LinkedIn: Fraudulent Job Detection

Anusha Ravichandran
University of California, San Diego

Pooja Sounder Rajan
University of California, San Diego

Abstract

Fraudulent job postings on LinkedIn have emerged as a growing threat, driven by advancements in technology and shifts toward remote work. These scams compromise the safety and trust of online professional platforms, often preying on job seekers' vulnerability during their employment search. This research introduces an automated system to detect and analyze fraudulent job advertisements, emphasizing the integration of machine learning (ML) and natural language processing (NLP) techniques. The proposed system evaluates job postings based on linguistic patterns, recruiter behaviors, and contextual indicators of legitimacy. Through feature extraction and classification algorithms, the system aims to distinguish between genuine opportunities and scams effectively. The findings from this study highlight key patterns of fraudulent activity and propose a scalable framework for fraud detection that can be implemented across various online platforms. By offering actionable insights, the system not only safeguards users from potential scams but also contributes to maintaining the integrity of the digital job market.

1 Introduction

The digital era has revolutionized how individuals connect, communicate, and pursue opportunities, with platforms like LinkedIn emerging as key enablers of professional networking and recruitment. As of 2023, LinkedIn boasts over 930 million users globally, facilitating the interaction between job seekers and employers across industries and borders. While this digital transformation has created unprecedented opportunities for career advancement, it has also introduced new risks, particularly the proliferation of fraudulent job postings. Reports reveal a 118% increase in job scams in 2023, a trend exacerbated by the rise of artificial intelligence and the expansion of remote work. These scams not only exploit the aspirations of job seekers but also pose significant threats to financial and data security, with victims losing an average of $2,000 and facing potential exposure of sensitive personal information.

Fraudulent job postings undermine trust in professional platforms, compromise user safety, and erode the credibility of the digital job market. Scammers often employ tactics such as vague job descriptions, exaggerated salary offers, and unverified recruiter identities to deceive job seekers, making detection increasingly challenging. This issue is not limited to a specific region, as job fraud cases have surged globally, affecting both developing and developed economies. The growing sophistication of scams, enabled by advancements in technology, highlights the urgent need for effective countermeasures to protect users and restore confidence in online professional platforms.

This paper addresses the critical issue of fraudulent job postings by proposing a comprehensive detection and analysis framework. Leveraging machine learning (ML) algorithms and natural language processing (NLP) techniques, the study aims to identify patterns in scam postings and distinguish them from legitimate job advertisements. By analyzing key indicators of fraudulent activity and employing advanced feature extraction methods, the proposed system not only detects scams but also provides valuable insights into their characteristics. The findings contribute to the development of robust fraud detection systems, fostering a safer and more trustworthy job market for users worldwide.

2 Related Works

The problem of online job scams has garnered increasing attention from researchers in recent years. Several studies have explored various aspects of this issue, employing different methodologies to detect and prevent such scams.

2.1 Machine Learning Approaches

Madhavi et al. (2022) proposed an automated tool using machine learning-based classification techniques to detect fraudulent job posts [4]. Their study compared different classifiers, including Naive Bayes, Decision Tree, K-nearest Neigh-

bor, and Random Forest. Their experimental results indicated that ensemble classifiers, particularly Random Forest, outperformed single classifiers in detecting job scams. This work provides a foundation for our approach, as we also intend to leverage machine learning algorithms for scam detection.
The growing prevalence of job scams has led to various studies investigating their increase and the methods used by scammers. According to Morris (2024), job scams more than doubled in 2023, highlighting the growing issue in online recruitment platforms [3]. The study discussed in Fortune indicates that scammers are increasingly targeting individuals through fraudulent job postings, with substantial growth in such incidents in recent years.

In a broader context of fraud detection, Chen et al. (2024) conducted a comprehensive literature review on financial fraud detection through machine learning techniques [2]. While their focus was not specifically on job scams, their findings on the effectiveness of various machine learning models in detecting fraudulent activities are relevant to our research. They highlighted the trend towards using real datasets and the importance of feature engineering in improving model accuracy.

2.2 Analysis of Scam Characteristics

A study by Heimdal Security (2023) analyzed 2,670 social media posts and comments related to employment scams [1]. Their findings revealed that the finance industry accounted for 35.45% of job scams, followed by IT at 30.43%. They also identified common red flags, such as requests for upfront payments (25.08%) and phishing attempts (18.81%). These insights into the characteristics of job scams will inform our feature selection and model development process.

