BAREKENG: Journal of Mathematics and Its Applications
March 2026 Volume 20 Issue 1 Page 0865–0880
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss1pp0865-0880

# VISUALIZING AND CLUSTERING FAKE JOB POSTINGS: DATA-DRIVEN INSIGHTS FOR FRAUD DETECTION

**Chee Keong Ch'ng** [1], **Xiang Yi Wong** [2*]

[1,2] *School of Quantitative Sciences, Universiti Utara Malaysia*
*Sintok, Kedah, 06010, Malaysia*

*Corresponding author's e-mail: * **xiangyiwong0705@gmail.com***

***Article Info***

***Article History:***

*Received: 16th June 2025*
*Revised: 15th July 2025*
*Accepted: 29th July 2025*
*Available online: 24th November 2025*

***Keywords:***

*Data visualization;*
*Fake job postings;*
*Fraud detection;*
*Job scams;*
*Text clustering.*

## *ABSTRACT*

*Online job platforms have made it easier to find jobs, but they have also made it easier for scammers to post fake job postings, posing risks to job seekers. These fraudulent activities can lead to severe consequences, such as identity theft, financial loss, and emotional distress for victims. To improve recruitment platform security and safeguard users, it is essential to spot trends in these fake job postings. This study focuses on visualizing patterns within fake job postings through data-driven insights, employing various data visualization techniques to reveal key attributes associated with fraudulent activity. A dataset contains both legitimate and fraudulent job postings. Exploratory data analysis (EDA) is conducted to examine variables including salary category, job function, industry, location, and other related features by using categorical distribution, geographical distribution, and word cloud. This study provides insights for recruitment platform controllers, raises user awareness, and facilitates the early detection of fraudulent job posts by displaying clear and actionable visual patterns. The results highlight how visualization and clustering are used to gain insight into characteristics of fraudulent job postings, like the fraudulent job postings predominantly target customer-facing roles in industries like Oil & Energy and Customer Service, which are concentrated in the United States (especially Texas and California), and rely on vague language and unrealistic promises. These findings contribute to more targeted fraud detection strategies and create safer online job search environments.*

***How to cite this article:***

C. K. Ch'ng and X. Y. Wong, "VISUALIZING AND CLUSTERING FAKE JOB POSTINGS: DATA-DRIVEN INSIGHTS FOR FRAUD DETECTION", *BAREKENG: J. Math. & App.*, vol. 20, iss. 1, pp. 0865-0880, Mar, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

**Research Article · Open Access**

## 1. INTRODUCTION

The pre-Internet era relied heavily on newspapers, where job seekers had to sift through classified ads to find employment opportunities. With the rise of the internet, the job search process shifted online, enabling rapid information exchange but also creating opportunities for scammers. As a result, fake job postings and fraudulent recruitment schemes have grown significantly, targeting vulnerable job seekers [1]. The COVID-19 pandemic further accelerated this trend, with a spike in bogus employment offers, overseas work scams, and multi-level marketing schemes since April 2020 [2]. In 2023, job scams ranked among the top 10 scam categories, with financial losses in Australia alone increasing by 150% compared to 2022 [3].

Work-from-home scams, deceptive emails, and social media job hoaxes are among the most common types of employment fraud, often aiming to steal money or personal data [4],[5]. Despite ongoing awareness campaigns, many individuals continue to fall victim, sometimes with severe consequences, including involvement in human trafficking [6],[7]. Financial losses are staggering across the globe, including $2.74 billion in Australia, over $27 million in Canada, and RM33.9 million in Malaysia [8]-[10]. At-risk groups often include students, flexible job seekers, and the long-term unemployed [11],[12].

In recent years, researchers have employed both machine learning and deep learning techniques to detect fraudulent job postings. Machine learning remains the foundation of most detection models, with classifiers such as Naïve Bayes, K-Nearest Neighbors, Decision Trees, and Logistic Regression frequently used. Ensemble methods like Random Forest, AdaBoost, Gradient Boosting, and boosted decision trees often outperform single classifiers, with reported accuracies above 98% [13]-[15]. Optimization approaches, including GridSearchCV, further enhance classifier performance, while genetic algorithm–based methods have leveraged SVM and Random Forest to reach up to 97% accuracy [16],[17]. These findings highlight the strong predictive ability of ensemble and optimized machine learning models in this domain.

Along with machine learning, deep learning methods have also shown competitive results. Studies have used neural architectures such as sequential neural networks with GloVe embeddings, Bi-Directional LSTM, GRU variants, and attention-based GRU models, achieving accuracies exceeding 97% [18]-[20]. Hybrid frameworks have also emerged, such as knowledge graph–driven deep neural networks (FRJD), which outperformed classical classifiers, including Random Forest, Logistic Regression, and SVM [21]. Despite these advancements, current research has primarily focused on classification performance. There remains limited analysis of the linguistic and structural characteristics of fraudulent job postings and little exploration of similarity-based clustering.

However, unlike these studies, this research adopts a different approach by focusing exclusively on text mining and data visualization techniques, rather than predictive classification models. Text mining, through content analysis and clustering, enables the extraction of meaningful patterns from unstructured job descriptions. Exploratory Data Analysis (EDA) supports this by visualizing categorical, geographical, and linguistic distributions, offering deeper insights into scam behaviors. Given the unstructured nature of job posting datasets, clustering techniques, such as hierarchical clustering, are essential for grouping similar postings and revealing how fraudulent listings mimic legitimate ones [22]. Rather than building a predictive model, this study emphasizes interpretable, descriptive analysis to understand the underlying strategies used in scam job postings.

Therefore, this study aims to leverage text mining, clustering, and visualization to identify key fraud indicators through content analysis, explore how fake postings mimic legitimate ones using text clustering, and visualize patterns using EDA to enhance understanding and detection. This approach offers actionable insights for researchers, job platforms, and policymakers, contributing to the development of more effective fraud prevention strategies beyond conventional classification techniques.

## 2. RESEARCH METHODS

This section discusses the process of data cleaning, data visualization, text preprocessing, and text clustering.

### 2.1 Research Design

According to Fig. 1, Phase 1 begins by defining the research questions, problems, and objectives. Once the objectives are clearly established, it is essential to review related works by other researchers to gain insights into their methodologies and findings. Following this, a dataset named Employment Scam Aegean Dataset (EMSCAD) is used, where all the observations in the dataset are collected from a job search website, Workable. Most prior EMSCAD-based studies rely on classification, a supervised learning approach where models are trained to distinguish between legitimate and fraudulent postings based on labeled data. In contrast, our study emphasizes clustering, an unsupervised method that groups postings by similarity without requiring labels. This approach not only complements classification but also reveals hidden structures, linguistic patterns, and subtypes of fraudulent postings, thereby providing a deeper and more interpretable understanding of scam behaviors. In this research, Orange will be used as it is an open-source toolkit for data visualization, machine learning, and data mining, featuring a visual programming interface for exploratory analysis and interactive visualization. The selected dataset is then loaded into Orange, where missing values in categorical variables are imputed. Phase 2 involves loading the imputed dataset into Tableau for exploratory data analysis (EDA). This step helps uncover meaningful insights into the characteristics of fake job postings.

In Phase 3, the dataset is reloaded into Orange for text preprocessing. This step ensures the data is clean and free of noise, such as stop words, delimiters, punctuation marks, and irrelevant patterns (e.g., regular expressions). Text preprocessing involves key tasks such as transformation, tokenization, normalization, and filtering. The result is a refined word cloud that excludes non-informative elements, such as stop words. Finally, Phase 4 focuses on text clustering. Since machine learning algorithms cannot directly interpret non-numeric data, the text must first be converted into numerical representations. Widgets such as “Bag of Words” and “Hierarchical Clustering” are used to identify potential clusters within the data. Once the clusters are identified, their meanings are interpreted and defined. Lastly, a SWOT analysis is carried out to understand the strengths, weaknesses, opportunities, and threats of this research.

