Journal of Crime and Justice

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rjcj20

Cyber victimization in hybrid space: an analysis of employment scams using natural language processing and machine learning models

Wenjing Gong, Claire Seungeun Lee, Shoujia Li, Daylon Adkison, Na Li, Ling Wu & Xinyue Ye

To cite this article: Wenjing Gong, Claire Seungeun Lee, Shoujia Li, Daylon Adkison, Na Li, Ling Wu & Xinyue Ye (08 Jan 2025): Cyber victimization in hybrid space: an analysis of employment scams using natural language processing and machine learning models, Journal of Crime and Justice, DOI: 10.1080/0735648X.2024.2448804

To link to this article: https://doi.org/10.1080/0735648X.2024.2448804

Published online: 08 Jan 2025.

JOURNAL OF CRIME AND JUSTICE
https://doi.org/10.1080/0735648X.2024.2448804

Routledge
Taylor & Francis Group

Cyber victimization in hybrid space: an analysis of employment scams using natural language processing and machine learning models

Wenjing Gong[a], Claire Seungeun Lee[b], Shoujia Li[a], Daylon Adkison[c], Na Li[c], Ling Wu[d] and Xinyue Ye[a]

[a]Department of Landscape Architecture and Urban Planning & Center for Geospatial Sciences, Applications and Technology, Texas A&M University, College Station, USA; [b]School of Criminology and Justice Studies, University of Massachusetts Lowell, Lowell, USA; [c]Department of Computer Science, Prairie View A&M University, Prairie View, USA; [d]Department of Justice Studies, Prairie View A&M University, Prairie View, USA

ABSTRACT

Employment scams are a growing and serious issue, increasingly targeting job seekers and exploiting vulnerabilities in the online recruitment process. Current employment scam studies focus primarily on virtual spaces during job searches, neglecting the impact of hybrid spaces (physical and virtual spaces combined) on cyber victimization. This study, leveraging advancements in artificial intelligence, assessed victimization and developed mechanisms to help job seekers reduce their risk of employment scams, considering the physical locations in the real world, as well as the virtual locations described in the job postings. The results show that the consistency of geographic information in the hybrid space of fake job postings is lower than that of legitimate job postings, and there is spatial heterogeneity in the distribution of the physical locations of fake job postings. This consistency, as well as detailed physical location, contributes significantly to the identification and classification of genuine and fake postings. Integrating multiple disciplines, this research enhances understanding of the prevalence, impact, contributing factors, and mitigation strategies associated with cyber victimization during employment. It also contributes to the development of novel methodologies and approaches for detecting, mitigating, and preventing cybercrime.

ARTICLE HISTORY
Received 17 July 2024
Accepted 29 December 2024

KEYWORDS
Cyber victimization; employment scams; hybrid space; machine learning

Introduction

The rapid development and widespread adoption of the Internet, social media, and technology have significantly transformed various aspects of daily life (Chayko 2020; Jadhav and Thepade 2019; Li et al. 2017; Reed 2018). Propelled by its extensive reach, cost-effectiveness, and flexibility, cyberspace has become indispensable for a wide range of activities, including job searching (Kircher 2020). Companies increasingly post job advertisements online, making online job searching an essential activity for job seekers. This trend is particularly evident as it provides a convenient and efficient way to find job opportunities that match their qualifications and interests (Fam, Hui Soo, and Imam Wahjono 2017; Hasan, Salehin, and Islam 2018).

Despite the advantages of online activities, several challenges have emerged, primarily due to advancements in internet technology, information asymmetry, globalization, the trend toward remote work, and the misuse of social media (Ribeiro Bezerra 2021; Richards 2012; Ye et al. 2021). That makes

CONTACT Ling Wu wuxianhaoshen@gmail.com; Xinyue Ye xinyue.ye@tamu.edu Department of Justice Studies, Prairie View A&M University, Prairie View, USA

cyber fraud victimization one of the most pressing issues (Lee 2021, 2022; Wu et al. 2024), with employment scams becoming increasingly prevalent (Vasist and Chatterjee 2023). Employment scams rank among the top 10 types of Internet crime complaints ("Internet Crime Complaint Center" 2023).

In recent years, the Federal Bureau of Investigation's Internet Crime Complaint Center (IC3) issued alerts warning about the increase in employment scams, which cybercriminals may use to obtain applicants' personally identifiable information ("FBI Warns Cyber Criminals Are Using Fake Job Listings to Target Applicants' Personally Identifiable Information" (2021); 2020). Common methods of such scams include online recruitment websites, social media platforms, emails, text messages, and fake company websites (Hijji and Alam 2021; Vidros et al. 2017). These fraudulent activities often exploit the anonymity and vast reach of the internet to deceive job seekers.

Employment scams typically involve the creation of fake job postings offering non-existent job opportunities. These scams can lead to the theft of personal and professional information from job seekers, and more seriously, result in financial loss and emotional distress for the victims (Vidros, Kolias, and Kambourakis 2016). This phenomenon has had a profoundly negative impact on job search experiences, heightening concerns about the risks associated with online job hunting. Consequently, identifying and avoiding fraudulent job postings has become a critical skill for job seekers (Wahid 2023).

Researchers have investigated employment scams, resulting in two distinct strands of literature divided by field of study. On the one hand, from social science perspectives, particularly criminology, previous research analyzes these online employment scams by examining trends, types, and defraud strategies (Cole 2022; Grant-Smith, Feldman, and Cross 2022; Jogalekar and Nanasaheb Jadhav 2022; Ravenelle, Janko, and Cai Kowalski 2022). On the other hand, researchers in data science and other fields have also recognized the importance of this issue and have employed various data science methods to analyze the characteristics of fake job postings in cyberspace (Meneses Silva, Silva Fontes, and Colaço Júnior 2021). Machine learning (ML) and deep learning (DL)[1] models are frequently used in this field. ML is a branch of artificial intelligence (AI) that focuses on creating algorithms and models to help computers improve performance on tasks through experience. DL, a subset of ML, uses neural networks with multiple layers to extract complex features from data. By leveraging ML and DL models, these scholars have developed techniques to classify and identify genuine versus fake job postings (Amaar et al. 2022; Li et al. 2022; Mahbub, Pardede, and Kayes 2022; Mishra, Abdul Rahmaan Ansari, and Mishra 2024). Typically, these studies have significantly contributed to our understanding of how to detect and prevent job scams. However, there is a noticeable gap in the research concerning the impact of hybrid spaces – where physical and virtual spaces are combined – on employment scams, as these spaces are normally neither differentiated nor discussed (Wu et al. 2024).

This consideration of the distinction between physical and virtual spaces is particularly critical when assessing the characteristics and authenticity of job postings. First, the physical location distribution of fake job postings may be uneven across different regions, influenced by numerous factors, including economic conditions, the job market, regulatory environments, and the prevalence of internet access (Malaichamy 2023). Additionally, some employment scams may involve fabricating fake company addresses to enhance the credibility of their recruitment postings (Vidros et al. 2017). Fraudsters employ a variety of strategies to remain undetected. They often operate in areas with minimal police presence, taking advantage of legal grey zones, and frequently change locations to evade detection (Button, Lewis, and Tapley 2009). This practice complicates job seekers' efforts to verify the authenticity of postings since they are not able to identify the actual company. Therefore, it is crucial to understand the consistency between the location information provided in online job postings and the actual physical locations from which these postings originate.

This study aims to address the following questions: (1) What are the main characteristics of fraudulent job postings in hybrid spaces? (2) How consistently does virtual space provide location information that aligns with the actual physical location information, and how does the hybrid geographic information aid in identifying fake postings? (3) How can we enhance our understanding

of employment scams using ML methods? In particular, which ML model achieves the highest accuracy in determining the authenticity of job postings in hybrid space? This research integrates multiple disciplines, utilizing natural language processing (NLP), ML, and Geographic Information System (GIS) to identify key characteristics of fraudulent job postings in hybrid spaces using the Employment Scam Aegean Dataset (EMSCAD). It particularly focuses on the consistency of location information between virtual and physical spaces and develops a classification model capable of accurately predicting the legitimacy of job postings. By highlighting the combination of virtual and physical spaces, this study contributes to the broader field of cybersecurity and helps protect job seekers from falling victim to fraudulent job postings.

The remaining sections are organized as follows. Section 2 reviews related works on employment scams. Section 3 introduces the data, analytical framework, and methodologies. Section 4 summarizes the key research findings. Finally, discussions, conclusions, and limitations are presented in Section 5.

Literature review

Cybercrime, criminology, and geography location

The digital age has transformed crime, redefining traditional concepts of space and place. Geographic location plays a crucial role in cybercrime operations, as the internet's anonymity and global connectivity have expanded criminal opportunities far beyond the physical world (Hall and Yarwood 2024; Wright 2023).

Fraudsters often exploit geographic inconsistencies by operating in regions with minimal law enforcement and weak regulations, posing significant challenges for tracking and prosecuting cybercrime (Williams 2016). Routine Activity Theory suggests that crime thrives where opportunities abound, and risks are low, which explains why cybercriminals target locations with weaker regulatory environments (Clarke and Felson 2017). In cybercrime, discrepancies between an entity's stated geographic location and its actual operating location often signal attempts to evade detection. For instance, fraudsters may post job ads appearing to originate from reputable cities while operating from jurisdictions with weaker cyber laws. These areas, often referred to as 'safe havens' for cybercriminals, provide reduced scrutiny and outdated or poorly enforced legal frameworks, enabling criminals to exploit vulnerabilities (Wall 2007; Yar 2005). This ability to misrepresent geographic locations makes geographic consistency analysis a crucial tool in detecting fraud and cybercrime.

