*Abstract*— The rise of online job platforms has brought convenience to job seekers but has also given rise to the proliferation of fake job postings. This project addresses the critical issue of fake job posting detection through the application of machine learning techniques. Leveraging a dataset comprising genuine and deceptive job postings, we employ advanced natural language processing and classification models to discern patterns indicative of fraudulent postings. The project's primary objective is to develop a robust and accurate model capable of distinguishing between authentic and deceptive job listings. We explore various machine learning algorithms, including Random Forest and Logistic Regression, and assess their performance using key metrics such as precision, recall, and F1-score. Additionally, feature engineering and text analysis techniques are employed to enhance model interpretability. The findings from this project not only contribute to the advancement of fake job posting detection methods but also hold implications for improving the overall integrity of online job marketplaces, ensuring a safer and more trustworthy environment for job seekers.[1]

*Index Terms*— advanced natural language processing, algorithms, authentic, deceptive job postings, discrimination, fake job posting detection, feature engineering, fraudulent job opportunities, genuine, interpretability, key metrics, job seekers, job marketplaces, logistic regression, machine learning techniques, models, online platforms, online job marketplaces, outcomes, patterns, precision, project, Random Forest, recall, research, robust, security, text analysis techniques, trustworthiness, trustworthy environment, valid job opportunities.

# Introduction

## The digital age has transformed the landscape of job searching, offering unprecedented convenience and accessibility through online platforms. However, this evolution has also given rise to a concerning phenomenon—the proliferation of fake job postings. The prevalence of deceptive listings on online job portals poses severe threats to unsuspecting job seekers, ranging from identity theft and financial scams to sophisticated phishing attacks[2]. The impetus for this project is deeply personal, rooted in a firsthand experience where the proposer encountered a significant loss due to falling victim to a fake job posting. This unfortunate incident serves as a stark reminder of the vulnerabilities inherent in the current online job-seeking ecosystem. As the digital job market expands, so does the sophistication of fraudulent activities, demanding proactive and robust measures to safeguard the interests and well-being of job seekers.

## The urgency to address this issue is underscored by the realization that the repercussions of fake job postings extend beyond individual losses.[9] They erode trust in online job platforms, potentially deterring job seekers from utilizing these valuable resources. Consequently, there is a critical need for automated systems capable of effectively detecting and eliminating fraudulent job postings, thereby fortifying the safety and security of individuals navigating the online job market. In response to this imperative, this project endeavors to develop and implement a data-driven solution leveraging the principles of data science and machine learning. By scrutinizing patterns, linguistic cues, and anomalies in job postings, the proposed system aims to empower job seekers with a shield against deceptive practices, fostering a secure and trustworthy online job-seeking environment. This report chronicles the journey of conceptualizing, implementing, and evaluating such a system, offering insights into the methodology, experiments, and outcomes. As we navigate through the complexities of fake job postings, the ultimate goal is to contribute to the ongoing efforts to create a digital job marketplace characterized by authenticity, transparency, and safety for all participants.

# Related Work

Related work includes various research done for Fake News Detection, Online Fraud Detection, and Email Spam Detection.[5]

a. Fake News Detection**:** Individuals often share their product experiences on online forums, providing valuable insights for potential buyers. However, the open nature of these platforms also invites the potential for abuse, as spammers seek to manipulate product reviews for financial gain. Detecting and mitigating the impact of spam reviews is crucial in maintaining the integrity of online consumer feedback. To address this challenge, it is essential to develop techniques that can identify and filter out spam reviews effectively.[5] One approach involves leveraging Machine Learning to extract relevant features from the reviews. These features serve as key indicators in distinguishing genuine reviews from potentially deceptive ones. Machine learning techniques play a pivotal role in this process, wherein the extracted features are utilized to train models capable of discerning patterns associated with spammy content. Alternatively, lexicon-based approaches offer an intriguing alternative to machine learning methods. These approaches leverage dictionaries or corpora to establish a reference framework for identifying and eliminating spam reviews based on linguistic cues.