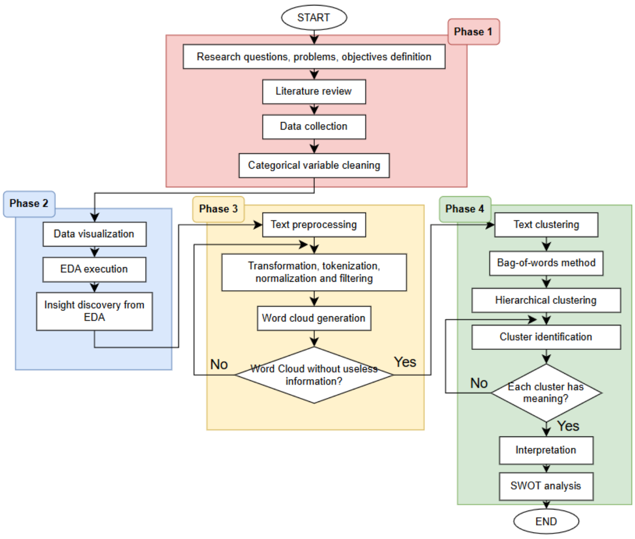


Figure 1. Research Flow Chart

## 2.2 Data Collection

This study utilizes secondary data from the Employment Scam Aegean Dataset (EMSCAD), which is publicly available on Kaggle. The original dataset consists of real job postings sourced from Workable and includes both qualitative and quantitative data. To derive deeper insights, it is essential to apply text mining techniques to convert the qualitative data into numerical values.

The dataset comprises 17,880 related to job advertisements. Of these, 17,014 are legitimate job postings, while 866 are identified as fraudulent. The dataset contains 18 features that can be analyzed to distinguish between genuine and fake job postings. The dataset contains eight categorical features: fraudulent, telecommuting, has company logo, has questions, employment type, required experience, required education, and function. It also includes one numerical feature, which is the job ID. Additionally, there are nine meta features: title, department, company profile, description, requirements, benefits, industry, location, and salary range. This comprehensive dataset provides a robust foundation for identifying patterns and characteristics that differentiate legitimate job postings from fraudulent ones.

### 2.3 Data Cleaning

The dataset, comprising 17,880 observations, contains a significant proportion of missing values, where 17.5% in its features and 51.9% in its meta variables. Fig. 2 shows the imputation of the categorical variables in different methods. Among the 8 categorical features, 4 contain missing values. The "employment_type" feature, which has 19% missing values and a skewed distribution, was imputed using the most frequent value. However, features such as "required_experience", "required_education", and "function" have 30-50% missing values. For these features, a model-based imputer, specifically a simple tree imputation method, was employed to effectively replace the missing values. For the meta variables, observations with null values were removed entirely, reducing the dataset to 900 entries.

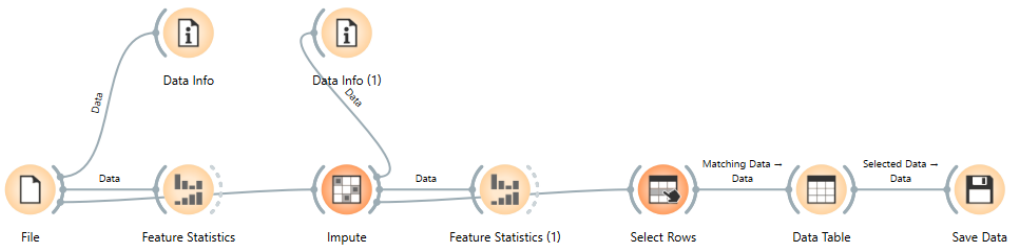


**Figure 2. Imputation Process During Data Preprocessing**

The "location" feature in the dataset, which combines country, state, and city in a single cell, was split into three separate columns: "location_country", "location_state", and "location_city". Additionally, the "salary_range" feature, initially recorded as an annual value in the meta variables, was split into "salary_min" and "salary_max". Using SPSS, a new variable, "salary_medium", was computed as the average of the min and max salaries. Salaries were categorized into three groups based on [23]. High salaries were defined as $106,501 and above, average salaries ranged from $67,001 to $106,500, and low salaries included those up to $67,000. These features and their details are presented in Table 1.

**Table 1. Summary of Data Transformation and Categorization**

| Feature | Transformation/Action | Details |
|---|---|---|
| Location | Split into three columns | Originally, the feature "Location" contained all three countries, state, and city together in a single column; the column is then split into 3 columns named "location_country", "location_state", and "location_city" |
| Salary Range | Split into two columns | The feature "Salary Range" contains the maximum and minimum values of annual salary. The column is then split into 2 columns named "salary_max" and "salary_min" |
| Salary Medium | Computed as the average of salary_min and salary_max | New variable salary_medium calculated in SPSS as $(salary\_min + \ salary\_max\ )/2$ |
| Salary Categories | Categorized into three groups based on [23] | The salary is categorized as high ($106,501 and above), average ($67,001 to $106,500) and low (Up to $67,000) |

The distribution of postings before sampling is shown in Table 2. A minimum sample size of 377 is required for a population of 17,880 at a 95% confidence level. However, 600 samples were selected for this study to ensure a more robust dataset, as more data generally leads to better insights. All 72 fraudulent job postings were retained, while a random sample of 528 legitimate postings was drawn from 816 available legitimate postings using the data sampler widget in Orange. The resulting dataset comprises 600 observations: 72 fraudulent and 528 legitimate.

**Table 2. Distribution of Postings**

| Type of Postings | Number of Postings | |
|---|---|---|
| | Before Sampling | After Sampling |
| Fraudulent Job Postings | 72 | 72 |
| Legitimate Job Postings | 816 | 528 |
| Total | 900 | 600 |

The differences between before and after sampling in the number of job postings for both fraudulent and legitimate ones are shown in Table 2. The process of sampling is shown in Fig. 3.

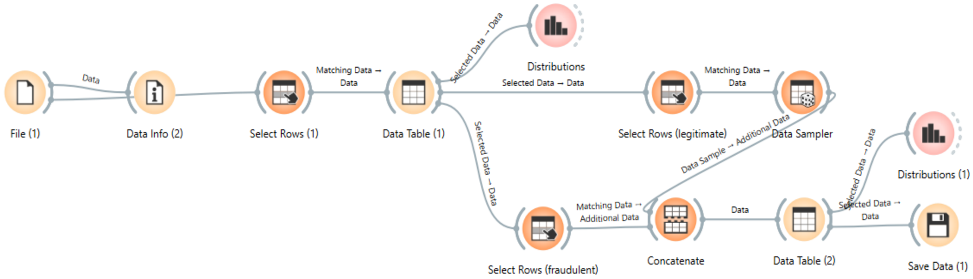


Figure 3. Process of Sampling

## 2.4 Data Visualization and Analysis

Tableau was used to visually explore and analyze the dataset through interactive charts and dashboards. It offers a range of functionalities that enhance data analysis, including support for complex calculations, integration with various data sources, and tools for collaboration and sharing insights. Additionally, Tableau's capabilities for predictive analytics and forecasting were utilized to uncover patterns and trends related to fake job postings.

Four interactive dashboards were created using various visualization types. A pie chart was used to show the proportion of fraudulent versus legitimate job postings. Vertical and horizontal bar charts were employed to display distributions and comparisons across categorical variables, such as employment type and required education. A choropleth map visualized the geographical distribution of job postings by country and state, while a bubble map represented the density of job postings in different cities. Lastly, a heatmap highlighted patterns within the salary categories.

Several calculated fields and parameters were developed to enhance the dashboards' functionality. These included filters, allowing users to interact with and drill down into specific subsets of the data for deeper insights. This dynamic and interactive approach to data visualization enabled a comprehensive understanding of the characteristics and patterns in fake job postings.

## 2.5 Text Preprocessing

After visualizing all the categorical variables, the next step was to focus on the meta variables. However, before visualizing the metadata, text preprocessing was required to prepare the data for analysis, as shown in Fig. 4.

