Job Listings Integrity Investigation

Pre-processing

In this analysis of identifying fraudulent job listings from genuine ones, we begin with data cleaning and apply NLP for text preparation as dataset pre-processing.

1. Cleaning

Acknowledging common missing data, we opted to handle them thoughtfully rather than exclude entries. For text data, missing values were replaced with empty strings, while others became "Not specified" to preserve job posting completeness, crucial for legitimacy assessment. Parsing country codes from the "location" column facilitated country-wise analysis. Despite challenges in the "salary_range" column, with over 84% missing values and diverse currencies, we retained it by addressing bad values. This ensures accurate representation of salary structures across locations. Overall, this cleaning approach establishes a robust foundation for insightful EDA.

2. Natural Language Processing

In text pre-processing, we developed functions to effectively prepare data, including removing stop words and anomalous words, standardizing text with lowercase conversion, punctuation removal, and tokenization. We further enhanced the text by stemming and lemmatization to extract meaningful tokens, aiding subsequent analysis. Additionally, we implemented TF-IDF vectorization for numerical representation, crucial for downstream tasks.

EDA & Visualisations

For exploratory data analysis, we delved into the categorical columns of our dataset to extract valuable insights and patterns by visualizing distributions and relationships among them, as well as explored our text columns to identify the most common words by plotting word frequency in the form of word clouds.

1. Exploratory Plots - In our exploration of categorical variables, we observed distinct trends shaping the job market landscape.

- The exploration of categorical variables revealed intriguing patterns, with 'Full-time' employment being prevalent, suggesting a strong inclination towards stable, long-term positions among companies.
- While 'Bachelor's Degree' emerged as the most common educational requirement, there was significant demand for 'High School or equivalent', albeit with a notably higher fraudulent rate.
- Industries such as 'Information Technology and Services' and 'Computer Software and Internet' dominated job listings, while the presence of logos implied credibility, with a higher proportion of fraudulent listings lacking a company logo.
- Telecommuting options were limited, with a relatively higher fraudulent rate among remote job listings, and a geographic bias was evident, with the United States leading in job listings followed by the United Kingdom and Canada.
- Notably, a large portion of entries lacked specified employment type and degree requirements, indicating the need for further analysis of the associated text columns for potential insights.

2. Word Clouds for Text Columns

The Word Cloud analysis offers insights into recurring themes in job postings, with 'team' emerging as a central focus.

- Company Culture: 'company_profile' emphasizes services, business, and team, reflecting a client-centric culture and supportive team environment.
- Job Description: In 'description', 'teamwork' is prominent alongside client focus, business, and sales.
- Qualifications: 'requirements' highlight experience, skills, and ability, reflecting a shift towards valuing practical expertise.
- Candidate Benefits: 'benefits' include opportunity, competitive salary, and insurance promises to candidates.

Analysis Plan

This analysis plan aims to enhance the job market platform's reliability by effectively identifying fraudulent job listings. To manage the large dataset, we extracted a sample for computation efficiency. We then developed a tailored pipeline that tokenizes, vectorizes, and splits text data for classification. Initially, we employed KMeans clustering to verify classification accuracy and utilized Principal Component Analysis (PCA) for efficiency. However, as clustering is unsupervised but we already have the labels, we opted for logistic regression, training it on TF-IDF features paired with the 'fraudulent' target variable. We evaluated the model's predictive performance using classification reports on the test set.

Preliminary Results

In our pursuit of classifying text data into 'fraudulent' versus 'non-fraudulent' job listings, we employed a basic Logistic Regression model on each of the four text columns ('company_profile', 'description', 'requirements', 'benefits') individually. By passing vectorized text to the model, we obtained the following preliminary results:

- The 'company_profile' feature exhibited the highest precision (1.00) and an F1-score of 0.49 in predicting fraudulent listings, while 'requirements' displayed the lowest precision (0.53) and an F1-score of 0.31.
- Although 'description' and 'benefits' achieved a precision of 0.96, they similarly struggled with recall in accurately predicting the 'fraudulent' class.

These results underscore the model's proficiency in labeling non-fraudulent postings but highlight challenges in detecting fraudulent ones. Consequently, there is a clear imperative for further hyperparameter tuning and exploration of alternative vectorization/dimensionality reduction methods to check if the model sees improvement.

Next Steps

In our next step, we will conduct a comparative analysis across different vectorization techniques, including bag-of-words, n-gram, GloVe, and Word2Vec for dimensionality reduction. These embeddings can significantly reduce dimensions while preserving semantic meaning, potentially enhancing model performance. Also, we will integrate non-text features into the classification model to build a more holistic and robust model that leverages the full spectrum of available data.

APPENDIX

Team Contribution

Group Member	Assignments	Contribution	Status
Dian Jin	• Importing Libraries & Load the Data • Insights of EDA • Text Explanation in Data Cleaning	4	Complete
Mitchell Wu	• Markdown & Introduction • Cleaning Categorical Data • Preprocessing Text Data with NLP - Tokenization & Vectorization	4	Complete
Tanvi Sheth	• EDA - Word Cloud for Text Data • Classification results • Insights	4	Complete
Jenil Shah	• EDA - Categorical Columns • Dimensionality Reduction	4	Complete
Sneha Sunil Ekka	• Cleaning Text Data • Preprocessing Text Data with NLP – Function • Classification function	4	Complete

Github Project Link
https://github.com/dianjin0407/BA820-Group4-Job-Listing-Integrity-Investigation.git

Colab Notebook Link
https://colab.research.google.com/github/dianjin0407/BA820-Group4-Job-Listing-Integrity-Investigation/blob/main/BA820_B1_Group4_Deliverable2.ipynb