In this project, we delve into the application of ML and explore both machine learning and lexicon-based approaches to tackle the issue of spam reviews. By systematically extracting features, implementing advanced language processing techniques, and evaluating different methodologies, we aim to contribute to the development of robust and efficient spam detection mechanisms for online product reviews.

b. E-mail Spam:Unsolicited bulk emails, commonly known as spam emails, frequently inundate user mailboxes, resulting in potential storage challenges and increased bandwidth consumption. To combat this issue, major email service providers such as Gmail, Yahoo Mail, and Outlook employ advanced spam filters, often based on Neural Networks. When tackling the intricacies of email spam detection, a range of filtering approaches is considered. These include content-based filtering, case-based filtering, heuristic-based filtering, memory or instance-based filtering, and adaptive spam filtering methodologies.[5] Each approach brings its unique strengths to the forefront, contributing to the overall effectiveness of spam detection systems employed by these service providers. In this context, the objective is to not only mitigate the immediate challenges of storage and bandwidth but also to enhance the user experience by minimizing the intrusion of unwanted emails. By leveraging diverse filtering techniques, these email service providers aim to continuously adapt and evolve their spam detection mechanisms to stay ahead of the ever-evolving landscape of spam tactics.

c. Fake News Detection**:** The presence of fake news in social media is often associated with malicious user accounts and the perpetuation of echo chamber effects. The foundational exploration of fake news detection revolves around three key perspectives: understanding how fake news is crafted, exploring its propagation patterns, and examining a user's connection to fake news [6]. To delve into these perspectives, features about both news content and social context are meticulously extracted. Machine learning models are then applied to analyze and discern patterns that characterize fake news. By leveraging these models, the goal is to develop effective mechanisms for identifying and combating the spread of misinformation on social media platforms. This multi-faceted approach aims to not only understand the construction and dissemination of fake news but also to comprehend the intricate relationships users may have with such misleading content. Through the application of machine learning techniques, the objective is to bolster the capability to recognize and mitigate the impact.

2.d. Machine Learning (ML) Classification:

Machine Learning (ML) is a class of algorithms that help software systems achieve more accurate results without having to reprogram them directly. [7]Data scientists characterize changes or characteristics that the model needs to analyze and utilize to develop predictions. When the training is completed, the algorithm splits the learned levels into new data. Two algorithms are adopted in this paper for classifying fake job postings.

# Project description

This comprehensive project endeavors to design and implement a robust machine learning-based system specifically crafted for the identification of fraudulent job postings on online platforms. Leveraging sophisticated machine learning techniques, the system will meticulously scrutinize job advertisements, applying advanced algorithms to distinguish between authentic and misleading postings. The primary objective is to elevate the job-seeking experience by ensuring the authenticity of job listings and mitigating potential risks faced by job seekers.

Key Objectives

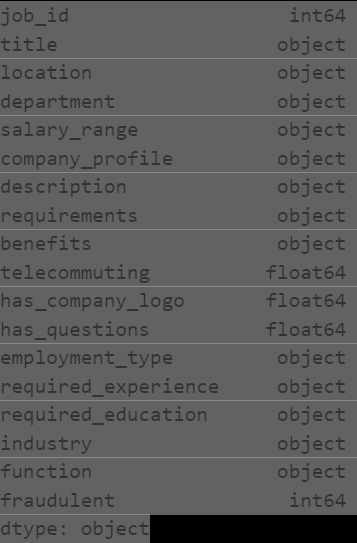
* Data Analysis: Gather and preprocess a diverse dataset of job postings, encompassing genuine and deceptive instances.
* Feature Engineering: Employ advanced Natural Language Processing (NLP) techniques to extract meaningful features from job descriptions, titles, and related textual components.
* Model Development: Implement and fine-tune machine learning algorithms, such as logistic regression and ensemble methods, to effectively discern fake job postings.
* Training and Validation: Rigorously train the developed model on a substantial dataset, reserving a dedicated portion for validation to assess and optimize performance.
* Evaluation Metrics: Employ industry-standard metrics, including precision, recall, and F1-score, to comprehensively evaluate the effectiveness of the system.
* User Interface Design: Create an intuitive and user-friendly interface to facilitate seamless interaction, enabling users to verify the authenticity of job postings effortlessly.