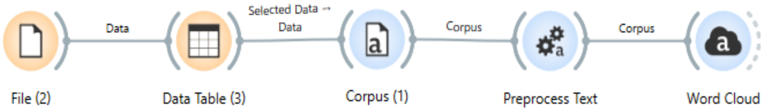


Figure 4. Process of Text Preprocessing

The meta variables in the dataset, such as title, department, company profile, description, requirements, benefits, and location city, were combined into a single column. This unified column was then converted into a corpus for further text analysis.

For text preprocessing, the first step is to transform the text by lowercasing it, removing accents, parsing HTML, and removing URLs. The tokenization process is then needed to split a text document into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters with single white spaces. This tokenized representation is then used for further processing. Followed by normalization of the text using lemmatization methods. This method tries to map verb forms to the infinite tense and nouns to the singular form. Lastly, filtering methods are needed to remove words from the dataset. The filtering methods that are going to be used are stop word filtering and regular expression filtering in order to remove words that bear little or no content information, like articles, conjunctions, prepositions, and delimiters. A limit of 200 most frequent tokens was set to ensure the word cloud effectively highlights the key terms in the combined metadata.

### 2.6 Text Clustering

Once the text data was pre-processed, a Bag of Words (BoW) model was generated to facilitate text clustering, as shown in Fig. 5. Text clustering is an unsupervised learning technique that groups similar text documents together based on their content. By converting text data into numerical format, clustering algorithms can be applied to find patterns and similarities within the text.

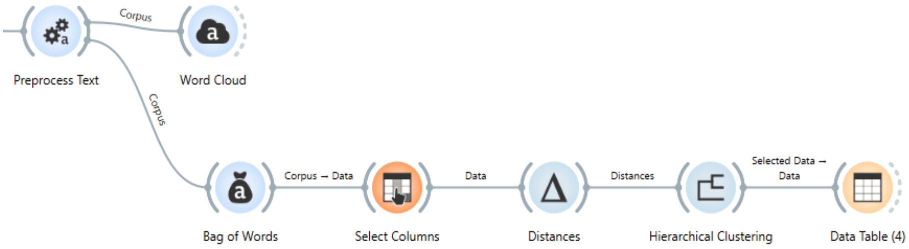


**Figure 5. Process of Text Clustering**

To measure the similarity between different text documents, cosine similarity is commonly used. Cosine similarity calculates the cosine of the angle between two vectors, which reflects how similar the two documents are in terms of their content. The formula for cosine similarity is:

$$Cosine\ similarity\ (A, B) = \frac{\sum_{i=1}^{n}(a_i \cdot b_i)}{\sqrt{\sum_{i=1}^{n} a_i{}^2} \cdot \sqrt{\sum_{i=1}^{n} b_i{}^2}} \quad (1)$$

Where:

$a_i : i^{th}\ components\ of\ vectors\ A.$
$b_i : i^{th}\ components\ of\ vectors\ B.$

Finally, hierarchical clustering can be performed.

## 3. RESULTS AND DISCUSSION

### 3.1 Result of Data Visualization

### 3.1.1 Categorical Distribution

This section begins with an overview of the categorical variables, as shown in Fig. 6. After data cleaning, 600 observations remain, with 88% (528) representing legitimate job postings and 12% (72) representing fraudulent job postings. While fraudulent job postings are prevalent, they are significantly outnumbered by genuine ones. Next, the Tree Map Chart shows that the majority of job postings offer low salaries, comprising 73.67% of the dataset. In contrast, 16.67% offer average salaries, and only 9.67% offer high salaries. This indicates that most job offers provide yearly incomes of less than $67,000. Job postings offering higher salaries are typically associated with higher education levels and experience, which will be discussed further below. On the Workable platform, most job postings (98.33%) feature a company logo, allowing job seekers to easily identify the company. However, not all companies require interviews; 33.50% of postings indicate that no interview is necessary. Additionally, most job postings do not offer telecommuting, reflecting the fact that remote work was uncommon during the 2012-2014 period.

The majority of job postings (91.50%) seek full-time workers, with the remaining postings offering contract, part-time, temporary, or other types of employment. Most job postings require high school or college graduates, suggesting that the target age group is likely between 18 and 24 years old. Vocational and professional roles are less common, indicating that employers tend to prefer candidates with more generalized educational backgrounds, such as Bachelor's or Master's degrees. Although 70% of job postings require high school or college graduates, 30.50% of the postings require mid-senior level experience, and 27.33% require associate-level skills. This suggests that skills and industry experience are often valued more than formal education. Roles such as executive and director positions in accounting account for only 2.67% each. These higher-level roles are typically filled through internal promotions rather than external hiring.

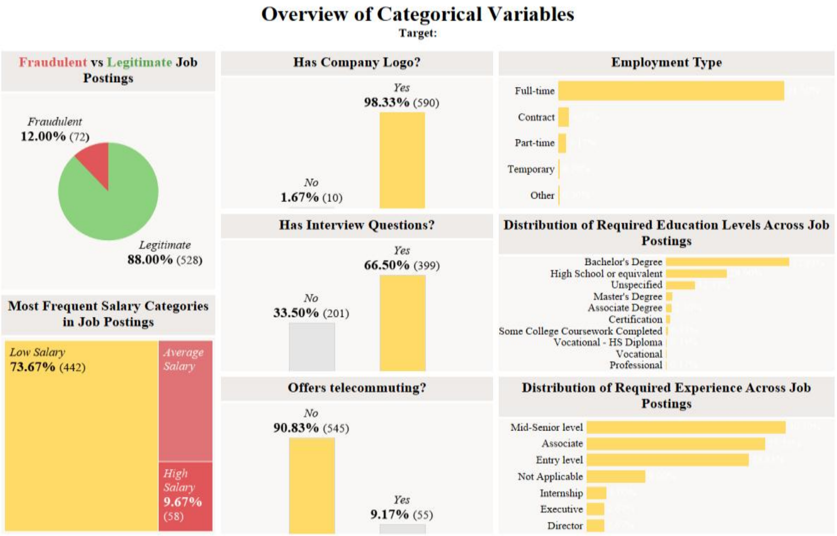


**Figure 6. Overview of Categorical Variables**
*(Source: Tableau 2024.1)*

The analysis presented in Fig.7 focuses specifically on fraudulent postings and highlights a key characteristic: 33.33% of all fraudulent job postings did not require an interview process. This lack of a screening procedure is a major indicator of scams, as legitimate job offers typically involve some form of interview or evaluation. Another notable feature of fraudulent job postings is that they often require only a high school education for entry-level roles. The percentage of job postings without a company logo and those offering telecommuting is slightly higher than that shown in Fig. 6, indicating that these characteristics might be more common in fraudulent job postings.

The detailed examination of fraudulent job postings is presented in Fig.7. While the share of postings without an interview process is nearly the same (33.50% across all postings in Fig. 6 versus 33.33% for fraudulent postings), more distinctive differences are observed. Compared with Fig. 6, fraudulent postings are more likely to require only a high school education, lack a company logo, and advertise telecommuting opportunities. These characteristics suggest that scams often appear more accessible and less formal, serving as potential red flags for job seekers.