The rise of virtual spaces has amplified this dynamic, providing cybercriminals with even greater opportunities to exploit geographic inconsistencies. Hayward emphasizes that the internet's anonymity and global reach create new avenues for criminal activity, blurring physical boundaries and enabling crimes like fraud, identity theft, and hacking to occur on a global scale (Hayward 2012). The blurring of geographical boundaries introduces geographical inconsistencies that serve as key indicators of cybercriminal behaviour. Recognizing and analysing these hybrid spaces is essential for developing effective regulatory frameworks and fostering international cooperation to combat cybercrime.

This understanding of hybrid spaces and geographic inconsistencies is particularly useful in identifying fraudulent job advertisements. The detection of such ads can be significantly enhanced by integrating insights from both cybersecurity and criminology. Cybersecurity offers powerful tools like machine learning, anomaly detection, and real-time monitoring to flag mismatches between the stated and actual locations in job posts while protecting job seekers' data. Meanwhile, criminology provides valuable insights into the motives and behaviours of fraudsters, explaining why they misrepresent geographic information to evade detection or target vulnerable populations. By combining these approaches, as suggested by Dupont and Whelan (Dupont and Whelan 2021), a more comprehensive fraud detection model can be developed. This interdisciplinary approach

strengthens fraud detection systems by incorporating both behavioural and technical indicators, enabling more effective differentiation between legitimate and fraudulent ads.

Identification of employment scams

Employment and recruitment scams are fraudulent practices that attract potential job applicants to apply for non-existent or deceptive positions. Criminologists understand the phenomenon of (online) employment scams in terms of trends, types, and defraud strategies among others (Cole 2022; Grant-Smith, Feldman, and Cross 2022). This phenomenon has been studied recently primarily in the context of Global North (e.g., the United States (Ravenelle, Janko, and Cai Kowalski 2022), Australia (Grant-Smith, Feldman, and Cross 2022)) and the Global South (e.g., India (Jogalekar and Nanasaheb Jadhav 2022), Nigeria (Obuene et al. 2024)). Despite the issue's breadth, depth, and importance, there remain a growing but limited body of literature from criminology, criminal justice, and social science perspectives addressing it. In contrast, researchers in data and computer science fields have explored the dynamics of online employment scams using public dataset and sophisticated methods. For instance, Vidros et al. made significant contributions to understanding online recruitment fraud explaining the role of Applicant Tracking Systems (ATS) in the recruitment process (Vidros, Kolias, and Kambourakis 2016). Subsequently, this research group defined and described the characteristics of online recruitment fraud, providing and evaluating the first publicly available EMSCAD dataset sourced from real-life systems (Vidros et al. 2017). EMSCAD has since become the primary data source for most research about fake job postings. Additionally, a small number of studies have utilized datasets other than EMSCAD, such as those sourced from Australia (Mahbub, Pardede, and Kayes 2022), Ghana (Dake 2023), and Bangladesh (Tabassum et al. 2021), among other regions.

Many studies have employed various advanced ML, DL, and NLP techniques, using categorical and textual data to identify employment scams and enhance job-seeking security. Some studies have used categorical data and employed ML or neural network models for prediction without involving complex text analysis. For example, several studies have utilized the categorical features of the EMSCAD dataset, such as 'employment type' and 'has company logo', and conducted comparative studies using models like support vector machine (SVM), random forest (RF), and neural networks (Habiba, Khairul Islam, and Tasnim 2021; Swetha et al. 2023).

Some literature has utilized textual data, leveraging NLP techniques and ML for predictions (Pratley and Masbaul Alam Polash 2023). For instance, Amaar et al. used textual data, including variables like location and title, applying feature extraction techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW) (Amaar et al. 2022). These techniques extracted features from textual data, such as splitting the location feature 'New York' into individual words 'New' and 'York', and employed six supervised ML classifiers. In another study, only four textual data variables were used as predictors: company profile, description, requirements, and benefits (Nessa et al. 2022).

Other studies performed data transformation and utilized various data formats within datasets for analysis. For example, one study categorized different data types based on the detail level of the location feature provided, converting data without location information into category 1, data with only country information into category 2, and so on; K-nearest neighbors (KNN), SVM, and decision trees (DT) were then used to identify fraudulent ads (Chiraratanasopha and Chay-Intr 2022). Sofy et al. translated the EMSCAD data from English to Arabic, converted variables such as education level, location, and industry into continuous variables suitable for classification models, and then used a set of different classifiers to detect fraudulent jobs (Sofy, Khafagy, and Badry 2023). They employed RF for weighting and feature selection of both categorical and textual features. Additionally, a study using artificial neural networks (ANN) for identification considered features like location and title to be irrelevant and removed them from the dataset (Naseer et al. 2021).

Despite the use of advanced NLP and ML techniques in previous studies to identify employment scams, significant research gaps remain. Most past research has relied on feature extraction techniques to tokenize the 'location' feature in the dataset or convert it into specific categories, overlooking the analysis of the consistency between physical and virtual location information. This gap is crucial, as analysing this consistency could play a key role in distinguishing between genuine and fraudulent job postings.

Data and methodology

Dataset

EMSCAD is a publicly available dataset that was obtained from Kaggle. It comprises 17,880 genuine job advertisements posted from 2012 to 2014, including 17,014 legitimate and 866 fake ads. It includes a diverse assortment of positions from around the world, including jobs that are to be completed remotely. Table 1 describes the information contained within all 18 presented features in the dataset. Among them, the 'telecommuting', 'has company logo', 'has questions', and 'fraudulent' features are binary values representing true or false, and the additional features are string-based. We removed the affected samples containing missing values before modelling because they comprised only approximately 2% of both legitimate and fraudulent job postings, which is unlikely to introduce significant bias or compromise the integrity of the dataset. Additionally, common imputation methods are inappropriate for location data, as they fail to capture the complexity and categorical nature of geographical information.

Table 1. Descriptive summary of EMSCAD features.

Features	Description
Job ID	The 'Job_ID' feature acts as the index of the dataset, increasing chronologically to 17,880.
Title	The 'Title' feature contains the name of the application being applied for.
Location	The 'Location' feature contains the published location of job postings, typically in a [Country, State, City] format.
Department	The 'Department' feature contains information regarding the specific unit within the company that is responsible for particular tasks, such as Sales, Marketing, or IT.
Salary Range	The 'Salary Range' feature contains monetary information about the advertised job.
Company Profile	The 'Company Profile' feature contains the company's sales pitch and description of services.
Description	The 'Description' feature contains the tasks required for the completion of the job, occasionally including locations, benefits, and technology that will be used.
Requirements	The 'Requirements' feature contains the necessary qualifications for the job, such as MS degrees, experience with certain technologies, and licenses.
Benefits	The 'Benefits' feature contains information regarding company culture, competitive pay, and other job perks.
Telecommuting	True or False value indicating if the job can be completed remotely.
Has Company Logo	True or False value indicating if the job posting has a company logo for their organization
Has Questions	True or False value indicating if screening questions are present.
Employment Type	The 'Employment Type' feature indicates if the job is full-time, part-time, contract, temporary, or other.
Required Experience	The 'Required Experience' feature contains the minimum experience level required for the job, such as entry-level, associate-level, or Mid-Senior level
Required Education	The 'Required Education' feature contains the minimum education level required for the job, such as High school or equivalent, Bachelor's Degree, etc.
Industry	The 'Industry' feature contains information describing their type of service, such as Oil & Energy, Accounting, Health Care, etc.
Function	The 'Function' feature contains information that is similar to 'department' and contains values such as IT, Customer Service, Design, etc.
Fraudulent	True or False value indicates if the job posting is fake and possibly malicious.

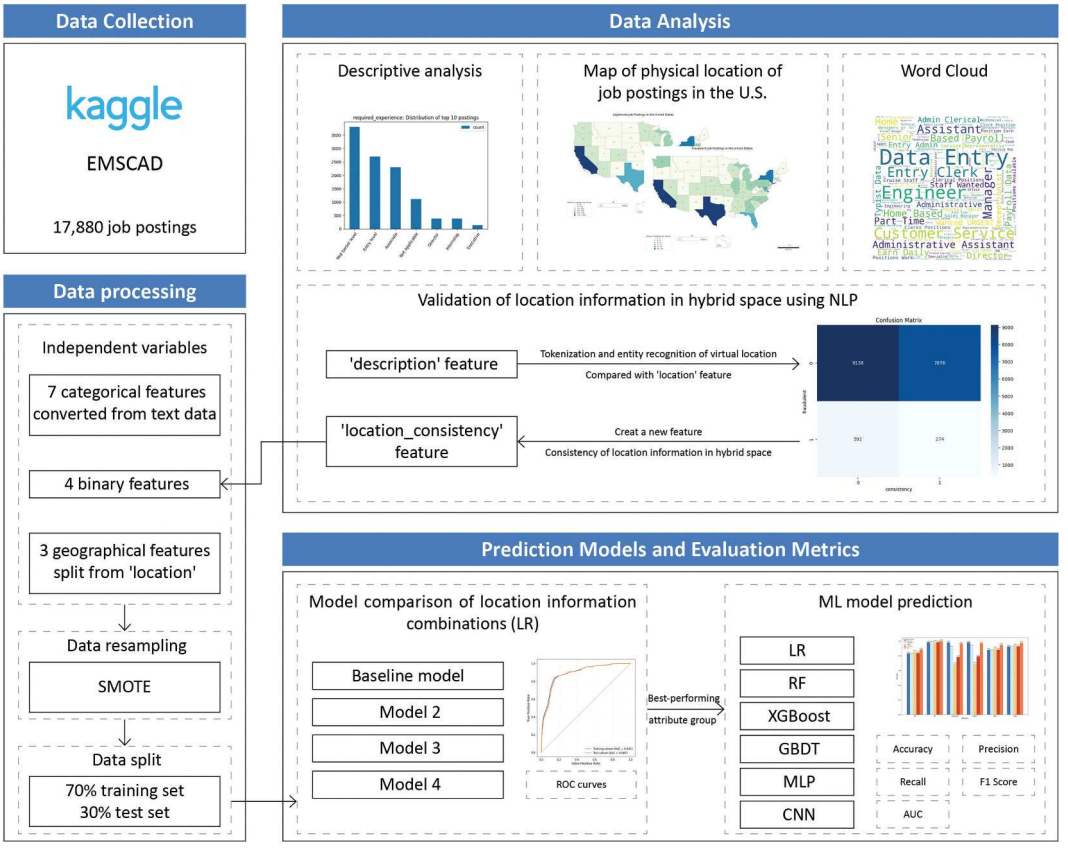

Figure 1. Analytical framework.