Our research builds upon these prior works, aiming to develop a comprehensive, LinkedIn-specific job scam detection system that leverages the latest advancements in machine learning while addressing the unique characteristics of job scams on professional networking platforms.

3 Hypotheses

To explore the characteristics and distribution of fraudulent job postings on LinkedIn, we hypothesize the following:

H1: Fraudulent job postings will have description lengths that closely mirror legitimate job postings. This is due to scammers attempting to mimic the appearance of genuine listings by keeping their job descriptions at a similar length to those of legitimate posts in order to avoid suspicion.

H2: Major metropolitan areas, particularly New York, Chicago, and Houston, will show a significantly higher concentration of fraudulent job postings compared to smaller cities. This is because scammers tend to target large job markets with higher population densities, maximizing their chances of reaching a larger pool of potential victims.

4 Experimental Design

In this study, data was collected from multiple sources to develop a machine learning model capable of detecting fraudulent job postings on LinkedIn. Due to platform restrictions preventing the direct scraping of LinkedIn data, two Kaggle datasets were used. The first dataset contained generic fraudulent job listings, while the second dataset included job postings similar to those found on LinkedIn. A classifier was first trained on the Kaggle datasets to identify fraudulent listings, and this classifier was subsequently tested on LinkedIn-style job postings to generate a new labeled dataset. This dataset was then used to train a specialized NLP model for detecting LinkedIn-specific job scams.

In addition to the Kaggle datasets, a survey was conducted over a one-week period to gather insights directly from users regarding their experiences with job scams. The survey aimed to identify common red flags and suspicious patterns in fraudulent job postings. The data collected from the survey, combined with the Kaggle datasets, enriched the training process of the model. Ultimately, this experimental design aimed to create an effective fraud detection system tailored to LinkedIn job postings, utilizing machine learning techniques and NLP to provide a valuable tool for identifying and mitigating online recruitment fraud.

5 Data Collection

5.1 Kaggle Datasets

- **D1: Employment Scam Aegean Dataset (EMSCAD)**
  This dataset contains 17,880 real-life job advertisements, manually annotated and classified into two categories: 17,014 legitimate ads and 866 fraudulent ads. The dataset, published between 2012 and 2014, provides a valuable testbed for researchers studying employment scams. It aims to provide the research community with insights into the employment scam problem and has been published in the MDPI Future Internet Journal.

- **D2: LinkedIn Job Postings Dataset**
  This dataset contains a comprehensive record of over 124,000 job postings listed on LinkedIn in 2023 and 2024. Each posting includes attributes such as job title, description, salary, location, application URL, and work types (e.g., remote, contract). It also features additional files with associated benefits, skills, and industries, as well as detailed company data such as headquarters location, number of employees, and follower count. This dataset enables exploration of trends in salaries, benefits,

remote work prevalence, and industry-level job offerings, making it a versatile resource for time-based trend analysis and predictive modeling.

5.2 Google Survey Form

Four main questions were asked to the respondents:

1. Have you come across a LinkedIn job posting that seemed suspicious or fake?
2. What position level did the suspicious job claim to offer?
3. What were some of the red flags in the suspicious job listing?
4. What industry was the suspicious job posting in?

6 Methodology

1. **The Employment Scam Aegean Dataset is used as the primary dataset, with two key features selected for training the model.**
   This dataset provides the foundation for training an initial classifier to distinguish between legitimate and fraudulent job postings.
2. **A MultinomialNB classifier is trained on the transformed dataset to classify job postings as legitimate or fraudulent.**
   The model leverages features from the dataset to build a probabilistic classifier suitable for text-based classification tasks.
3. **The trained MultinomialNB model is applied to a new dataset of LinkedIn job postings to detect potentially fraudulent listings.**
   Predictions from this model allow the identification of suspicious job postings in the LinkedIn dataset.
4. **A Logistic Regression classifier is trained on the LinkedIn dataset with added fraud predictions, improving accuracy and addressing class imbalance using data balancing techniques.**
   This step enhances the performance of the classification process by incorporating additional data and mitigating bias caused by imbalanced classes.
5. **The model's performance is evaluated using: A Confusion Matrix to assess accuracy and a Classification Report to analyze precision, recall, and F1-score.**
   These evaluation metrics provide a comprehensive assessment of the model's effectiveness in identifying fraudulent job postings.