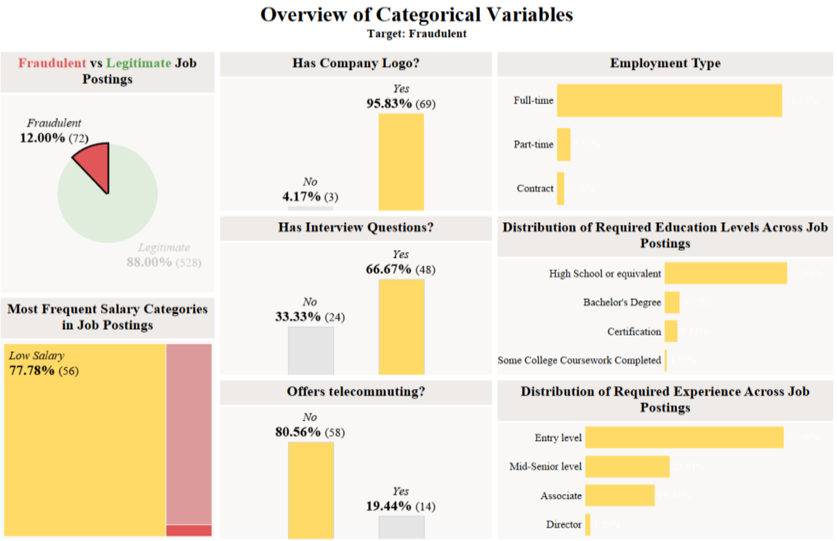


**Figure 7. Overview of Categorical Variables Filter by Fraudulent Job Postings**
*(Source: Tableau 2024.1)*

The common job functions and industries are displayed in Fig. 8. Job functions that were not specifically categorized are grouped under “Other”, while those functions occupying less than 1% of the data are categorized under “All Other Functions”, using a threshold percentage of 1%. The same approach was applied to the industries. The Common Job Functions plot shows that Sales is the most common job function, followed by technical roles such as Engineering and Information Technology, and then Customer Service. In the Common Industries plot, technical industries such as Internet, IT, and Computer Software dominate the job postings. These two plots highlight that customer-facing roles and professional jobs are in high demand. However, the question arises: Does this mean these job functions and industries are more susceptible to fake job postings?

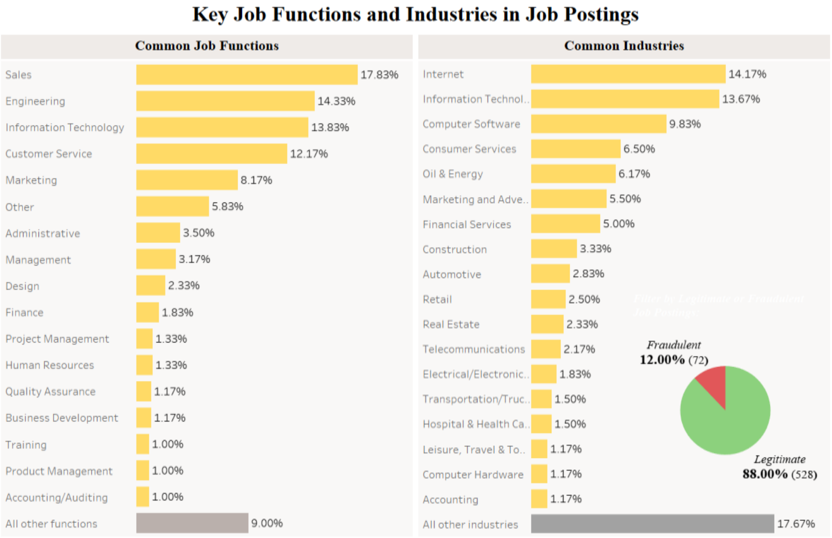


**Figure 8. Key Job Functions and Industries in Job Postings**
*(Source: Tableau 2024.1)*

Gaining insight into the functions and industries of fake job postings is crucial to identifying which are most vulnerable to scams. Fig. 9 shows that, contrary to expectations, Customer-Facing Roles rather than professional jobs are most targeted by scammers. Customer Service makes up 47.22% of fake job postings, followed by Other (22.22%) and Sales (11.11%). Surprisingly, the Oil & Energy and Real Estate industries dominate fake job postings, followed by the Customer Service industry.

In summary, scammers tend to offer jobs without interviews and target high school graduates seeking entry-level positions. The top 3 vulnerable industries are Oil & Energy, Real Estate, and Customer Service, while the top 3 vulnerable job functions are Customer Service, Other, and Sales. This pattern serves as a red flag for job seekers to be cautious of fraudulent opportunities.

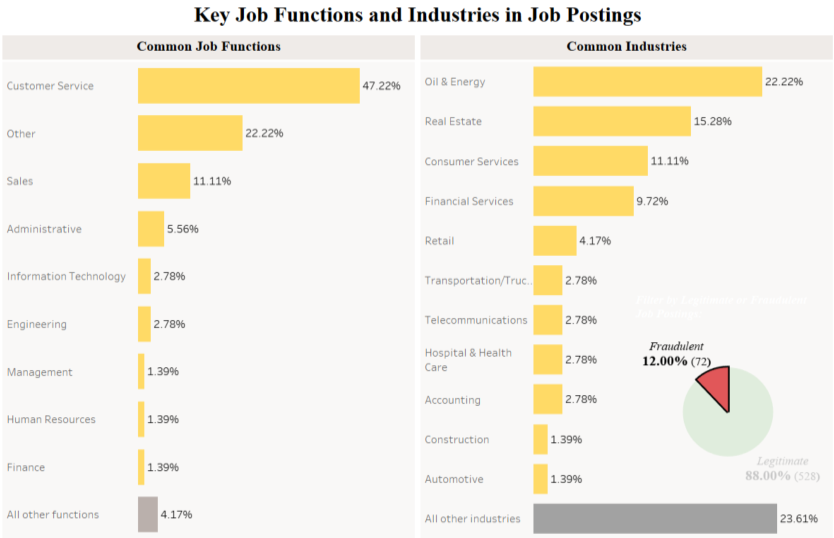


**Figure 9. Key Job Functions and Industries Filter by Fraudulent Job Postings**
*(Source: Tableau 2024.1)*

### 3.1.2 Geographical Distribution

A choropleth map that shows job postings in each country is presented in Fig.10. This map involved 16 countries which are Australia (AU), Brazil (BR), Canada (CA), Germany (DE), Egypt (EG), France (FR), United Kingdom (GB), Greece (GR), Indonesia (ID), India (IN), Iraq (IQ), Netherlands (NL), New Zealand (NZ), Pakistan (PK), Taiwan (TW) and United States (US). The map highlights that the United States (US) accounts for the highest percentage of job postings on Workable at 53.5%, followed by the United Kingdom (GB) with 22% and New Zealand (NZ) with 4.8%. Half of the job postings originate from the United States (US), suggesting it may have a higher likelihood of being targeted by scammers due to its larger user base, which increases the potential for scams.

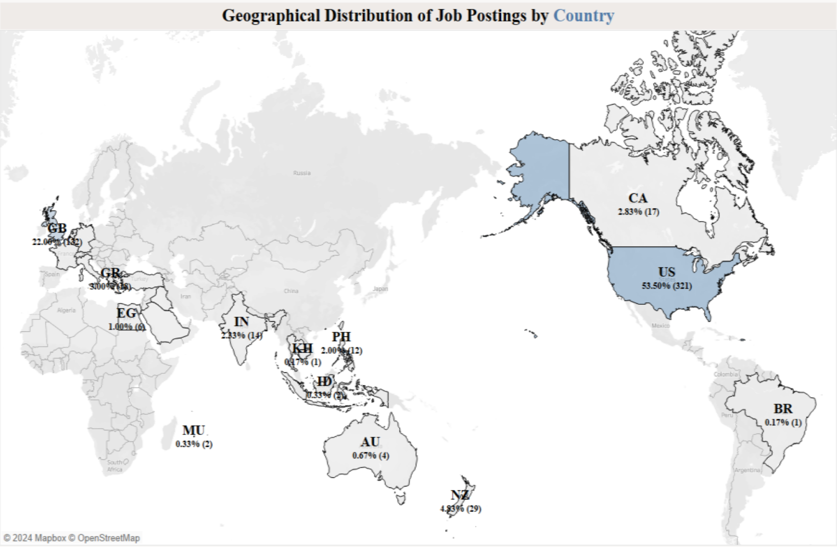
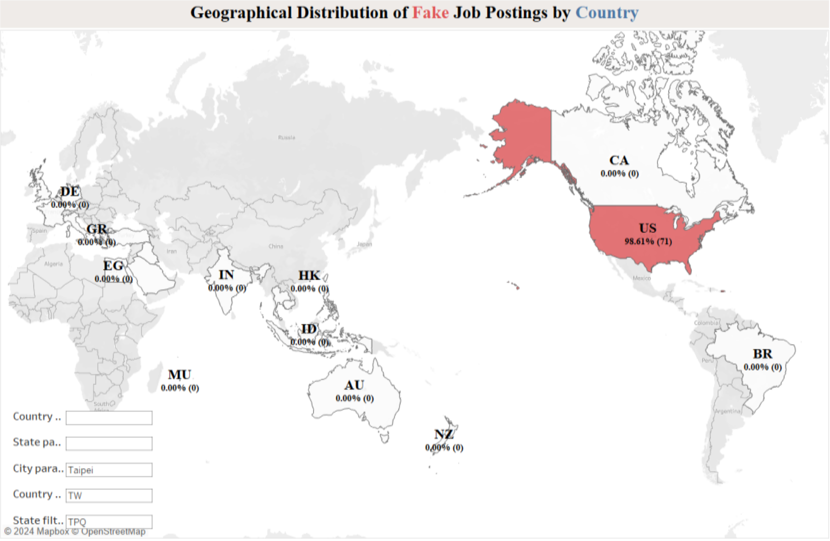