Analytical framework

The analytical framework is shown in Figure 1. We first conducted descriptive analysis, including statistical evaluation of various features and geographical mapping of job postings across the U.S., the region with the highest concentration. Keyword analysis was performed using a Word Cloud for preliminary NLP exploration. Then, we compared only the 'description' feature in the dataset with 'location' through NLP processing to create a 'location_consistency' feature, visualized with a confusion matrix.

Before ML prediction, the Synthetic Minority Over-sampling Technique (SMOTE) was used to address the class imbalance, and the dataset was split into 70% training and 30% test sets. Logistic Regression (LR) models, a type of supervised learning algorithm for classification, were implemented via the Scikit-learn package in Python and compared based on different combinations of location information. The best-performing group was identified using Receiver Operating Characteristic (ROC) curves and other detailed metrics. Machine learning models, including LR, RF, eXtreme Gradient Boosting (XGBoost), Gradient Boosting Decision Trees (GBDT), Multilayer Perceptron (MLP), and Convolutional Neural Networks (CNN), were employed. Their performance was evaluated using key metrics such as Accuracy, Precision, Recall, F1 Score, and Area Under the Curve (AUC).

Natural language processing

Word cloud

To facilitate the preliminary exploration of the text data, we employed Word Cloud, a widely used data visualization tool in NLP (Burch et al. 2013; Sinclair and Cardew-Hall 2008), to analyze the terms in the job postings. Word Cloud effectively highlights the frequency and significance of words by visually representing them in varying sizes and colors, with more frequently occurring words displayed in larger fonts. This method effectively captures the prominent terms and their relative importance, aiding in understanding the text's structure.

Tokenization and entity recognition

A text-processing approach was employed to analyze the information within the 'description' feature. The process involved tokenization, a fundamental technique in NLP, to segment the text into meaningful units and identify significant location-related information. Tokenization was performed using the SpaCy library, a powerful NLP tool providing pre-trained models to facilitate various text analyses (Altinok 2021; Vasiliev 2020).

After tokenization, SpaCy's NER capabilities, which are widely used in NLP for accurate and efficient extraction of named entities in large-scale text data (Shelar et al. 2020), were leveraged to extract entities from the text. SpaCy's NER is trusted across various industries and research fields for its ability to accurately identify entities such as locations, organizations, and people, making it a powerful tool for our task (Jiang, Banchs, and Li 2016; Yanti, Santoso, and Hulliyyatus Suadaa 2021Naseer et al. 2021). The focus was on identifying geographical entities, labeled as 'GPE' (Geopolitical Entities). These entities include locations such as cities, countries, and other significant geographical markers.

Validation of location information in hybrid space

The extracted virtual location tokens from the 'Description' were then compared with the physical location in the 'Location' feature in both fraudulent and legitimate postings using a confusion matrix. By assessing whether these two locations matched, we could determine the consistency of the location information in hybrid space. Fraudulent job postings often use fake or misleading location details to appear more legitimate, making this inconsistency a potentially valuable clue for identifying scams. Therefore, consistency between these two sources of location information was assessed to analyze the accuracy and reliability of the post's content. This assessment led to the creation of a new feature, 'location_consistency', which labels the consistency of the location information in hybrid space. This new feature was added to the dataset for further ML prediction.

Machine learning models

Data processing

We decomposed the feature 'location' into three distinct features: country, state, and city. These three features, along with the employment type, required experience, required education, industry, function, salary range, and department, were converted to categorical variables. In addition, the dataset included three original binary variables and a newly created binary variable, location_consistency, which was derived from the NLP analysis. This preprocessing resulted in a total of 14 independent variables for subsequent analysis.

To address the imbalance issue of legitimate and fake job postings in the dataset, we employed the SMOTE, an over-sampling method that generates synthetic samples for the minority class by interpolating between existing minority class samples (Blagus and Lusa 2013; Chawla et al. 2002). SMOTE is widely used in machine learning for its effectiveness in handling class imbalance across various domains, including fraud detection, and has been applied in numerous studies that utilized the same dataset as ours (Amaar et al. 2022; Bhatia and Meena 2022; Chiraratanasopha and Chay-Intr

2022). This approach helps in balancing the class distribution and enhancing the performance and robustness of classification models.

Model architecture

The LR model was initially employed to compare different attribute groups, with particular attention given to the contribution of geographic information in the hybrid space to the model's predictions. Subsequently, we developed three models: one incorporating the location_consistency variable, another including the country, state, and city variables, and a third integrating all attribute groups with hybrid location information. The dataset was divided into a 70% training set and a 30% testing set for all the models in this study. To evaluate the performance of the LR models, ROC curves were employed. These curves are useful for understanding the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) (Hoo, Candlish, and Teare 2017). Other detailed metrics such as Accuracy, Precision, Recall, F1 Score, Specificity, and Effect Strength of Sensitivity were also employed to provide a more granular understanding of model performance.

Building on the best-performing attribute group combinations identified by LR, we extended our analysis by employing five additional ML models widely applied in numerous studies: RF, XGBoost, GBDT, MLP, and CNN (Gong et al. 2023; Gong, Rui, and Li 2024). RF is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. XGBoost is an efficient and scalable implementation of gradient boosting that optimizes predictive performance by combining multiple weak learners (decision trees) in an additive fashion. GBDT is another ensemble technique that iteratively adds models to correct errors from previous models, using gradient descent to minimize loss. MLP is a type of feedforward neural network consisting of multiple fully connected layers of neurons, commonly used for classification and regression tasks. Lastly, CNN is a deep learning model that uses convolutional layers to automatically learn hierarchical patterns and features from input data, making it effective for various types of structured data.

Since the selection of hyperparameters significantly impacts ML models' performance and predictive capability, we employed GridSearchCV in conjunction with 5-fold cross-validation. GridSearchCV is a hyperparameter tuning technique that exhaustively searches through a predefined grid of hyperparameters to find the best combination for model optimization, while 5-fold cross-validation divides the dataset into five equal parts to train and validate the model, ensuring that the performance is consistent and less prone to overfitting (Adnan et al. 2022; Chadha and Kaushik 2022). To assess and compare the effectiveness of these models, we utilized a comprehensive set of evaluation metrics, including Accuracy, Precision, Recall, F1 Score, and AUC. These metrics provide insight into different aspects of model performance, such as overall correctness, the ability to correctly classify positive instances, the ability to capture all positive instances, the balance between precision and recall, and the model's ability to distinguish between classes.

Results

Descriptive results

Descriptive statistics play a crucial role in understanding the basic features of the dataset, laying the foundation for more in-depth analysis. Therefore, we performed a descriptive analysis of four key data categories: industries, required education, required experience, and employment types, as well as the physical location of job postings. Examining these data points not only provided an overview of the current state and trends in the job postings but also revealed potential relationships and impacts among different factors.

Category features

The EMSCAD dataset contains 131 different job industries. For more clear and understandable industry information, we recoded the 131 industries into 19 categories based on the North American Industry Classification System (NAICS), which is the standard used by Federal statistical agencies to allow for a high level of comparability in business statistics among the North American countries ("North American Industry Classification System (NAICS) U.S. Census Bureau" 2024). The mapping relationship between NAICS definitions and EMSCAD industry features in this study can be found in Table A1. The industry value counts of job postings after recoding are illustrated in Figure 2(a). The data reveals a significant concentration in the 'Information' industry, which accounts for the majority of job postings with a count of 4,685. Figure 2(b) shows the top 10 industries for fake job postings after recoding. The highest number of fraudulent postings is seen in 'Mining, Quarrying, and Oil and Gas Extraction', with over 100 instances. The information industry also ranks high in fraudulent postings, making up roughly 10% and placing third overall, suggesting a high level of vulnerability to fraud. The high volume of job postings in the Information industry, combined with its reliance on remote and freelance work, maybe the reason that makes it an attractive target for fraudsters who exploit the industry's rapid growth and various job needs. Tables A2 to A4 show the descriptive analysis of required education (Table A2), required experience (Table A3), and employment type (Table A4).

Physical location of job postings

The dataset contains job postings unevenly distributed across the globe, with a very large percentage of job postings in the U.S. and a smaller percentage in other regions. Specifically, fake ads in the U.S. make up about 90% of all fakes worldwide. We therefore performed further detailed mapping of the job posting distribution for the U.S. where they are most concentrated. Figure 3 illustrates the distribution of fraudulent and legitimate job postings across various U.S. states, mainly using GeoPandas and Matplotlib libraries in Python. They were color-coded based on numbers and classified using the natural breaks (Jenks) method.