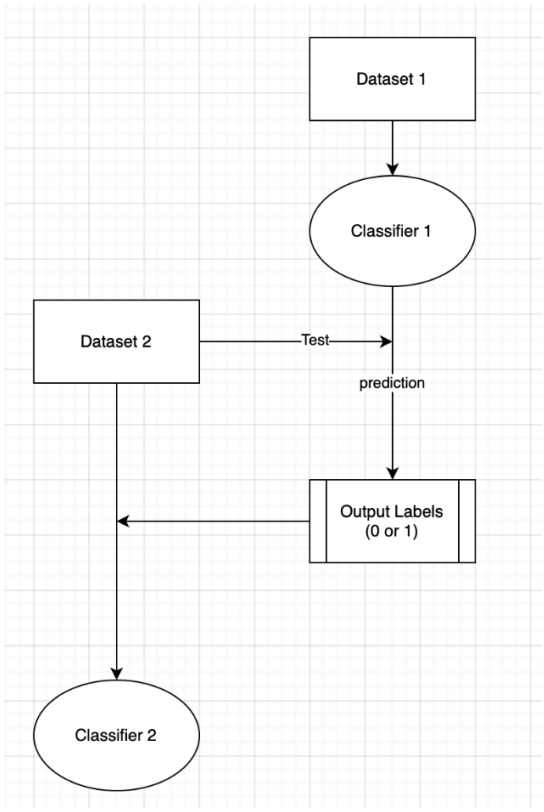

Figure 1: Workflow of the methodology.

7 Results

7.1 Findings

7.1.1 Hypothesis 1: Description Length

The analysis confirmed that fraudulent job postings have description lengths that closely mirror legitimate job postings. This finding suggests that scammers are crafting job descriptions that match the length patterns of genuine listings, making it necessary to look beyond just the description length to identify fraudulent postings.

7.1.2 Hypothesis 2: Geographic Distribution

The results partially supported the hypothesis regarding the concentration of fraudulent job postings in major metropolitan areas. Key findings include:

- **New York, NY** emerged as the most targeted location, with significantly higher numbers of both legitimate and fraudulent postings compared to other cities.
- Major metropolitan areas dominated the top 10 locations for job postings.
- Interestingly, some smaller cities like **Phoenix** and **Charlotte** showed proportionally higher rates of fraudulent

postings relative to their total posting volume.

These findings indicate that while major cities are indeed targets for scammers, the distribution of fraudulent postings is not limited to the largest metropolitan areas, and some smaller cities may also be disproportionately affected.

7.2 Analysis

7.2.1 Key Terminologies and Legitimacy Tactics

- Fraudulent job postings often include terms related to equal opportunity employment and workplace policies, such as:
  - *"sexual orientation," "gender identity," "equal opportunity,"* and *"national origin."*
- These terms appear prominently, signaling an attempt by fraudsters to present their scams as legitimate by mimicking the language of reputable employers.
- Additionally, terms like *"customer service," "full-time,"* and *"veteran status"* are frequently used, further attempting to replicate legitimate professional job descriptions to enhance credibility.

Figure 2: Common words in fraudulent job descriptions

7.2.2 Indicators of Fraud

- The most common scam indicators in fraudulent job postings include:
  - *"Urgent"* and *"guaranteed"*, each appearing over 1000 times.
  - Other terms such as *"easy money"* and *"no experience"* appear less frequently, suggesting fraudsters avoid more obvious red flags.
- These indicators align with previous research on fraud detection in job listings [1, 4].

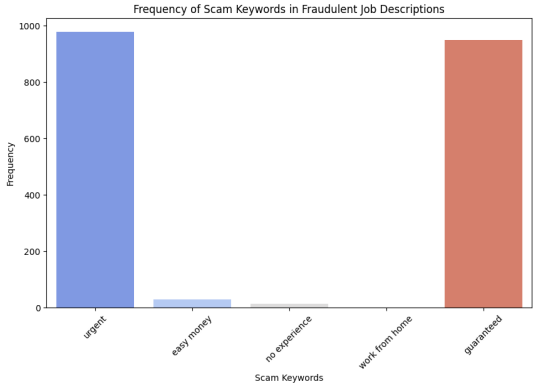

Figure 3: Frequency of Scam Keywords

7.2.3 Geographic Targeting of Fraudulent Postings

- **New York, NY** emerges as the city most targeted by fraudulent job postings, followed by major metropolitan areas like Chicago and Houston.
- Smaller cities such as **Phoenix** and **Charlotte** exhibit disproportionately high rates of fraudulent postings relative to their overall posting volume, which may suggest that scammers are targeting less competitive markets for increased victim engagement.