**Figure 10. Geographical Distribution by Country**
*(Source: Tableau 2024.1)*

After analyzing the distribution of job postings across countries, attention shifts to the distribution of fake job postings. Fig. 11 reveals that 71 out of 72 fraudulent job postings originate from the United States (US), underscoring a significant concentration of scams in this region. This is unsurprising, given that the US accounts for half of the platform's users. One contributing factor is the flourishing online job market in the US during challenging global economic conditions. For instance, in 2011, 650,000 new jobs were posted online in the US [23], creating a lucrative environment for scammers to exploit job seekers through fraudulent postings.



Country ..
State pa..
City para.. Taipei
Country .. TW
State filt.. TPQ

**Figure 11. Geographical Distribution of Fake Job Postings by Country**
*(Source: Tableau 2024.1)*

When the map is zoomed into city-level details, as shown in Fig. 12, it reveals that the remaining fraudulent job posting is located in Taipei, Taiwan. Among cities with fraudulent postings, Austin accounts for 25%, Bakersfield for 19.44%, and Dallas for 11.11%. Notably, both Austin and Dallas are in Texas, while Bakersfield is in California. This indicates that job postings in the states of Texas and California have a higher likelihood of being fraudulent.

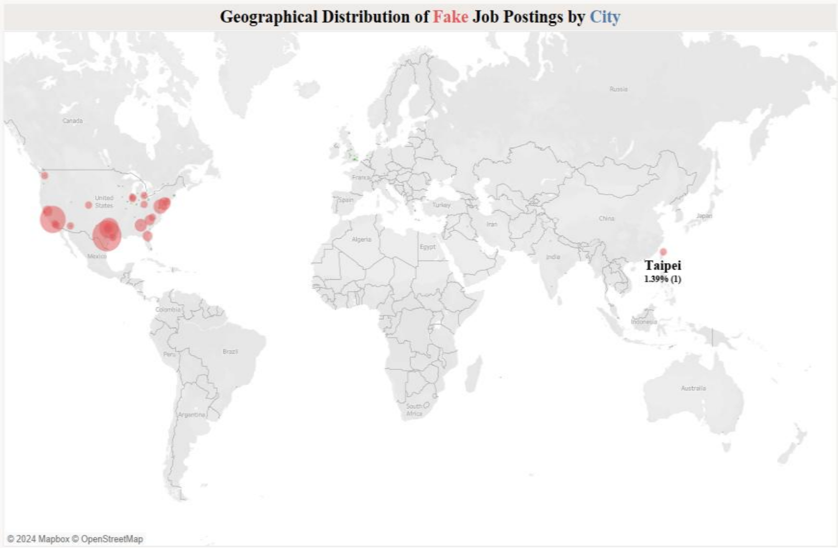
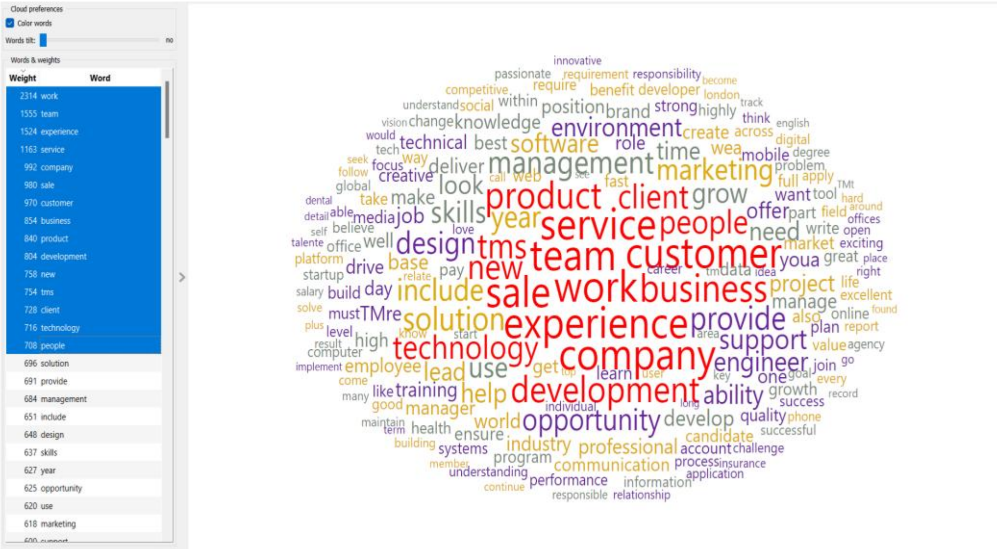


**Figure 12. Geographical Distribution of Fake Job Postings by City**
*(Source: Tableau 2024.1)*

### 3.1.3 Word Cloud

The following word cloud provides a visual representation of the 200 most frequently used words in various fields, including job title, department, company profile description, job description, job requirements, job benefits, and location city.

A significant difference between the fraudulent and non-fraudulent word clouds is that the non-fraudulent job postings include more precise requirements and comprehensive descriptions of the tasks involved, as illustrated in Fig. 13. The terms specific to legitimate job postings (e.g., "team", "business", "technology", "development", "client", and "opportunity") indicate that these positions are better structured, career-focused, and emphasize long-term growth in professionalism. Legitimate job postings often include terms that suggest organized advancement, teamwork, defined duties, and a professional workplace atmosphere.


Cloud preferences

☑ Color words

Words tilt: no

Words & weights

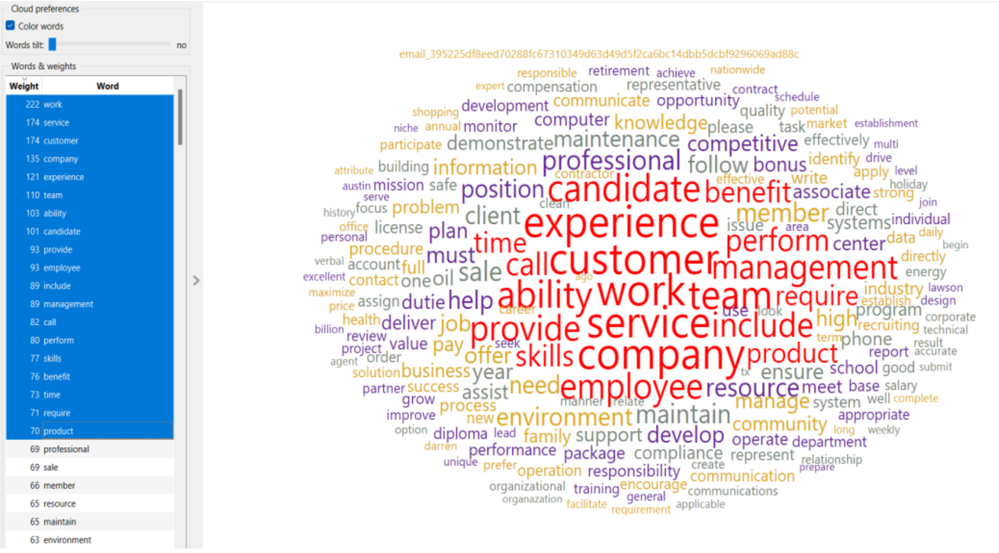
| Weight | Word |
|---|---|
| 2314 | work |
| 1555 | team |
| 1524 | experience |
| 1163 | service |
| 992 | company |
| 980 | sale |
| 970 | customer |
| 854 | business |
| 840 | product |
| 804 | development |
| 758 | new |
| 754 | tms |
| 728 | client |
| 716 | technology |
| 708 | people |
| 696 | solution |
| 691 | provide |
| 684 | management |
| 651 | include |
| 648 | design |
| 637 | skills |
| 627 | year |
| 625 | opportunity |
| 620 | use |
| 618 | marketing |
| 600 | support |