Texas and California have the highest number of fraudulent postings, ranging from 65 to 150. New York, Maryland, and Florida also show significant numbers, with 64, 33, and 27 respectively, indicating key areas for regulatory focus and job seeker caution. Moderate levels of fraudulent

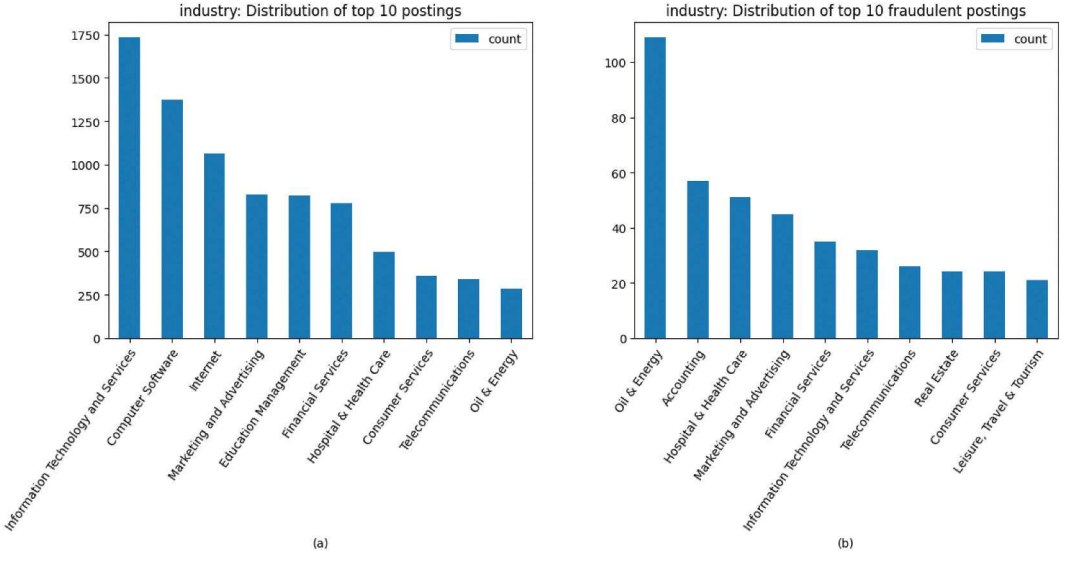

Figure 2. Top 10 industries of (a) posting and (b) fraudulent postings after recoding.

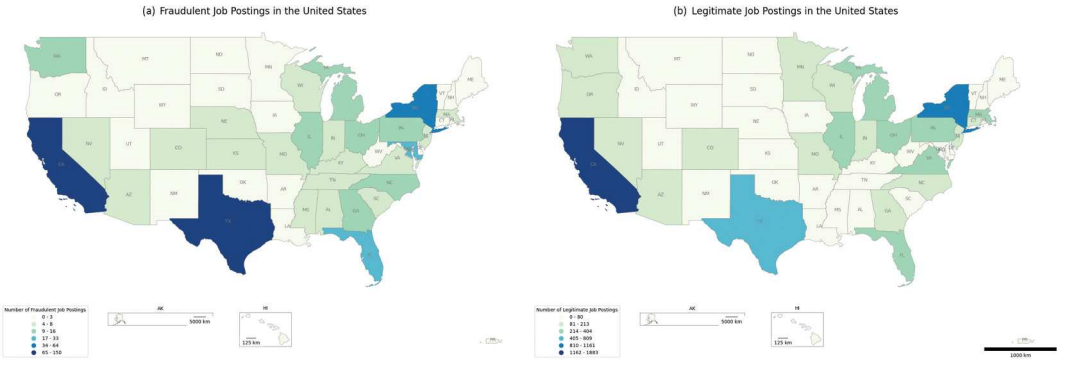

Figure 3. Distribution of (a) fraudulent and (b) legitimate job postings in the United States.

postings (9 to 16) are observed in 7 states, such as Washington, Pennsylvania, and Michigan. Some states like Montana, Wyoming, South Dakota, North Dakota, and Hawaii report minimal to no data on fraudulent job postings, with fewer than 3 cases. This may suggest either genuinely low incidences or potential underreporting.

California shows the highest number of legitimate job postings, with a count of 1883. New York and Texas also report high numbers, with 1161 and 809 respectively. Similar to the fraudulent postings, many states fall into the lower ranges for legitimate postings, with several reporting less than 80. Overall, while both fraudulent and legitimate job postings are more prevalent in states like California and Texas, the spread of fraudulent postings is somewhat more uneven compared to legitimate postings.

Results of natural language processing

Keyword analysis

The word cloud in Figure A1(a) highlights the most frequent terms found in the title of fraudulent job postings. Prominent words include 'Data Entry', 'Engineer', 'Customer Service', 'Assistant', and 'Clerk'. These terms suggest that fraudulent job postings often advertise positions that require minimal qualifications or are typically associated with remote or flexible work environments. The presence of words like 'Home', 'Based', and 'Part Time' further supports the notion that such ads target individuals seeking work-from-home opportunities or part-time employment.

Figure A1(b) displays the word cloud for legitimate job postings. Key terms include 'Customer Service', 'Engineer', 'Teacher', 'Developer', and 'Manager'. Unlike fraudulent job postings, legitimate ads emphasize positions that typically require higher qualifications and specialized skills. The findings also indicate several keywords that are common across both categories, such as 'Customer Service', 'Manager', and 'Assistant'. These roles have wide applicability and high demand, which may make them natural targets for both genuine job ads and scams.

Consistency of geographic information in hybrid space

We analyzed the content within the 'description' label to extract the virtual location information mentioned using NLP. By comparing this virtual location with the physical location in the 'location' label, we identified patterns of consistency and inconsistency in geographic information within hybrid spaces (Figure 4). Overall, about 54.4% of all postings (9,730 out of 17,880) did not match the real-world location.

Specifically, out of 17,014 legitimate postings, 7,876 postings (about 46.3%) had consistent location information, while 9,138 postings (about 53.7%) showed inconsistency. In addition, of the 866 fake postings, 274 postings (about 31.6%) had consistent location information, whereas 592 postings (about 68.4%) were inconsistent. This highlights the significant prevalence of inconsistency in hybrid geographic information, suggesting areas for further investigation and verification.

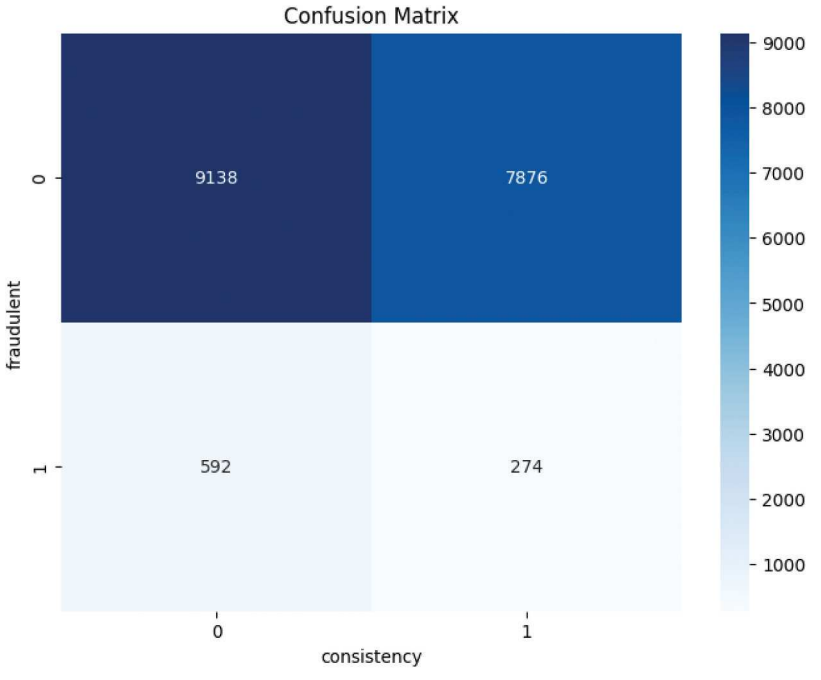

Figure 4. Confusion matrix of geographic information consistency.

Results of machine learning models

Model comparison of geographic information combinations

Figure 5 and Table 2 show the performance of models with different combinations of geographic information. Without the geographic information, the baseline model provided a strong foundational prediction with AUC values over 0.85 for both training and testing cohorts (Figure 5(a)). The baseline model's Accuracy was 0.812, and while Precision (0.796) and Recall (0.838) were fairly balanced, the model had a Specificity of 0.786, indicating some room for improvement in identifying non-fraudulent cases.

Incorporating the location_consistency variable improved the model's performance compared to the baseline (Figure 5(b)). The AUC values for the testing cohort increased to 0.865, and this was reflected in better metrics in Model 2: Accuracy improved to 0.816, Precision to 0.805, and Specificity to 0.799, suggesting that the consistency of location information provided additional predictive power. In addition, the Effect Strength of Sensitivity of 0.826 for Model 2 demonstrates a clear improvement beyond chance performance, indicating that the inclusion of geographic consistency is not sensitive to baseline class proportions and enhances the model's sensitivity to detecting fraudulent ads. When the country, state, and city variables were added to the model, there was also an improvement in both AUC values (Figure 5(c)) and other metrics: Accuracy increased further to 0.826, Recall to 0.848, and Specificity to 0.805.