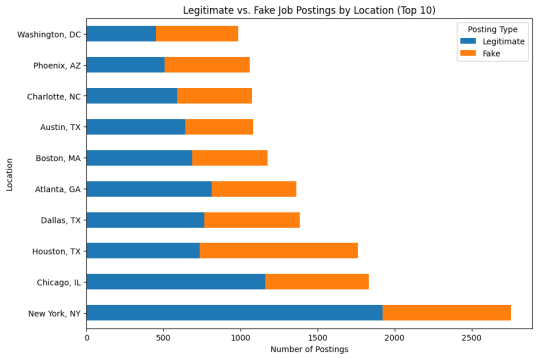

Figure 4: Location-wise Postings

7.2.4 Mimicking Legitimate Job Descriptions

- Fraudsters often craft job descriptions that closely mirror the length patterns of legitimate job postings.
  - This suggests that while description length is a useful feature for identifying fraud, it alone is insufficient for differentiation.
  - Further analysis and features are required to reliably flag fraudulent listings.

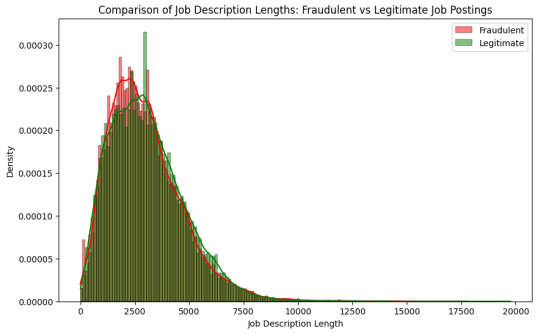

Figure 5: Description Length

7.2.5 Suspicious Contact Information

- The most common form of contact information in fraudulent job postings is suspicious phone numbers, particularly toll-free numbers (e.g., *1-888*), which are more prevalent than email addresses.
- The lack of email addresses or the use of inconsistent phone numbers are significant red flags.
- This suggests that phone numbers are a key element in identifying fraudulent job listings.

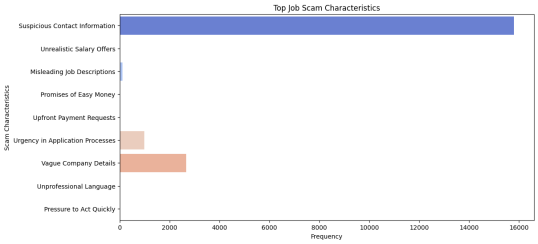

Figure 6: Scam Characteristics

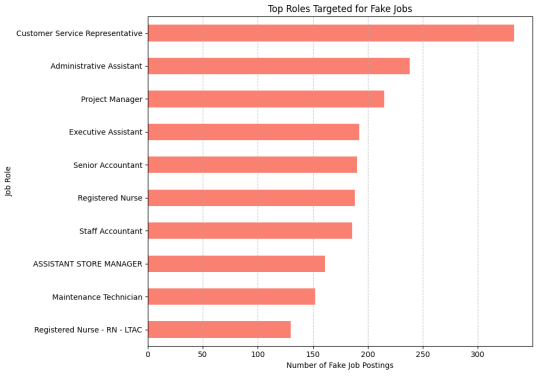

Figure 7: Contact Information in Fraudulent Jobs - A

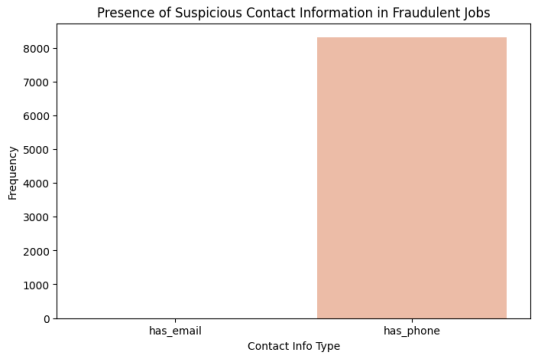

Figure 8: Contact Information in Fraudulent Jobs - B

7.2.6 Targeted Job Roles

- The most targeted role for fraudulent job postings is **Customer Service Representative**, with approximately 300 fraudulent listings, significantly outnumbering other positions.
- Scammers appear to focus on entry-level positions due to their accessibility, but also create sophisticated scams targeting higher-paying roles, particularly in professional and healthcare sectors, to maximize potential gains.