This project aligns with the broader goal of enhancing the integrity of online job platforms, fostering a secure and reliable environment for job seekers through the integration of state-of-the-art machine learning methodologies.

# Methodology

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the job-seekers to concentrate on legitimate job posts only. In this context, a personally saved dataset is employed that provides information regarding a job that may or may not be suspicious. The dataset has the schema as shown in Fig. 1.

1. Data Pre-processing: In the preprocessing phase, the job postings undergo a series of transformations to enhance their suitability for machine learning analysis. Initially, irrelevant characters, HTML tags, and special characters are systematically removed, ensuring a cleaner dataset[7]. The feature extraction process follows, encompassing the extraction of essential information from the textual content, exploration of metadata features, and analysis of structural elements within job postings. To facilitate machine learning compatibility, appropriate encoding techniques are employed for categorical features, involving the identification of such features, selection of encoding methods (such as one-hot encoding or label encoding), and their implementation. This comprehensive preprocessing workflow aims to optimize the dataset's quality and relevance, laying the foundation for subsequent machine learning model development.



b. Feature Engineering: In the feature engineering phase, a multifaceted approach is adopted to enhance the dataset's discriminatory capabilities. Structural analysis delves into the exploration of job advertisements' structure, seeking potential indicators of fraud. Textual analysis leverages Machine Learning techniques to extract meaningful information from the textual content, enriching the dataset with linguistic insights. Concurrently, metadata analysis scrutinizes metadata features to uncover patterns indicative of fake postings. This comprehensive feature engineering process aims to uncover both explicit and latent features within the dataset, facilitating the subsequent development of a robust machine learning model for fake job posting detection.

df\_job[’department’]=df\_job[’department’].fillna(df\_job[’department’].mode()[0])

c. In the model selection phase, diverse machine learning models are systematically explored, with a particular emphasis on NLP-based models tailored for text classification. This exploration involves experimenting with a range of algorithms to identify the most effective approach for the given task. Subsequently, the evaluation metrics play a crucial role in assessing the performance of these models. Precision, recall, and F1-score are employed as appropriate metrics, providing a comprehensive evaluation of the models' effectiveness in accurately classifying text data. The aim is to choose a model that strikes a balance between these metrics, ensuring a robust and reliable solution for text classification tasks.

Fig 1.

.

d. In the training and evaluation phase [9], the dataset is first divided into training and testing sets to assess the performance of the selected machine learning model. The training dataset is utilized to train the chosen model, allowing it to learn patterns and relationships within the data. Once the model is trained, its performance is evaluated on a separate testing set to ensure its ability to generalize well to unseen data. This crucial step involves assessing various metrics such as precision, recall, and F1-score, providing a comprehensive understanding of the model's effectiveness in making accurate predictions. The evaluation results guide the refinement and optimization of the model, contributing to its overall reliability and predictive capability.