The model that integrates all attribute groups (attributes in the baseline model, location_consistency, and country, state, and city), as depicted in Figure 5(d), demonstrated the highest AUC values among the four models tested. Specifically, this model achieved an AUC of 0.892 for the training cohort and 0.887 for the test cohort. Correspondingly, Model 4 displayed the best overall performance, with the highest Accuracy (0.832), Precision (0.818), and Specificity (0.811), indicating the superior performance of hybrid geographic information in distinguishing between fraudulent and non-fraudulent job postings compared to the other models. With an Effect Strength of Sensitivity of 0.847, Model 4 clearly shows the strongest improvement beyond random chance, confirming that

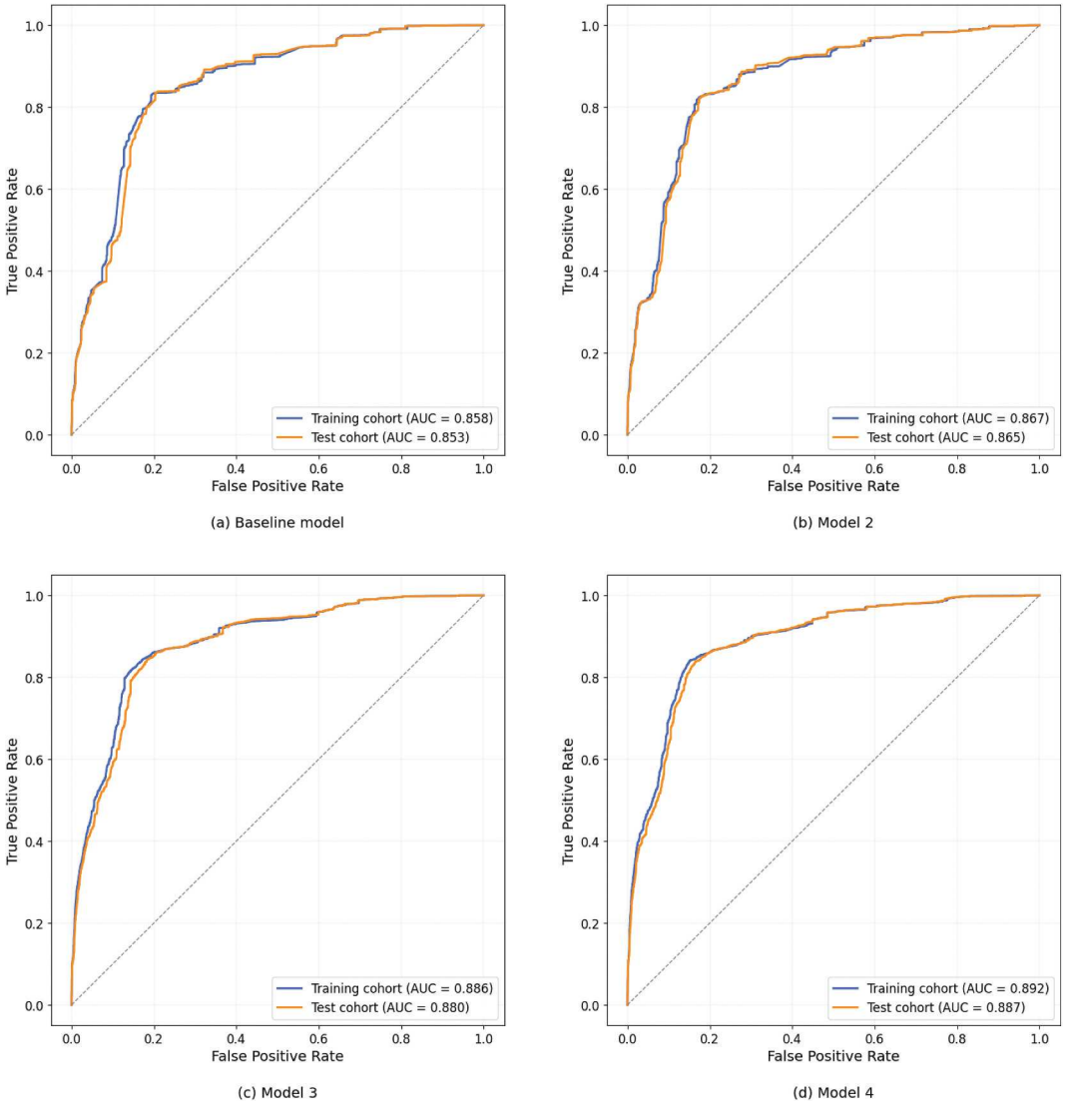

Figure 5. ROC curves of different geographic information combination models.

the integration of detailed geographic information consistently enhances the model's sensitivity while remaining unaffected by baseline class proportions.

Table A5 shows the coefficients and odd ratios of Model 4. The negative coefficient indicates that location consistency has a protective effect against fraud, making an ad less likely to be fraudulent. The odds ratio of 0.472 shows that consistent location information significantly reduces the odds of a job ad being fraudulent, with the odds decreasing by about 52.8% compared to ads with inconsistent location information.

Prediction performance of different models

The performance of various models integrating all geographical information attribute groups is illustrated in Figure 6, which compares their evaluation metrics: Accuracy, Precision, Recall, F1 Score, and AUC. RF demonstrated superior performance across all the metrics, with particularly high scores in Accuracy (0.983), Recall (0.989), and AUC (0.998), indicating its robustness in handling the dataset. This showed its effectiveness in capturing the complexities of the location data in the hybrid space.

Table 2. Prediction performance of models with different geographical location combinations.

Metric	Baseline model	Model 2	Model 3	Model 4
Accuracy	0.812	0.816	0.826	0.832
Precision	0.796	0.805	0.812	0.818
Recall	0.828	0.834	0.848	0.854
F1 Score	0.816	0.818	0.830	0.836
Specificity	0.786	0.799	0.805	0.811
Effect Strength of Sensitivity	0.819	0.826	0.840	0.847

1. Model 2 incorporates the location_consistency variable, Model 3 includes the country, state, and city variables, and Model 4 integrates all hybrid location information variables (attributes added in Model 2 and Model 3) based on the baseline model.
2. Accuracy represents the proportion of correct predictions among the total predictions. Precision measures the proportion of true positives out of all positive predictions, indicating how accurate the model is when predicting positive labels. Recall reflects the proportion of true positives identified out of all actual positives, showing the model's ability to detect positive instances. The F1 Score is the harmonic mean of precision and recall, balancing the two when they are uneven. Specificity represents the proportion of true negatives identified out of all actual negatives, measuring the model's ability to correctly identify negative instances. Effect Strength of Sensitivity measures how much better a model's sensitivity (recall) is at identifying positive cases compared to random guessing, with values closer to 1 indicating strong performance beyond chance.

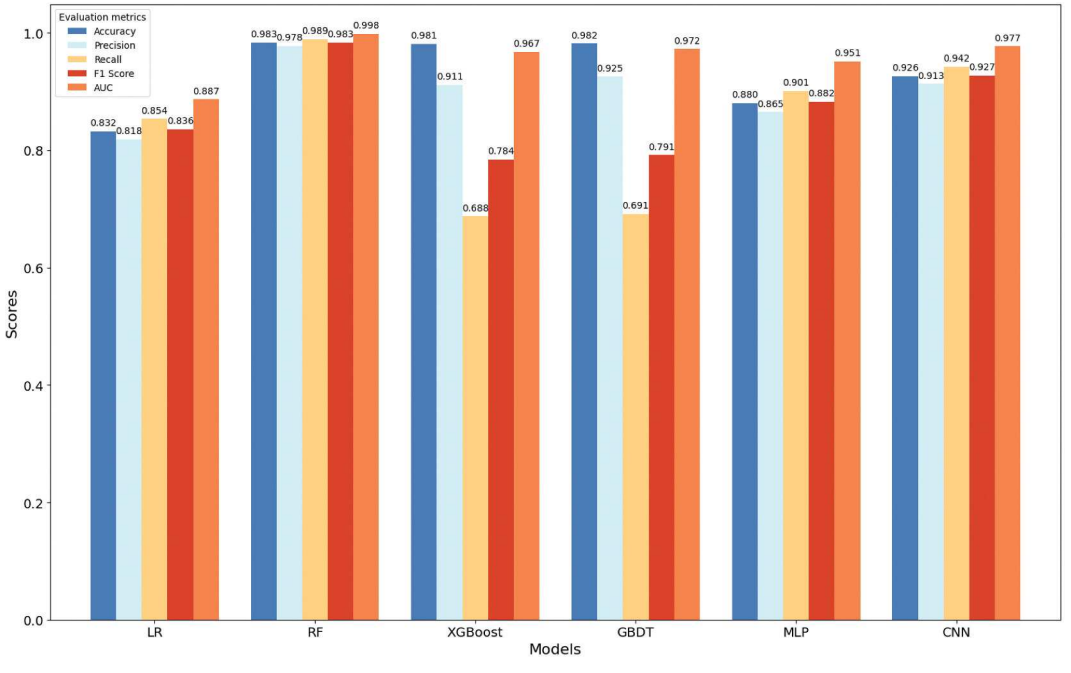

Figure 6. Prediction performance of ML models.

XGBoost also showed competitive performance, particularly in terms of Accuracy (0.981) and AUC (0.982). MLP and CNN presented solid results as well. CNN achieved notable scores with an Accuracy of 0.926, Recall of 0.942, and AUC of 0.977, demonstrating its capability to handle non-linear patterns in the data. In contrast, LR exhibited comparatively lower scores across all metrics, with an Accuracy of 0.832, Precision of 0.818, Recall of 0.854, F1 Score of 0.836, and AUC of 0.887. These results suggest that LR may be less effective in capturing the intricate relationships within the hybrid geographical data compared to more complex models like RF and CNN.

Discussion

The identification of employment scams has become a major problem and presents a challenging task worldwide. To enhance the in-depth understanding of the influence of hybrid spaces, this study explored the possibility of NLP and ML for modeling fake job identification using a public dataset. This study makes three main contributions. First, by analyzing and integrating both physical and virtual geographical information, as well as the consistency between them, the research demonstrated how such data can substantially enhance the performance of prediction models. This improvement in ML model accuracy provides a powerful new insight into safeguarding job seekers from fraudulent job postings and can be beneficial for job listing platforms to implement more robust fraud detection mechanisms.

Second, this multidisciplinary approach, which combines elements of GIS, data science, AI, and cybersecurity, highlights the importance of cross-field collaboration in addressing complex cyber issues. The insights gained from this study pave the way for future research in hybrid spaces, refining and expanding these methods, and offering new strategies from the location information for combating cybercrime and protecting vulnerable populations in the digital age.

Third, the framework established in this study is not limited to employment scams alone. It has the potential to be extended and adapted to other forms of cyber victimization. For instance, online dating scams (Buchanan and Whitty 2014), online fraud (Lee 2021, 2022), e-commerce fraud (Mutemi and Bacao 2024), and other types of cybercrime could benefit from similar approaches that leverage the analysis of location data. By applying this framework, researchers and practitioners can develop more robust models for identifying and preventing various types of online fraud, ultimately contributing to a safer and more secure environment for users across different domains.