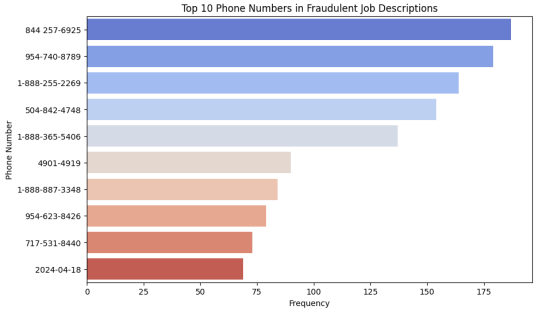

Figure 9: Targeted Roles

8 Limitations

While the methodology and datasets used in this study provide a solid foundation for detecting fraudulent job postings, several limitations were identified:

1. **Limited Generalizability of the Models:**
   The models trained on the Employment Scam Aegean Dataset (EMSCAD) and LinkedIn datasets may not generalize well to job postings outside these datasets. Job postings vary significantly across industries, regions, and platforms, which could reduce model effectiveness in other contexts.

2. **Imbalanced Dataset Issues:**
   Despite efforts to address class imbalance through data balancing techniques, the relatively small number of fraudulent postings compared to legitimate ones could still bias the model, particularly when applied to unseen data.

3. **Feature Selection Constraints:**
   Only a limited set of features were used for model training. Including additional features, such as semantic analysis or context-based embeddings, might improve model accuracy and robustness.

4. **Static Nature of Datasets:**
   The datasets used in this study are static and may not capture the evolving tactics of scammers. Real-time updates and retraining would be required to keep the models effective against new fraud patterns.

5. **Survey Form Limitations:**
   The Google survey used in this study relies on self-reported data, which may be subject to biases, such as misinterpretation of questions or incomplete responses.

9 Future Works

The field of fake job post detection using machine learning is rapidly evolving, and several promising avenues for future research and development exist. Below are key areas warranting further exploration:

9.1 Development of Predictive Tools - Web Application

A crucial next step is the development of a user-friendly web application that allows job seekers to input job postings for scam predictions. This tool would leverage machine learning algorithms to assess the legitimacy of postings based on historical data. By providing an accessible interface, job seekers can quickly verify the authenticity of job listings, reducing their vulnerability to scams.

9.2 Continuous Learning Models - Adaptive Algorithms

Implementing machine learning models that continuously learn from new data is essential for staying ahead of evolving scam tactics. Continuous Training (CT), a component of the MLOps practice model, aims to retrain models automatically and constantly to respond to changes in the data. This approach prevents models from becoming unreliable and inaccurate, ensuring they remain effective in detecting new types of job scams as they emerge.

9.3 Enhanced Data Analysis

Future research should investigate how economic cycles and job market trends influence the prevalence of fraudulent postings. This analysis could help identify peak times for scams and improve response strategies.

By pursuing these future works, we can significantly enhance the effectiveness of fake job post detection systems, ultimately providing better protection for job seekers in an increasingly complex digital landscape.

10 Conclusion

Job seekers need to be vigilant, especially when applying for entry-level positions or roles in customer service. Awareness of red flags, such as suspicious contact information and urgency in job postings, is crucial for users to avoid falling victim to scams.

Furthermore, platforms like LinkedIn and other job portals should implement robust detection systems to protect their users and maintain platform credibility. By integrating advanced machine learning techniques and continuously updating their algorithms, these platforms can effectively combat fraudulent job postings.

Ultimately, a combined effort from both job seekers and platform providers is essential to ensure a safer job search environment in the digital age.

References

[1] Heimdal Security. 2024. *Job Scams Exposed - Insights from Analyzing 2,670 Social Media Posts and Comments*. Retrieved November 4, 2024, from https://heimdalsecurity.com/blog/job-scam-social-media-study/.

[2] Chen, X., Liu, Y., and Zhang, L. 2024. A Comprehensive Review of Financial Fraud Detection Using Machine Learning. *Journal of Financial Technology* 15, 2 (February 2024), 200-210. DOI:https://doi.org/10.22214/jft.2024.2305.

[3] Morris, Chris. 2024. *Job scams more than doubled in 2023*. *Fortune* (June 2024). Retrieved November 4, 2024, from https://fortune.com/2024/06/27/job-scams-increase-in-2023/.

[4] Madhavi, D., Reddy, M. Sri Manisha, Ramya, M., and Sanjana, G. 2022. Detection of Online Employment Scam through Fake Jobs Using Machine Learning Techniques. *International Journal for Research in Applied Science and Engineering Technology* 10, 5 (May 2022), 1234-1240. DOI:https://doi.org/10.22214/ijraset.2022.44384.