innovative
passionate requirement responsibility become
competitive require benefit developer london
understand social within position brand strong highly track
vision change knowledge environment create across think english
would technical best software role time wea mobile digital
tech way deliver management marketing degree
seek focus creative problem
follow global look call web fast full apply TMt
dental take make product client grow want tool hard
detail able media job skills year service people offer part field around
self believe love need write offices
talente office well design tms team customer market open exciting
platform drive base pay new career tmdata idea youa great place
startup relate right
salary build day include sale work business project life
solve must TMre manage excellent
plus level know start solution experience provide also online found
result high area plan report
computer technology support value agency
implement employee lead use get top company engineer join go
come learn user key one goal every
many like training help development ability growth record
good manager individual long success
maintain term health world opportunity develop quality phone
building ensure industry professional candidate successful
systems program account challenge
member understanding communication process insurance
continue performance information application
responsible relationship

**Figure 13. Word Cloud of Legitimate Job Postings**
*(Source: Orange)*

In contrast, the fraudulent word cloud mainly showcases common and widespread terms that fail to specify the job's nature, the precise requirements, or the duties, as illustrated in Fig.14. These job postings frequently depend on ambiguous phrasing and language tricks, complicating job seekers' ability to assess the authenticity of the offer. The distinctive keywords found in the fraudulent job postings (such as "ability", "candidate","call", "perform", "benefit", "require", and "member") indicate that these fraudulent job postings frequently employ general, appealing language to attract job seekers, yet they lack detailed and precise information regarding the true nature of the position. These job postings may guarantee unattainable advantages or roles that do not meet professional criteria, often focusing on recruitment or sales roles that lead to scams like multi-level marketing or commission-only schemes.

The overlap of keywords found in Figs. 13 and 14, such as "work", "service," "customer," "company", "experience", "team", and "sale", suggests that both legitimate and fake job postings have a similar surface structure. Fraudulent job listings frequently utilize typical job-related vocabulary to seem legitimate. Nonetheless, the actual job content is often vague or misleading, and such terms are occasionally employed to mislead job seekers into applying for positions that are not what they seem.



Cloud preferences
☑ Color words
Words tilt: no

Words & weights

| Weight | Word |
|---|---|
| 222 | work |
| 174 | service |
| 174 | customer |
| 135 | company |
| 121 | experience |
| 110 | team |
| 103 | ability |
| 101 | candidate |
| 93 | provide |
| 93 | employee |
| 89 | include |
| 89 | management |
| 82 | call |
| 80 | perform |
| 77 | skills |
| 76 | benefit |
| 73 | time |
| 71 | require |
| 70 | product |
| 69 | professional |
| 69 | sale |
| 66 | member |
| 65 | resource |
| 65 | maintain |
| 63 | environment |

email_395225df8eed70288fc67310349d63d49d5f2ca6bc14dbb5dcbf9296069ad88c
responsible retirement achieve nationwide
expert compensation representative contract schedule
shopping development communicate opportunity quality potential
niche annual monitor computer knowledge please task market establishment
participate demonstrate maintenance competitive effectively multi
attribute building information professional follow bonus identify drive apply level
austin mission safe contractor effective write holiday
serve position candidate benefit associate strong join
history focus problem clean client member direct
office license plan experience issue area systems individual
personal procedure time perform center data daily begin
verbal account full must directly
excellent contact one oil sale call customer management energy
maximize ago industry lawson
price assign dutie help ability work team require establish design
career use look program corporate
billion health deliver job provide service include high recruiting technical
review term phone result
project value pay seek offer skills company product report accurate
agent order
solution business year tx ensure school good submit
partner success assist need employee resource meet base salary
grow process manner relate manage system well complete
improve new environment maintain community appropriate
option diploma lead family support develop long weekly
darren performance package compliance operate department
unique prefer represent relationship create prepare
operation responsibility communication
organizational training encourage communications
organazation facilitate general applicable
requirement

**Figure 14. Word Cloud of Fraudulent Job Postings**
*(Source: Orange)*

Six distinct categories of words within the word cloud of fraudulent job postings were identified, as summarized in Table 3. These categories represent common patterns and recurring themes in the language used within fraudulent job postings. Each category encompasses a specific grouping of keywords that reflect the tactics often employed by scammers to deceive job seekers.

In conclusion, legitimate job postings feature more precise, professional, and organized language, emphasizing distinct career trajectories, technical expertise, and roles tailored to the company. Conversely, fraudulent job postings frequently employ ambiguous, general, or overly optimistic language, usually highlighting wide "opportunities" while failing to offer specific information about the position or its criteria. The inclusion of terms like "bonus", "attractive benefits", and "no experience needed" serves as a warning sign in fraudulent job postings, signaling a likely scam.

**Table 3. Content Analysis of Fraudulent Job Postings**

| No. | Category | Keywords | Possible Implications |
|---|---|---|---|
| 1 | Generic and Vague Role Descriptions | *candidate, employee, position, work, ability, associate, individual, deliver, ensure, operate, facilitate, establish, maintain, implement* | Using vague or ambiguous terminology may indicate a lack of transparency or even an intention to mislead job seekers. |

| No. | Category | Keywords | Possible Implications |
|---|---|---|---|
| 2 | Overemphasis on Benefits and Rewards | *benefit, offer, pay, package, bonus, compensation, retirement, salary, reward, success, growth, recognition* | Emphasizing appealing yet ambiguous incentives might attract job seekers while masking the authenticity of the position. |
| 3 | Ambiguous Skills, Experience, and Education | *skills, experience, perform, knowledge, must, require, training, diploma, school, license* | Lack of specific or measurable requirements might indicate fraudulent or an overly generic job description. |
| 4 | Aspirational and Positive Language | *opportunity, potential, growth, success, unique, competitive, achievement, excellent, good, effective, strong* | The overuse of motivational or exaggerated adjectives in descriptions can create an illusion of value or promise, potentially diverting attention from the absence of detailed or defined opportunities that are essential for informed decision-making. |
| 5 | Customer Service, Sales and Marketing Focus | *customer, client, service, call, contact, support, provide, help, team, member, environment, community, associate, department, sale, product, price, marketing, business, develop, market* | Roles highlighting customer service, sales or marketing without detail may target naive job seekers or be scams. Overemphasis on teamwork without clear responsibilities could also point to fabricated or unclear roles. |
| 6 | Flexible Work Scam | *part-time, flexible, remote* | Target individuals looking for convenient and adaptable work options |

### 3.2 Result of Text Clustering

A partial dendrogram of hierarchical clustering, with three distinct clusters represented by blue, red, and green colours, is presented in Fig. 15. Each cluster comprises a combination of both legitimate and fraudulent job postings that share certain common characteristics. The analysis of these clusters reveals underlying patterns or similarities among the postings, which will be discussed in detail below. This clustering approach is able to discover shared traits between fraudulent and legitimate postings, helping to identify subtle signs of fraud while also uncovering overlaps in legitimate job descriptions.