The study's results highlight the potential for preventing employment scams. First, job postings that are skilfully crafted – featuring a company logo, a Q&A section, and realistic education and experience requirements – are more likely to be legitimate. In contrast, postings that omit entire sections, promise high monetary rewards with daily earnings, or focus heavily on remote data entry or administrative assistance roles, are more likely to be fraudulent. This pattern suggests that detailed and error-free information serves as an indicator of authenticity, while the lack of such details, combined with unrealistic incentives, presents potential red flags for fraudulent activity.

Second, in the physical space, there is a significant concentration of fraudulent job postings in states like Texas and California, while the Midwest and many Western states show lower incidences. This geographical disparity in the distribution of fraudulent job postings reveals important patterns that can inform both job seekers and regulatory bodies. The concentration in Texas and California suggests these states are particularly targeted, potentially due to their large populations and strong job markets, making them attractive to scammers seeking to exploit job seekers. The imbalances highlight areas where increased caution is necessary. Job seekers in high-incidence areas should be particularly cautious when evaluating job postings and should look for signs of potential fraud, such as inconsistencies in the geographical information provided.

Furthermore, in the context of the hybrid space, the study compared the extracted virtual location information through NLP with the posting's physical location. This method proved to be effective in identifying inconsistencies between the stated virtual and physical locations, thereby aiding ML models, especially the RF model which achieved the best performance, in determining the legitimacy of job postings. The inclusion of detailed consistency and geographic information underscores the critical role of location-related data in the identification process. This approach not only highlights the patterns that are often indicative of fraudulent job postings but also provides an understanding of how scammers manipulate geographical data to deceive job seekers. It demonstrates the potential to serve as a powerful tool for accurately identifying fraudulent job postings, thereby enhancing the reliability and security of job listing platforms.

Policy implications

Geographical inconsistencies are pivotal for identifying fraudulent job ads and informing broader prevention and intervention strategies, particularly in the fields of cybersecurity and criminology. Fraud detection systems can leverage these inconsistencies by flagging transactions or communications that present anomalous geographical data for further investigation. For instance, this approach could be integrated into real-time monitoring systems to (1) flag discrepancies in IP addresses, billing addresses, or physical locations as potential indicators of suspicious behavior; (2) develop fraud detection algorithms that combine geographical data with other behavioral features (e.g., login times and device usage patterns) to identify anomalies indicative of cybercrime; and (3) incorporate geographical inconsistencies into multi-factor authentication (MFA) mechanisms, triggering additional verification if a user's location does not align with the expected region.

Building on these insights, this study offers several key policy implications. First, governments and regulatory bodies could establish policies that require online job boards and platforms to enforce minimum transparency standards, such as displaying verifiable company information, job descriptions, and salary details. Such measures would not only help prevent fraudulent job postings but also enhance accountability. Second, public education campaigns should also be launched to educate job seekers on recognizing common warning signs, such as unrealistic salary offers, vague job descriptions, and promises of high-reward, low-effort work. Third, online job platforms should be encouraged to adopt proactive reporting of suspected fraudulent listings using anomaly detection systems to flag unusual patterns. This type of proactive reporting can expedite investigations and enable timely legal action against fraudsters. Fourth, the public could also be informed to check the consistency of the geolocation of job advertisements in hybrid space. This could be facilitated by leveraging already available tools or through the development of a dedicated app for public use, empowering job seekers to identify geographical inconsistencies effectively. Finally, collaboration between law enforcement agencies, cybersecurity experts, and job platforms is also crucial. Developing shared databases of known fraudsters and suspicious activities can significantly improve intelligence sharing and strengthen efforts to track and prosecute individuals involved in job-related scams. By integrating geographical inconsistency detection with enhanced transparency, proactive anomaly reporting, and collaborative efforts, these entities can create a safer digital environment for job seekers.

Limitations

Despite the research findings, contributions, and implications, this study is not without limitations. First, due to the limited availability of public datasets, this study relied on EMSCAD, which contains a relatively small proportion of fake ads. Although techniques such as SMOTE were employed to mitigate this imbalance during the modelling process, future research would benefit from collecting more balanced datasets to validate the findings of this study. Second, this study did not examine the specific factors that influence the creation and distribution of fake job postings. Understanding these factors is crucial for developing more effective prevention and detection strategies. Future research could benefit from integrating additional datasets, such as those containing information on economic conditions, job market trends, and regional employment patterns, to gain a more comprehensive understanding of the dynamics behind employment scams. Third, integrating qualitative data, such as interviews with victims of job scams or insights from cybersecurity experts, could provide valuable context that quantitative data alone might not reveal. This multidimensional approach could lead to the development of more nuanced strategies for combating employment scams.

Conclusions

In summary, this current study tackles the increasing issue of identifying employment scams by examining the use of NLP and ML to model fake job identification with a public dataset. The research provides three main contributions: (1) improving prediction model accuracy through the integration of physical and virtual geographical information, (2) emphasizing the importance of interdisciplinary collaboration across GIS, data science, AI, and cybersecurity/criminology, and (3) creating a framework that can be applied to other types of cybercrime such as online dating scams, fraud, and e-commerce fraud. The findings show that well-crafted job postings are more likely to be legitimate, while those offering unrealistic incentives and lacking details tend to be fraudulent. Geographical differences reveal a higher concentration of fraudulent postings in Texas and California, indicating these areas as targets. By comparing virtual and physical location data, the study underscores the importance of consistency and geographic information in fraud detection, offering valuable insights to enhance the reliability and security of job listing platforms.

Note

1. The main distinction between ML and DL models lies in their complexity and structure. Some advanced ML models can approach the complexity of DL, but DL remains better suited for tasks involving hierarchical data processing such as image recognition and natural language processing (NLP). In the context of AI, scripted models follow preset rules, while generative models, often utilizing DL, can create new content based on learned patterns. This is particularly evident in NLP tasks, where DL-based generative AI produces original outputs rather than following predefined scripts.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This material is partially based upon work supported by the National Science Foundation under Grant No. 2331984 and the start-up Grant from Department of Justice Studies at Prairie View A&M University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and the funders have no role in the study design, data collection, analysis, or preparation of this article.

Notes on contributors

Wenjing Gong and Shoujia Li are PhD students in the Department of Landscape Architecture and Urban Planning & Center for Geospatial Sciences, Applications and Technology at Texas A&M University. Their research interests are on urban analytics and computational social science.

Claire Seungeun Lee is an Associate Professor in the School of Criminology and Justice Studies and a Member of the Center for Internet Security and Forensics Education and Research (iSAFER) at University of Massachusetts-Lowell. She is a Core Personnel of the Center for Asian American Studies and a Fellow of the Center for Public Opinion.

Daylon Adkison is a master's student in the Department of Computer Science at Prairie View A&M University.

Na Li is an Associate Professor in the Department of Computer Science at Prairie View A&M University. Her research is focused on cybersecurity, including privacy and security in online social networks, Network Security, and Security Education.

Xinyue Ye is the Harold Adams Endowed Professor in Urban Informatics and Geospatial AI at Texas A&M University, where he directs the Center for Geospatial Sciences, Applications, and Technology established by the Texas A&M Board of Regents.

Ling Wu is an Associate Professor in the Department of Justice Studies at Prairie View A&M University. Her research is focused on Criminal Justice, Quantitative Methods, Juvenile Delinquency, and Juvenile Justice.

References

Adnan, M., M. I. U. Alaa Abdul Salam Alarood, I. Ur Rehman, and I. Ur Rehman. 2022. "Utilizing Grid Search Cross-Validation with Adaptive Boosting for Augmenting Performance of Machine Learning Models." *PeerJ Computer Science* 8 (February): e803. https://doi.org/10.7717/peerj-cs.803.

Altinok, D. 2021. *Mastering spaCy: An End-To-End Practical Guide to Implementing NLP Applications Using the Python Ecosystem*. Birmingham, UK: Packt Publishing Ltd.

Amaar, A., W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, and S. Ludi. 2022. "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches." *Neural Processing Letters* 54 (3): 2219–2247. https://doi.org/10.1007/s11063-021-10727-z.

Bhatia, T., and J. Meena. 2022. "Detection of Fake Online Recruitment Using Machine Learning Techniques." *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, 300–304. https://doi.org/10.1109/ICAC3N56670.2022.10074276.

Blagus, R., and L. Lusa. 2013. "SMOTE for High-Dimensional Class-Imbalanced Data." *BMC Bioinformatics* 14 (1): 106. https://doi.org/10.1186/1471-2105-14-106.

Buchanan, T., and M. T. Whitty. 2014. "The Online Dating Romance Scam: Causes and Consequences of Victimhood." *Psychology Crime & Law* 20 (3): 261–283. Routledge. https://doi.org/10.1080/1068316X.2013.772180.

Burch, M., S. Lohmann, D. Pompe, and D. Weiskopf. 2013. "Prefix Tag Clouds." *2013 17th International Conference on Information Visualisation*, 45–50. https://doi.org/10.1109/IV.2013.5.

Button, M., C. Lewis, and J. Tapley. 2009. "Fraud Typologies and Victims of Fraud." *National Fraud Authority*. https://researchportal.port.ac.uk/en/publications/fraud-typologies-and-the-victims-of-fraud-literature-review.

Chadha, A., and B. Kaushik. 2022. "A Hybrid Deep Learning Model Using Grid Search and Cross-Validation for Effective Classification and Prediction of Suicidal Ideation from Social Network Data." *New Generation Computing* 40 (4): 889–914. https://doi.org/10.1007/s00354-022-00191-1.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *The Journal of Artificial Intelligence Research* 16 (June): 321–357. https://doi.org/10.1613/jair.953.