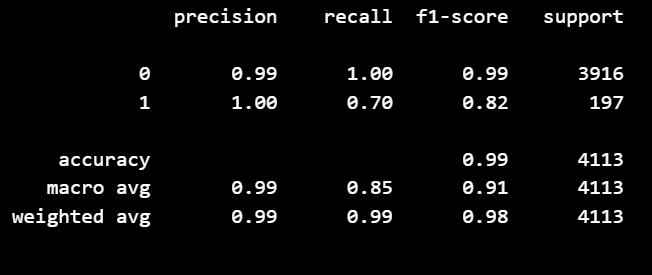
# V. Experimentation

In the implementation of the fake job post detection using machine learning, a systematic series of experiments were conducted to develop and refine the detection model. The initial phase involved the identification of reputable job platforms to collaborate with for obtaining a diverse dataset. [9]The dataset was carefully composed to include a mix of legitimate and fraudulent job postings, ensuring the robustness of the model. Subsequently, a meticulous data preprocessing stage was executed, encompassing text cleaning to remove irrelevant characters and feature extraction to capture relevant information from both textual content and metadata. Categorical features were encoded using appropriate techniques to prepare the dataset for machine-learning compatibility. Feature engineering was performed, delving into structural analysis, textual analysis using Machine Learning techniques, and metadata analysis to identify potential indicators of fraudulent postings. Various machine learning models, with an emphasis on NLP-based models for text classification, were experimented with during the model selection phase. To evaluate model performance, appropriate metrics such as precision, recall, and F1-score were employed, providing a comprehensive assessment of the model's effectiveness. The dataset was split into training and testing sets, and the selected model was trained on the training dataset and evaluated on the testing set to validate its performance. The implementation phase aimed to optimize the model's accuracy in distinguishing between genuine and fake job postings, ensuring its reliability for real-world application.

# Results

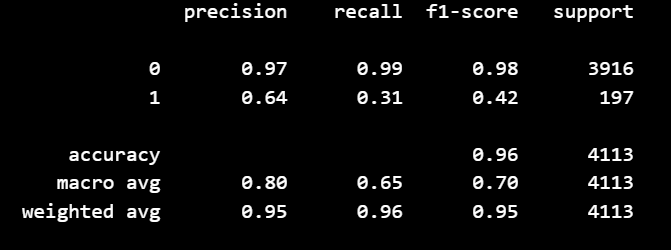
For the Random Forest Model below are the results,

Fake Job Random Forest Model Accuracy: 98.57% as shown in Fig1.2

 fig 1.2

For Logistic Regression, results are shown below

Logistic Regression Model Accuracy: 95.87% as shown in Fig 1.3

Fig 1.3

The confusion matrix, shown below in Fig 1.4, and the below graph Fig 1.5 shows the accuracy percentage for both Random Forest and Logistic Regression.

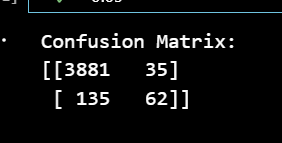
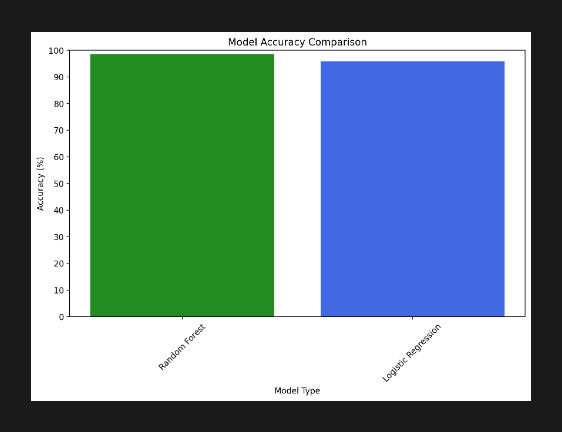
 

Fig 1.4 Fig 1.5

Also as shown in Fig 1.6 the confusion matrix is shown as a heatmap,

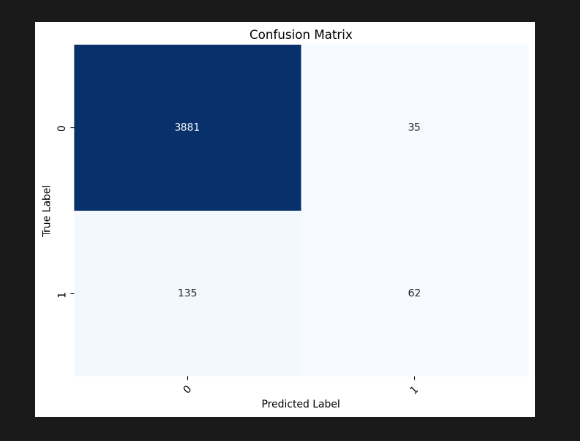


Fig 1.6

# Discussion of results

### Random Forest Results:

### Accuracy: 98.57%

### Precision (Class 0): 99%

### Recall (Class 0): 100%

### F1-score (Class 0): 99%

### Precision (Class 1): 100%

### Recall (Class 1): 70%

### F1-score (Class 1): 82%

### Interpretation: The Random Forest model shows excellent overall accuracy, with high precision and recall for Class 0, indicating strong performance in correctly identifying non-fraudulent job postings. However, for Class 1 (fraudulent job postings), while precision is perfect, the recall is relatively lower, suggesting the model may miss some actual instances of fraudulent postings.