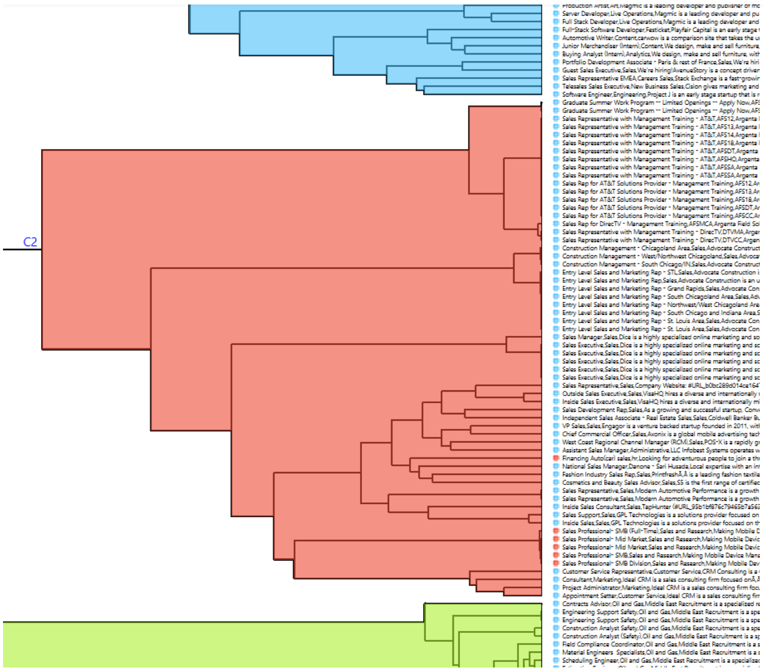


**Figure 15. Dendrogram of Hierarchical Clustering**
*(Source: Orange)*

The distribution of job postings across the three clusters is presented in Table 4. Cluster 1, represented in blue in Fig. 15, has the lowest number of fraudulent job postings, indicating it primarily comprises legitimate job postings. Cluster 2, shown in red, has the least total number of postings and a slightly higher proportion of fraudulent postings than Cluster 1. Lastly, Cluster 3, represented in green, has the highest number of fraudulent postings, reflecting a greater concentration of potentially deceptive job postings in this cluster.

**Table 4. Distribution of Job Postings Across Clusters**

| Type of Postings | Name of Cluster | | |
|---|---|---|---|
| | Cluster 1 - Innovation and Career Development | Cluster 2 - Operations and Sales | Cluster 3 - Client Relations and Support |
| Legitimate Job Postings | 253 | 56 | 219 |
| Fraudulent Job Postings | 3 | 6 | 63 |
| Total | 256 | 62 | 282 |

Cluster 1 is named Innovation and Career Development as it emphasizes roles and job descriptions related to innovation, product advancement, and technology. Cluster 1 appears to be dominated by professional services industries, which tend to have fewer fraudulent job postings. Most of the postings in Cluster 1 emphasize the creation of innovative products, encouraging creativity, and cultivating technology-oriented solutions while supporting skill enhancement and collaboration. Companies in this cluster likely operate in fast-paced environments that rely on the latest technologies to drive growth. However, for the fraudulent postings in Cluster 1, two job postings from the same company, one for a security officer and the other for a warehouse associate, use detailed-sounding duties but lack transparency about company operations, a common sign of fake job postings. Another suspicious posting for a Product Development Engineer in Taiwan offers excessive compensation, including stock options and relocation perks, but is flagged for vague contact details and unrealistic offers, typical of fraudulent listings. While the cluster primarily focuses on technology and innovation, the fake job postings are unrelated, instead targeting security, warehouse logistics, and engineering roles. The postings for Security Officer, Warehouse Associate, and Product Development Engineer exhibit common fraud indicators, including vague company details, excessive compensation, and unrealistic offers, suggesting they may be fraudulent

Cluster 2 focuses on operations, sales, and training roles where the job postings are mostly position-oriented towards sales performance, customer interaction, and training programs. Many postings also mention ‘compensation’ and ‘environment’, suggesting that these jobs probably promise competitive benefits or a workplace culture. This might be the reason why Cluster 2 has a slightly high fraud ratio due to enticing offers used in scams. For the six fake job postings identified in Cluster 2, five were found to be associated with sales jobs that require at least a high school education for an entry-level position. All five postings fall under the Computer & Network Security Industry, and all of them provide benefits like 401k (a retirement savings plan), health insurance, paid time off, vacation time, and even bonuses for employees. The sixth posting is for a contract worker associated with the automotive industry. All six fraudulent job postings share common features, such as no transparency, vague details, and inconsistent compensation structures. In this cluster, most of the fraudulent job postings are sales-related, which aligns with the legitimate postings found in cluster 2. This suggests that the more frequently a job title appears on job search platforms, the higher the likelihood it is being used for fraudulent job postings.

Cluster 3 highlights fraudulent job postings that primarily target client-facing roles such as customer service, administrative support, and retail positions. These postings are characterized by vague and overly general job descriptions that appeal to a wide audience, often listing tasks like responding to calls or preparing reports. Unrealistic perks, such as high starting salaries, guaranteed raises, biannual bonuses, and “100% dental and life insurance”, are used to lure applicants, particularly for entry- and mid-level roles where such benefits are uncommon. They frequently emphasize teamwork and integrity but fail to provide specific details, reducing their credibility. Another common tactic includes duplication of postings under different job IDs or locations, such as Austin and Dallas, to increase visibility and attract more applicants. The focus on computer proficiency and web-based applications suggests a potential aim to collect personal information rather than to recruit candidates genuinely. Overall, these postings rely on a combination of generic descriptions, exaggerated compensation, and vague promises to deceive and exploit job seekers, particularly in administrative and customer service roles. The prevalence of these tactics in Cluster 3 underscores the importance of vigilance and detailed scrutiny in identifying fraudulent job postings.

## 4. CONCLUSION

In conclusion, the Exploratory Data Analysis revealed significant insights into the characteristics of fraudulent job postings. Customer-facing roles, particularly in Customer Service (47.22%), were identified as the most targeted, surpassing professional jobs. The most vulnerable industries included Oil & Energy, Real Estate, and Customer Service. These fraudulent postings often appeal to high school graduates seeking entry-level positions and typically bypass the interview process, making them particularly deceptive. From a geographical perspective, the majority of fraudulent job postings (71 out of 72) originated in the United States, with a smaller number found in Taiwan (Taipei). Within the U.S., Texas was the leading state, accounting for 42.25% of fraudulent postings, with Austin (60%) and Dallas (26.67%) as the primary hotspots. California followed with 25.35% of fraudulent postings, most of which were concentrated in Bakersfield (77.78%). These cities may serve as hubs for job scams due to their large populations and active job markets, which provide scammers with ample opportunities to target potential victims. The word cloud analysis revealed that fraudulent job postings frequently rely on vague and generic language, while avoiding specific details about job responsibilities or requirements. Such postings often use broad, appealing language to attract candidates and make unrealistic promises about benefits or roles, adding to their deceptive nature. Moreover, clustering analysis identified three main categories of job postings. The first cluster, Innovation and Career Development, contains the least fraudulent postings as it primarily includes professional jobs. The second cluster, Operations and Sales, showed a slightly higher prevalence of fraudulent postings, mainly related to sales roles. The third cluster, Client Relations and Support, had the highest number of fraudulent postings, focusing on administrative and customer service roles that require minimal qualifications.

## Author Contributions

Chee Keong Ch'ng: Supervision, validation, writing – review & editing, project Administration. Xiang Yi Wong: Conceptualization, methodology, data curation, formal analysis, software, visualization, writing – original draft. All authors discussed the results and contributed to the final manuscript.

## Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

## Acknowledgement

The authors extend sincere thanks to the School of Quantitative Sciences, Universiti Utara Malaysia, for the kind support and assistance that led to this article. Also, thanks to Reviewers and other parties who gave some constructive suggestions to make this article more significant to contribute to academic things.

## Declarations

The authors declare no conflict of interest.