Chayko, M. 2020. *Superconnected: The Internet, Digital Media, and Techno-Social Life*. Thousand Oaks, California, USA: SAGE Publications.

Chiraratanasopha, B., and T. Chay-Intr. 2022. "Detecting Fraud Job Recruitment Using Features Reflecting from Real-World Knowledge of Fraud." *Current Applied Science and Technology* 22 (6). https://doi.org/10.55003/cast.2022.06.22.008.

Clarke, R. V., and M. Felson, eds. 2017. *Routine Activity and Rational Choice: Volume 5*. New York: Routledge. https://doi.org/10.4324/9781315128788.

Cole, T. 2022. "Exploring Fraudsters Strategies to Defraud Users on Online Employment Databases." *International Journal of Cyber Criminology* 16 (2): 61–83.

Dake, D. 2023. "Online Recruitment Fraud Detection: A Machine Learning-Based Model for Ghanaian Job Websites." *International Journal of Computer Applications* 184 (51): 20–28. https://doi.org/10.5120/ijca2023922639.

Dupont, B., and C. Whelan. 2021. "Enhancing Relationships Between Criminology and Cybersecurity." *Journal of Criminology* 54 (1): 76–92. https://doi.org/10.1177/00048658211003925.

Fam, S.-F., J. Hui Soo, and S. Imam Wahjono. 2017. "Online Job Search Among Millennial Students in Malaysia." *JDM (Jurnal Dinamika Manajemen)* 8 (1): 1–10. https://doi.org/10.15294/jdm.v8i1.10406.

FBI Warns Cyber Criminals Are Using Fake Job Listings to Target Applicants' Personally Identifiable Information. 2021. *Press Release*, Federal Bureau of Investigation. https://www.fbi.gov/contact-us/field-offices/elpaso/news/press-releases/fbi-warns-cyber-criminals-are-using-fake-job-listings-to-target-applicants-personally-identifiable-information.

Gong, W., X. Huang, M. White, and N. Langenheim. 2023. "Walkability Perceptions and Gender Differences in Urban Fringe New Towns: A Case Study of Shanghai." *The Land* 12 (7): 1339. https://doi.org/10.3390/land12071339.

Gong, W., J. Rui, and T. Li. 2024. "Deciphering Urban Bike-Sharing Patterns: An In-Depth Analysis of Natural Environment and Visual Quality in New York's Citi Bike System." *Journal of Transport Geography* 115 (February): 103799. https://doi.org/10.1016/j.jtrangeo.2024.103799.

Grant-Smith, D., A. Feldman, and C. Cross. 2022. ""Key Trends in Employment Scams in Australia: What are the Gaps in Knowledge About Recruitment Fraud?" Contribution to Newspaper, Magazine or Website." QUT Centre for Justice Briefing Papers. QUT Centre for Justice. https://eprints.qut.edu.au/228500/.

Habiba, S. U., M. Khairul Islam, and F. Tasnim. 2021. "A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques." *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 543–546. https://doi.org/10.1109/ICREST51555.2021.9331230.

Hall, T., and R. Yarwood. 2024. "New Geographies of Crime? Cybercrime, Southern Criminology and Diversifying Research Agendas." *Progress in Human Geography* 48 (4): 437–457. SAGE Publications Ltd https://doi.org/10.1177/03091325241246015.

Hasan, M., M. Salehin, and M. Islam. 2018. "Determinants of Graduate Students' Usage of Online Social Media as a Job Searching Tool: The Changing Landscape of Labor Market in Bangladesh." *Review of Public Administration and Management* 6 (01). https://doi.org/10.4172/2315-7844.1000245.

Hayward, K. J. 2012. "Five Spaces of Cultural Criminology." *The British Journal of Criminology* 52 (3): 441–462. https://doi.org/10.1093/bjc/azs008.

Hijji, M., and G. Alam. 2021. "A Multivocal Literature Review on Growing Social Engineering Based Cyber-Attacks/threats During the COVID-19 Pandemic: Challenges and Prospective Solutions." *Institute of Electrical and Electronics Engineers Access* 9:7152–7169. https://doi.org/10.1109/ACCESS.2020.3048839.

Hoo, Z. H., J. Candlish, and D. Teare. 2017. "What is an ROC Curve?" *Emergency Medicine Journal* 34 (6): 357–359. BMJ Publishing Group Ltd and the British Association for Accident & Emergency Medicine https://doi.org/10.1136/emermed-2017-206735.

IC3 Issues Alert on Employment Scams. 2020. https://www.cisa.gov/news-events/alerts/2020/01/22/ic3-issues-alert-employment-scams.

Internet Crime Complaint Center. 2023. https://www.ic3.gov/.

Jadhav, S. S., and S. D. Thepade. 2019. "Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier." *Applied Artificial Intelligence* 33 (12): 1058–1068. Taylor & Francis https://doi.org/10.1080/08839514.2019.1661579.

Jiang, R., R. E. Banchs, and H. Li. 2016. "Evaluating and Combining Name Entity Recognition Systems." In *Proceedings of the Sixth Named Entity Workshop*, edited by X. Duan, R. E. Banchs, M. Zhang, H. Li, and A. Kumaran, 21–27. Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2703.

Jogalekar, J., and A. Nanasaheb Jadhav. 2022. "The CON-MEN of EMPLOYMENT in India: 'job SCAMS, a CURSE." no. 4.

Kircher, P. A. T. 2020. "Search Design and Online Job Search – New Avenues for Applied and Experimental Research." *Labour Economics* 64 (June): 101820. https://doi.org/10.1016/j.labeco.2020.101820.

Lee, C. S. 2021. "Online Fraud Victimization in China: A Case Study of Baidu Tieba." *Victims & Offenders* 16 (3): 343–362. Routledge https://doi.org/10.1080/15564886.2020.1838372.

Lee, C. S. 2022. "How Online Fraud Victims are Targeted in China: A Crime Script Analysis of Baidu Tieba C2C Fraud." *Crime & Delinquency* 68 (13–14): 2529–2553. SAGE Publications Inc https://doi.org/10.1177/00111287211029862.

Li, J., Y. Li, H. Han, and X. Lu. 2022. "Exploratory Methods for Imbalanced Data Classification in Online Recruitment Fraud Detection: A Comparative Analysis." *2021 4th International Conference on Computing and Big Data*, 75–81. ICCBD 2021, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3507524.3507537.

Li, Q., W. Wei, N. Xiong, D. Feng, X. Ye, and Y. Jiang. 2017. "Social Media Research, Human Behavior, and Sustainable Society." *Sustainability* 9 (3): 384.

Mahbub, S., E. Pardede, and A. S. M. Kayes. 2022. "Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries." *Institute of Electrical and Electronics Engineers Access* 10:82776–82787. https://doi.org/10.1109/ACCESS.2022.3197225.

Malaichamy, G. 2023. "Online Job Posting Authenticity Prediction Using Machine and Deep Learning Techniques." Masters, Dublin, National College of Ireland. https://norma.ncirl.ie/6611/ .

Meneses Silva, C. V., R. Silva Fontes, and M. Colaço Júnior. 2021. "Intelligent Fake News Detection: A Systematic Mapping." *Journal of Applied Security Research* 16 (2): 168–189. Routledge https://doi.org/10.1080/19361610.2020.1761224.

Mishra, R. K., J. A. A. J. Abdul Rahmaan Ansari, and V. Mishra. 2024. "Analysis of Criminal Landscape by Utilizing Statistical Analysis and Deep Learning Techniques." *Journal of Applied Security Research* 1–26. Routledge https://doi.org/10.1080/19361610.2024.2314392.

Mutemi, A., and F. Bacao. 2024. "E-Commerce Fraud Detection Based on Machine Learning Techniques: Systematic Literature Review." *Big Data Mining & Analytics* 7 (2): 419–444. https://doi.org/10.26599/BDMA.2023.9020023.

Naseer, S., M. Mudasar Ghafoor, S. Bin Khalid Alvi, A. Kiran, G. Murtazae Shafique Ur Rahmand, and G. Murtaza. 2021. "Named Entity Recognition (NER) in NLP Techniques, Tools Accuracy and Performance." *Pakistan Journal of Multidisciplinary Research* 2 (2): 293–308.

Nessa, I., B. Zabin, K. Omar Faruk, A. Rahman, K. Nahar, S. Iqbal, M. Sabbir Hossain, M. Humaion Kabir Mehedi, and A. Alim Rasel. 2022. "Recruitment Scam Detection Using Gated Recurrent Unit." *2022 IEEE 10th Region 10 humanitarian Technology Conference* (R10-HTC), 445–449. https://doi.org/10.1109/R10-HTC54060.2022.9929928.

North American Industry Classification System (NAICS) U.S. Census Bureau. 2024. Accessed July 8. https://www.census.gov/naics/?58967?yearbck=2022.

Obuene, H. U., O. Tade, B. Rasak, O. Arisukwu, and E. E. Okafor. 2024. "Job Advertisements and Lived Experiences of Victims of Job Scams in Ibadan, Nigeria." *International Journal of Sociology & Social Policy* 44 (11/12): 1123–1136. https://doi.org/10.1108/IJSSP-03-2024-0142.

Pratley, J. D., and M. Masbaul Alam Polash. 2023. "Fraudulent Jobs Prediction Using Natural Language Processing and Deep Learning Sequential Models." In *Web Information Systems Engineering – WISE 2023*, edited by F. Zhang, H. Wang, M. Barhamgi, L. Chen, and R. Zhou, 509–519. Singapore: Springer Nature. https://doi.org/10.1007/978-981-99-7254-8_39.

Ravenelle, A. J., E. Janko, and K. Cai Kowalski. 2022. "Good Jobs, Scam Jobs: Detecting, Normalizing, and Internalizing Online Job Scams During the COVID-19 Pandemic." In *New Media & Society*, London, England: SAGE PublicationsSage UK. https://doi.org/10.1177/14614448221099223.

