Job Listings Integrity Investigation

I. Introduction

Problem Statement: The main objective of this project is to leverage Natural Language Processing algorithms to process textual job postings and draw out patterns that distinguish fraudulent jobs from real ones. There is a critical need for an automated, reliable solution that can enhance the detection of fraudulent postings, improve platform integrity, and ensure a safe, trustworthy environment for both recruiters and job seekers.

Business Relevance: From a business standpoint, implementing a solution is crucial for platforms like LinkedIn and Handshake. By enhancing fraud detection accuracy, these platforms can create a safer environment for both - recruiters and job seekers, trust and satisfaction. Furthermore, these insights can guide strategic decisions and market trend analysis, facilitating business growth and enhancing goodwill among users.

Dataset Overview: The dataset has 17880 job postings and 18 features including job descriptions, company profiles, benefits, requirements, and a binary indicator of whether a posting is real or fake. The data consists of both textual and other relevant information about the listings. The data dictionary is mentioned in Appendix A.

II. Data Wrangling

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.

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. Although the “salary_range” column contains over 84% missing values and diverse currencies, we cleaned the data on a high level to not interfere with our analysis. Overall, this cleaning approach establishes a robust foundation for insightful EDA.

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.

III. EDA & Visualizations

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 exploring our text columns to identify the most common words by plotting word frequency in the form of word clouds.

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.

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

- Job Title: The most popular job “titles” are those of manager and developer, followed by sales, service, engineer.
- 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.

IV. Methodology

This analysis 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.

V. Data Analysis & Modeling

Vectorization and Fine-tune: After conducting data cleaning, tokenization, and exploring data insight from the EDA process, we embarked on a series of steps to enhance the effectiveness of text-based data for classification tasks. To better capture the semantic nuances, we decided to use Word2Vec for vectorization since TF-IDF represents words only based on their frequency of occurrence within and across documents but Word2Vec embeds words into a continuous vector space where similar words are positioned closer to each other. We fine-tuned the model's sensitivity to word co-occurrences by systematically tuning Word2Vec parameters such as vector size, window size, and min_count of the word, aiming to optimize the quality of embeddings. These embeddings capture semantic similarities and relationships between words or phrases, enabling a better understanding of the main textual columns in our dataset and ultimately enhancing the effectiveness of downstream classification tasks.

Dimensionality Reduction using PCA: After obtaining embeddings for these text columns, we have employed dimensionality reduction techniques such as Principal Component Analysis (PCA) to condense the high-dimensional embeddings into a lower-dimensional space. We plotted the explained variance ratios to visualize the components that explain 95% of the variance. This not only helps in reducing computational complexity but also aids in visualizing and interpreting the data more effectively.

Combining features and SMOTE: Subsequently, we integrated the other features in the dataset with the reduced-dimensional embeddings we got. This holistic representation captures both textual and non-textual information, providing a comprehensive input for classification models. However, we noticed an imbalance in the dataset concerning the distribution of labels since there are only 4.8% of all job postings classified as fraudulent, which can adversely affect model performance. To address this issue, we employed SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples of the minority class, thereby balancing the dataset and mitigating bias towards the majority class.

Classification: After considering various classification models and hyper-tuning them with HalvingRandomSearchCV to find the best parameters. While the stacking classifier has yielded a near-perfect cross-validation accuracy of 99.9%, we need to consider this with caution given the possibility of overfitting. We have a more realistic result with the Stacking Classifier yielding a balanced accuracy of 80%.

VI. Challenges

The project faced several challenges, notably with the "salary_range" column, which was largely composed of invalid or missing data, comprising 85% of the dataset, and therefore couldn't be used for analysis. Additionally, the variability of salary standards across different countries presented a challenge in normalizing salary data to a uniform currency for global comparison. These issues hindered the incorporation of potentially insightful salary data into the analysis, which could have enriched the analysis.

VII. Limitations

Firstly, due to the limited time and resources, we were unable to check all possible combinations of the word2vec hyper-parameters and evaluate them using classifiers. Therefore, instead of an exhaustive analysis of the hyper-parameters, we considered some predefined value combinations for arguments as discussed in class. What's more, although the significant class imbalance within the dataset was addressed using SMOTE, the effectiveness of this method on the model's performance remains uncertain. With appropriate time and resources, this could be explored further to ensure we train the model on the best version of our dataset. Last but not least, we had planned on comparing our model across vectorizers - word2vec and GloVe to evaluate their performance, but we were unable to find an appropriate pre-trained GloVe model on job listings. It would have been more insightful if we would have been able to implement the vectorizer comparison.

VIII. Conclusion

With this project, we have successfully implemented Natural Language Processing (NLP) techniques to examine textual data and integrated those findings with other techniques including dimensionality reduction, handling class imbalance and finally with classification models. In terms of the business implications, we believe that this model is a great example and a viable starting point for a much more complex project dealing with the integrity of job listings on any website or platform. By implementing this model, the platforms can very easily verify the legitimacy of a job posting without compromising the trust of not only the job seekers, but also other companies posting real job listings.

APPENDIX A

Data Dictionary

Column Name / Features	Description
job_id	Unique Job ID
title	The title of the job ad entry
location	Geographical location of the job ad
department	Corporate department (e.g. sales)
salary_range	Indicative salary range
company_profile	A brief company description
description	The details description of the job ad
requirements	Enlisted requirements for the job opening
benefits	Enlisted offered benefits by the employer
telecommuting	True for telecommuting positions
has_company_logo	True if company logo is present
has_questions	True if screening questions are present
employment_type	Full-type, Part-time, Contract, etc.
required_experience	Executive, Entry level, Intern, etc.
required_education	Doctorate, Master's Degree, Bachelor, etc.
industry	Automotive, IT, Health care, Real estate, etc.
function	Consulting, Engineering, Research, Sales etc.
fraudulent	target - Classification attribute

APPENDIX B

Team Contribution for Coding

Group Member	Assignments	Final assignments	Contribution	Status
Dian Jin	• Importing Libraries & Load the Data • Insights of EDA • Text Explanation in Data Cleaning	• Mega Function & Combine Embeddings • SVC model • Add "Title" Column for Word-cloud & Classification	7	Complete
Mingze Wu	• Markdown & Introduction • Cleaning Categorical Data • Preprocessing Text Data with NLP - Tokenization & Vectorization	• Token & Word2Vec Functions • Random Forest Model • Explanation for 'Next Steps' & Basic Functions	7	Complete
Tanvi Sheth	• EDA - Word Cloud for Text Data	• Import Packages & Define Model Functions • Logistic Regression Model • Result Table & Challenges and Limitations & Conclusion	7	Complete
Jenil Shah	• EDA - Categorical Columns	• PCA & Variance Plot • XGB Model • Explanation for Hyper-parameter Tuning & Classification Model	7	Complete
Sneha Sunil Ekka	• Cleaning Text Data • Preprocessing Text Data with NLP - Function	• Combining All Features & Smote • Stacking Model • Explanation for Word2Vec & Combining Data & SMOTE	7	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_Deliverable3.ipynb