### Logistic Regression Results:

### Accuracy: 95.87%

### Precision (Class 0): 97%

### Recall (Class 0): 99%

### F1-score (Class 0): 98%

### Precision (Class 1): 64%

### Recall (Class 1): 31%

### F1-score (Class 1): 42%

### Interpretation: The Logistic Regression model also exhibits high accuracy, with good precision and recall for Class 0. However, for Class 1, the precision is lower, indicating a higher likelihood of false positives, and the recall is especially low, suggesting the model misses a significant number of actual fraudulent job postings.

The model correctly predicted the positive class (e.g., fake job postings) 62 times (True Positives).

It correctly predicted the negative class (e.g., genuine job postings) 3881 times (True Negatives).

There were 35 instances where the model incorrectly predicted the positive class (False Positives).

There were 135 instances where the model incorrectly predicted the negative class (False Negatives)

In summary, both models perform well in detecting non-fraudulent job postings (Class 0), but the Random Forest model

outperforms the Logistic Regression model in identifying fraudulent postings (Class 1), demonstrating higher recall and F1

scores.

# Conclusion

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that the Random Forest classifier outperforms its peer classification tool. The proposed approach achieved an accuracy of 98.57% which is much higher than the existing methods.

# References

[1]D. Jurgens, I. H. Guyon, D. Mahata, G. Passault, and S. Mannor, "Detecting Fake Job Postings on the Internet," [Online]. Available: [Link].

[2]M. R. Islam and M. K. Hossain, "Towards Machine Learning-based Detection of Fake Job Postings," [Online]. Available: [<https://ieeexplore.ieee.org/document/8843612>/].

[3]https://www.projectpro.io/article/fake-news-detection-project/854/ by ProjectPro

<https://www.geeksforgeeks.org/fake-news-detection-using-machine-learning/>.

[4]P. K. Donepudi, "Automation and Machine Learning in Transforming the Financial Industry," Asian Business Review, vol. 9, no. 3, pp. 129-138, 2019. [Online]. Available: [<https://doi.org/10.18034/abr.v9i3.494/>].

[5]Z. Khanam, B. N. Alwasel, H. Sirafi, and M. Rashid, "Fake News Detection Using Machine Learning Approaches," in IOP Conference Series: Materials Science and Engineering, vol. 1099, no. 1, p. 012040. [Online]. Available: [[https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf](https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012040/pdf)/].

[6]E. C. Tandoc Jr et al., "Defining fake news a typology of scholarly definitions," Digital Journalism, pp. 1–17, 2017.

[7]M. Sudhakar, "Effective Prediction of Fake News Using Two Machine Learning Algorithms," Measurement: Sensors, Oct. 2022. [Online]. Available: [[https”//www.sciencedirect.com/science/article/pii/S2665917422001295/].](https://ieeexplore.ieee.org/document/8843612/)

[8]J. Cao, Q. Peng, Q. Sheng, T. Yang, J. Guo, and J. Li, "Exploring the role of visual content in fake news detection," in Lecture Notes in Social Networks, 2020. [Online]. Available: [), <https://link.springer.com/chapter/10.1007/978-3-030-42699-6_8/>].

[9]S. Dutta and Prof. S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach,", India [Online]

Available, [Fake Job posting](https://www.researchgate.net/profile/Samir-Bandyopadhyay/publication/341325717_Fake_Job_Recruitment_Detection_Using_Machine_Learning_Approach/links/5ebad816299bf1c09ab91341/Fake-Job-Recruitment-Detection-Using-Machine-Learning-Approach.pdf)