Reed, T. V. 2018. *Digitized Lives: Culture, Power and Social Change in the Internet Era*. 2nd ed. New York: Routledge. https://doi.org/10.4324/9781315143415.

Ribeiro Bezerra, J. F. 2021. "Content-Based Fake News Classification Through Modified Voting Ensemble." *Journal of Information and Telecommunication* 5 (4): 499–513. Taylor & Francis https://doi.org/10.1080/24751839.2021.1963912.

Richards, J. 2012. "What Has the Internet Ever Done for Employees? A Review, Map and Research Agenda." *Employee Relations* 34 (1): 22–43. Emerald Group Publishing Limited https://doi.org/10.1108/01425451211183246.

Shelar, H., G. Kaur, N. Heda, and P. Agrawal. 2020. "Named Entity Recognition Approaches and Their Comparison for Custom NER Model." *Science & Technology Libraries* 39 (3): 324–337. Routledge https://doi.org/10.1080/0194262X.2020.1759479.

Sinclair, J., and M. Cardew-Hall. 2008. "The Folksonomy Tag Cloud: When is it Useful?" *Journal of Information Science* 34 (1): 15–29. SAGE Publications Ltd https://doi.org/10.1177/0165551506078083.

Sofy, M. A., M. H. Khafagy, and R. M. Badry. 2023. "An Intelligent Arabic Model for Recruitment Fraud Detection Using Machine Learning." *Journal of Advances in Information Technology*. https://doi.org/10.12720/jait.14.1.102-111.

Swetha, K., M. Tharun Reddy, K. Sravani, and B. Subramanyam. 2023. "Fake Job Detection Using Machine Learning Approach." *Journal of Engineering Sciences* 14 (2):67–74.

Tabassum, H., G. Ghosh, A. Atika, and A. Chakrabarty. 2021. "Detecting Online Recruitment Fraud Using Machine Learning." *2021 9th International Conference on Information and Communication Technology (ICoICT)*, 472–477. https://doi.org/10.1109/ICoICT52021.2021.9527477.

Vasiliev, Y. 2020. *Natural Language Processing with Python and spaCy: A Practical Introduction*. San Francisco, California, USA: No Starch Press.

Vasist, P. N., and D. Chatterjee. 2023. "Combating Fake News and Digital Deception at the Workplace: An Integrative Review and Open Systems Theory-Led Framework for Future Research." In *IIM Kozhikode Society & Management Review*, 22779752231163360. SAGE Publications India. https://doi.org/10.1177/22779752231163360.

Vidros, S., C. Kolias, and G. Kambourakis. 2016. "Online Recruitment Services: Another Playground for Fraudsters." *Computer Fraud & Security* 2016 (3): 8–13. https://doi.org/10.1016/S1361-3723(16)30025-2.

Vidros, S., C. Kolias, G. Kambourakis, and L. Akoglu. 2017. "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset." *Future Internet* 9 (1): 6. Multidisciplinary Digital Publishing Institute https://doi.org/10.3390/fi9010006.

Wahid, A. 2023. "Prevalence of Employment Fraud in Indonesia: Highlighting Criminal Strategies and Government Efforts." *International Journal of Criminal Justice Sciences* 18 (1): 52–63.

Wall, D. S. 2007. "Cybercrime: The Transformation of Crime in the Information Age." *SSRN Scholarly Paper*, Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=1066922.

Williams, M. L. 2016. "Guardians Upon High: An Application of Routine Activities Theory to Online Identity Theft in Europe at the Country and Individual Level." *The British Journal of Criminology* 56 (1): 21–48. https://doi.org/10.1093/bjc/azv011.

Wright, D. C. 2023. "Geographical Aspects of Cybercrime: A Literature Review." *SSRN Electronic Journal*. https://coingeek.com/geographical-aspects-of-cybercrime-a-literature-review/.

Wu, L., S. Jamonnak, X. Ye, S.-L. Shaw, N. Saxena, and K.-S. Choi. 2024. "Theoretical and Experimental Framework for Estimating Cyber Victimization Risk in a Hybrid Physical-Virtual World." *Journal of Applied Security Research*: 1–25. https://doi.org/10.1080/19361610.2024.2368969.

Yanti, R. M., I. Santoso, and L. Hulliyyatus Suadaa. 2021. "Application of Named Entity Recognition via Twitter on SpaCy in Indonesian (Case Study: Power Failure in the Special Region of Yogyakarta)." *Indonesian Journal of Information Systems* 4 (1): 76–86. https://doi.org/10.24002/ijis.v4i1.4677.

Yar, M. 2005. "The Novelty of 'Cybercrime': An Assessment in Light of Routine Activity Theory." *European Journal of Criminology* 2 (4): 407–427. SAGE Publications https://doi.org/10.1177/147737080556056.

Ye, X., W. Wang, X. Zhang, Z. Li, D. Yu, J. Du, and Z. Chen. 2021. "Reconstructing Spatial Information Diffusion Networks with Heterogeneous Agents and Text Contents." *Transactions in GIS* 25 (4): 1654–1673.

Appendix

Figure A1. Word cloud in the title of (a) fraudulent and (b) legitimate job postings.

Table A1. Mapping relationship between NAICS definitions and EMSCAD industry features.

NAICS_sector	NAICS_definition	Industries in EMSCAD
72	Accommodation and Food Services	Hospitality, Leisure, Travel & Tourism, Food & Beverages, Restaurants
11	Agriculture, Forestry, Fishing and Hunting	Farming, Fishery, Ranching
71	Arts, Entertainment, and Recreation	Online Media, Computer Games, Entertainment, Broadcast Media, Media Production, Gambling & Casinos, Sports, Music, Motion Pictures and Film, Performing Arts, Museums and Institutions
23	Construction	Construction, Building Materials
61	Educational Services	Education Management, E-Learning, Professional Training & Coaching, Higher Education, Primary/Secondary Education
52	Finance and Insurance	Financial Services, Accounting, Insurance, Banking, Venture Capital & Private Equity, Investment Management, Capital Markets, Investment Banking
62	Health Care and Social Assistance	Hospital & Health Care, Health, Wellness and Fitness, Medical Practice, Mental Health Care, Individual & Family Services
51	Information	Information Technology and Services, Computer Software, Internet, Telecommunications, Computer & Network Security, Computer Networking, Publishing, Information Services, Animation, Wireless, Libraries
31–33	Manufacturing	Textiles, Wine and Spirits, Printing, Chemicals, Packaging and Containers, Plastics, Electrical/Electronic Manufacturing, Computer Hardware, Aviation & Aerospace, Medical Devices, Semiconductors, Machinery, Defense & Space, Industrial Automation, Shipbuilding, Automotive, Apparel & Fashion, Cosmetics, Food Production, Pharmaceuticals, Furniture
21	Mining, Quarrying, and Oil and Gas Extraction	Oil & Energy, Mining & Metals
81	Other Services (except Public Administration)	Consumer Services, Nonprofit Organization Management, Fund-Raising, Religious Institutions, Philanthropy, Civic & Social Organization
54	Professional, Scientific, and Technical Services	Marketing and Advertising, Management Consulting, Design, Legal Services, Public Relations and Communications, Market Research, Biotechnology, Mechanical or Industrial Engineering, Graphic Design, Research, Writing and Editing, Law Practice, Government Relations, Translation and Localization, Architecture & Planning, Civil Engineering, Renewables & Environment, Program Development, International Trade and Development, Veterinary, Photography, Public Policy, Nanotechnology, Alternative Dispute Resolution
92	Public Administration	Government Administration, Law Enforcement, Executive Office, Public Safety, Military
53	Real Estate and Rental and Leasing	Real Estate, Commercial Real Estate
44–45	Retail Trade	Luxury Goods & Jewelry, Business Supplies and Equipment, Sporting Goods, Retail, Consumer Goods, Consumer Electronics
48–49	Transportation and Warehousing	Maritime, Package/Freight Delivery, Logistics and Supply Chain, Airlines/Aviation, Transportation/Trucking/Railroad, Warehousing
22	Utilities	Utilities
56	Waste Management and Remediation Services	Staffing and Recruiting, Human Resources, Facilities Services, Events Services, Environmental Services, Security and Investigations, Outsourcing/Offshoring
42	Wholesale Trade	Wholesale, Import and Export

Table A2. Count of top 10 required education in postings and fraudulent postings.

required_education	posting count	fraudulent posting count
Bachelor's Degree	5145	100
High School or equivalent	2080	170
Unspecified	1397	61
Master's Degree	416	31
Associate Degree	274	6
Certification	170	19
Some College Coursework Completed	102	3
Professional	74	4
Vocational	49	0
Some High School Coursework	27	20
Doctorate	26	1
Vocational – HS Diploma	9	0
Vocational – Degree	6	0

Table A3. Count of top 10 required experience in postings and fraudulent postings.

required_experience	posting count	fraudulent posting count
Mid-Senior level	3809	113
Entry level	2697	179
Associate	2297	42
Not Applicable	1116	60
Director	389	17
Internship	381	10
Executive	141	10

Table A4. Count of top 10 employment types in postings and fraudulent postings.

employment_type	posting count	fraudulent posting count
Full-time	11620	490
Contract	1524	44
Part-time	797	74
Temporary	241	2
Other	227	15

Table A5. Coefficients and odds ratios from logistic regression Model 4 for predicting fraudulent job postings.

Variable	Coefficient	Odds Ratio
telecommuting	−0.464	0.629
has_company_logo	−2.696	0.067
has_questions	−0.941	0.390
employment_type	−0.304	0.738
required_experience	−0.069	0.933
required_education	0.035	1.036
industry	−0.002	0.998
function	0.002	1.002
department	0.002	1.002
salary_range	0.001	1.001
country	−0.011	0.989
state	−0.007	0.993
city	0.000	1.000
location_consistency	−0.750	0.472