## REFERENCES

[1] Federal Trade Commission, "AMERICANS LOSE $450 MILLION TO FAKE JOB SCAMS," *Newsweek*, Apr. 30, 2024. [Online]. Available: https://www.newsweek.com/americans-lose-450-million-fake-job-scams-1895739 [Accessed: 7 December 2024]

[2] C. Reinicke, “JOB SCAMS HAVE INCREASED AS COVID-19 PUT MILLIONS OF AMERICANS OUT OF WORK. HERE’S HOW TO AVOID ONE,” *CNBC*, Oct. 6, 2020. [Online]. Available: https://www.cnbc.com/2020/10/06/job-scams-have-increased-during-the-covid-19-crisis-how-to-one.html [Accessed: 7 December 2024]

[3] PTI, “JOB SCAMS ARE ON THE RISE. WHAT ARE THEY, AND HOW CAN YOU PROTECT YOURSELF?”, *ETHRWorld.com.* May 3, 2024. [Online]. Available: https://hrsea.economictimes.indiatimes.com/news/job-scams-are-on-the-rise-what-are-they-and-how-can-you-protect-yourself/109780955#:~:text=on%20the%20rise.-,What%20are%20they%2C%20and%20how%20can%20you%20protect%20yourself%3F,compared%20to%20the%20year%20before. [Accessed: 8 December 2024]

[4] A. J. Ravenelle, E. Janko, and K. C. Kowalski, “GOOD JOBS, SCAM JOBS: DETECTING, NORMALIZING, AND INTERNALIZING ONLINE JOB SCAMS DURING THE COVID-19 PANDEMIC,” *new media & society*, vol. 24, no. 7, 1591-1610, Jul. 2022. doi: https://doi.org/10.1177/14614448221099223

[5] A. Kurtuy, “5+ COMMON JOB SCAMS IN 2024 [& HOW TO AVOID THEM!],” *Novorésumé*, Dec. 27, 2023. [Online]. Available: https://novoresume.com/career-blog/job-scams

[6] Australian Government National Anti-Scam Centre, “WARNING ISSUED ON SOCIAL MEDIA SCAMS, AS CRACKDOWN ON FAKE JOB LISTINGS CONTINUES,” *National Anti-Scam Centre*, Dec. 9, 2024. [Online]. Available: https://www.nasc.gov.au/news/warning-issued-on-social-media-scams-as-crackdown-on-fake-job-listings-continues

[7] FMT Reporters, “JOB SCAM VICTIM FORCED TO PAY RM30,000 TO RETURN TO MALAYSIA,” *Free Malaysia Today*, Feb. 9, 2024. [Online]. Available: https://www.freemalaysiatoday.com/category/nation/2024/02/09/job-scam-victim-forced-to-pay-rm30000-to-return-to-malaysia/

[8] D. Salampasis, “JOB SCAMS ARE ON THE RISE. WHAT ARE THEY, AND HOW CAN YOU PROTECT YOURSELF?”, *The Conversation*, May 1, 2024. [Online]. Available: https://theconversation.com/job-scams-are-on-the-risewhat-are-they-and-how-can-you-protect-yourself-228996

[9] P. Foran, “‘IT WAS ALL MY SAVINGS’: ONTARIO WOMAN LOSSES $15K TO FAKE WALMART JOB SCAM,” *Toronto*, Apr. 19, 2024. [Online]. Available: https://toronto.ctvnews.ca/ontario-woman-loses-15-000-to-fake-walmart-job-scam-1.6853204#:~:text=According%20to%20the%20Canadian%20Anti,to%20employment%20scams%20in%202023

[10] CNA, “MALAYSIA ARRESTS 5 PEOPLE LINKED TO JOB SCAM SYNDICATE TARGETING SINGAPOREANS,” *CNA.* Mar. 27, 2024. [Online]. Available: https://www.channelnewsasia.com/asia/malaysia-arrests-suspects-jobscam-syndicate-4225336

[11] SBS News, “EMPLOYMENT SCAMS ARE ON THE RISE. HERE’S WHAT TO LOOK OUT FOR,” *SBS News*, Oct. 23, 2023. [Online]. Available: https://www.sbs.com.au/news/article/employment-scams-are-on-the-rise-hereswhat-to-look-out-for/2xgyuapu0

[12] ET Online, “JOB SCAMS: WHO IS VULNERABLE? HOW TO PROTECT YOURSELF FROM JOB SCAMS? - job scams on the rise,” *The Economic Times*, May 3, 2024. [Online]. Available: https://economictimes.indiatimes.com/jobs/hr-policies-trends/job-scams-who-is-vulnerable-how-to-protect-yourself-from-job-scams/job-scams-on-the-rise/slideshow/109816792.cms?from=mdr

[13] A. Dutta, “ENSEMBLE CLASSIFIER: DATA MINING,” *GeeksforGeeks*, Jan. 10, 2022. [Online]. Available: https://www.geeksforgeeks.org/ensemble-classifier-data-mining/

[14] F. H. A. Shibly, S. Uzzal, and H. M. M. Naleer, “PERFORMANCE COMPARISON OF TWO CLASS BOOSTED DECISION TREE AND TWO CLASS DECISION FOREST ALGORITHMS IN PREDICTING FAKE JOB POSTINGS,” *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, pp. 2462 – 2472, Apr. 2021. http://ir.lib.seu.ac.lk/handle/123456789/5611

[15] Z. Ullah and M. Jamjoom, “A SMART SECURED FRAMEWORK FOR DETECTING AND AVERTING ONLINE RECRUITMENT FRAUD USING ENSEMBLE MACHINE LEARNING TECHNIQUES,” *PeerJ Computer Science*, vol. 9, p. e1234, Feb. 2023. doi: https://doi.org/10.7717/peerj-cs.1234

[16] D. Choudhury and T. Acharjee, “A NOVEL APPROACH TO FAKE NEWS DETECTION IN SOCIAL NETWORKS USING GENETIC ALGORITHM APPLYING MACHINE LEARNING CLASSIFIERS,” *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9029-9045, Mar. 2023. doi: https://doi.org/10.1007/s11042-022-12788-1

[17] R. Rofik, R. A. Hakim, J. Unjung, B. Prasetiyo, and M. A. Muslim, “OPTIMIZATION OF SVM AND GRADIENT BOOSTING MODELS USING GRIDSEARCHCV IN DETECTING FAKE JOB POSTINGS,” *MATRIK: Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, vol. 23, no. 2, pp. 419-430. 2024. doi: https://doi.org/10.30812/matrik.v23i2.3566

[18] D. Ranparia, S. Kumari, and A. Sahani, “FAKE JOB PREDICTION USING SEQUENTIAL NETWORK,” *in Proc. IEEE 15th Int. Conf. Industrial and Information Systems (ICIIS), Nov. 2020, pp. 339–343.* doi: https://doi.org/10.1109/ICIIS51140.2020.9342738

[19] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, “FAKE JOB DETECTION AND ANALYSIS USING MACHINE LEARNING AND DEEP LEARNING ALGORITHMS,” *Revista Geintec-Gestao Inovacao e Tecnologias*, vol. 11, no. 2, pp. 642–650, Jun. 2021. https://doi.org/10.47059/revistageintec.v11i2.1701

[20] A. Kumar, “SELF-ATTENTION GRU NETWORKS FOR FAKE JOB CLASSIFICATION,” *International Journal of Innovative Science and Research Technology*, vol. 6, no. 11, Nov. 2021. [Online]. Available: https://ijisrt.com/assets/upload/files/IJISRT21NOV109.pdf

[21] N. Goyal, N. Sachdeva, and P. Kumaraguru, “SPY THE LIE: FRAUDULENT JOBS DETECTION IN RECRUITMENT DOMAIN USING KNOWLEDGE GRAPHS,” *in Knowledge Science, Engineering and Management (KSEM 2021)*, Tokyo, Japan, Aug. 2021, pp. 612–623. https://doi.org/10.1007/978-3-030-82147-0_50

[22] Y. V. Reddy, B. S. Neeraj, K. P. Reddy, and P. B. Reddy, “ONLINE FAKE JOB ADVERT DETECTION APPLICATION USING MACHINE LEARNING,” *Journal of Engineering Sciences*, vol. 14, no. 3, pp. 310–320, Apr. 2022. doi: https://doi.org/10.1109/DELCON54057.2022.9752784

[23] ZipRecruiter, “SALARY: USD UNITED STATES,” ZipRecruiter, Sep. 2024. [Online]. Available: https://www.ziprecruiter.com/Salaries/Usd-Salary