COMBATING EMPLOYMENT SCAMS: TRUE AND FAKE JOB CLASSIFICATION WITH NLP

Authors Pravin Tahiliani
Romil Jain
Harmish Doshi
Harvinder Singh Laliya

Course:
IST 668 / CIS 664 - Natural Language Processing

01 Introduction

Employment scams have been on the rise, with the number of cases doubling in 2018 compared to 2017. This project aims to create a classifier using Machine Learning and NLP to identify real and fake job postings, ensuring users are provided with reliable and legitimate employment opportunities.

Overview of Employment Scams

Market Situation
High unemployment and fewer job opportunities make people more vulnerable to scammers.

1. **Rising Scam Cases**
The number of employment scam cases doubled in 2018 compared to 2017.

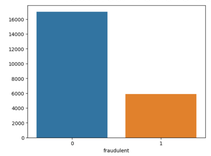
2. **Reputation Impact**
Fraudulent job listings mislead job seekers and affect the reputation of job portals.

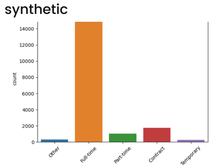
02 Objective

Addressing the issue of employment scams through advanced technologies like Machine Learning and NLP. Creating a classifier to accurately distinguish between genuine and fake job advertisements, protecting job seekers from giving out personal information to scammers.

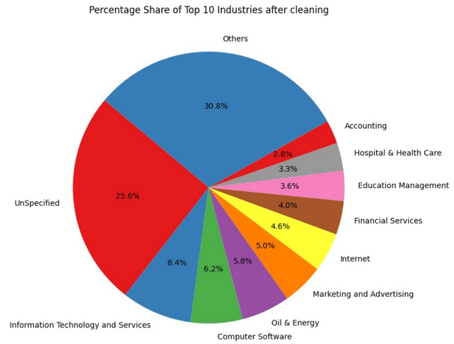
03 Dataset

- Used a diverse dataset from University of the Aegean dataset containing job postings categorized as fraudulent or not.
- Dataset includes job title, locations, departments, salary ranges, employment types, company profiles, and detailed descriptions
- The dataset consists of 17,880 observations and 18 features. We have added 5000 synthetic

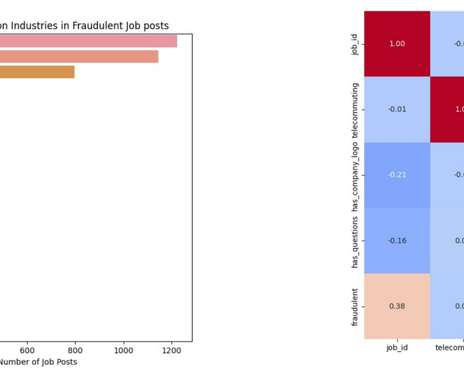




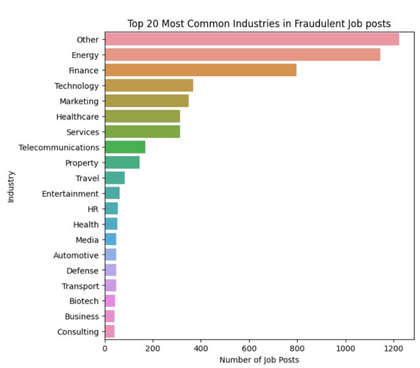
04 Exploratory Data Analysis



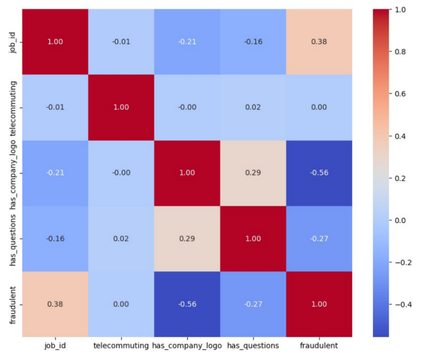
Percentage Share of Top 10 Industries



Top 20 Most Common Industries In Real Job Posts



Top 20 Most Common Industries In Fraud Job Posts



Correlation Matrix
The correlation matric for numeric data showed no strong correlations

05 Data Cleaning

- Data Augmentation – Employed data augmentation by integrating synthetic data to enhance the diversity and volume of the dataset for improved model training and generalization.
- Handling Null Values - Null values in columns excluding 'job_id,' 'telecommuting,' 'has_company_logo,' 'has_questions,' and 'salary_range' were replaced with 'Unspecified,' while null values in the specified columns were retained as NaN.
- Mapping Education Level - A custom function was applied to standardize diverse education labels in job postings, grouping similar qualifications (e.g., "Bachelor's Degree" and "Bachelor's or Equivalent") into a unified column for streamlined analysis.
- Location - A dedicated column was created in the job postings dataset to store the extracted two-character country codes from the inconsistent location field, facilitating standardized analysis and categorization based on country codes.
- Industry Mapping- Job descriptions in the 'Technology' category, specifically in 'Computer Software,' 'Information Technology and Services,' and 'Computer Hardware,' were reorganized based on common traits to facilitate easier understanding and comparison of roles in the tech field.
- Number to Text Transformation - Mapping binary values in job-related columns to descriptive labels like 'is_absent'/'is_present' and 'not_allowed'/'is_allowed' for telecommuting and work-from-home permissions.
- Merging of Text Columns – We consolidated text columns into a unified 'full_text' column, excluding the numeric columns, to enhance the efficiency of textual analysis.

Data Preprocessing

- Tokenization: Segmented the full text into individual words to facilitate further analysis.
- Lowercasing: Ensured consistency by converting all text to lowercase.
- Removal of Non-Alphabetic Characters: Filtered out non-alphabetic characters to retain only meaningful words.
- Stop Word Removal: Eliminated common English stop words to enhance the relevance of the text.
- Lemmatization: Reduced words to their base or root form for improved consistency in language.
- Part-of-Speech Tagging: Assigned grammatical categories to each word, providing insights into syntactic structure.
- Detokenization: Reconstructed preprocessed tokens into coherent text for further analysis.
- Vectorization- Vectorized the data with different methods.
  Improved Data Representation-

1. Vectorization methods like TF-IDF, count vectorization, word2vec, and BERT were implemented to enhance analysis by converting text into numerical vectors, and capturing semantic relationships and context for better data representation.
2. Versatile Model Compatibility-Vectorization enabled compatibility with most of the machine learning and neural network models, fostering flexibility and enhancing model performance by providing meaningful numerical representations of textual data.

06 Results/Findings

Algorithms and Evaluation

	Naïve Bayes	Random Forest	SVM	RNN	LSTM
TFIDF	Accuracy 0.97 Precision 0.97 Recall 0.97 F1 Score 0.97	Accuracy 0.95 Precision 0.95 Recall 0.95 F1 Score 0.94	Accuracy 0.98 Precision 0.98 Recall 0.98 F1 Score 0.98	Accuracy 0.95 Precision 0.95 Recall 0.95 F1 Score 0.95	Accuracy 0.97 Precision 0.97 Recall 0.97 F1 Score 0.96
Count Vector	Accuracy 0.97 Precision 0.97 Recall 0.97 F1 Score 0.97	Accuracy 0.91 Precision 0.92 Recall 0.91 F1 Score 0.90	Accuracy 0.98 Precision 0.98 Recall 0.98 F1 Score 0.98	Accuracy 0.96 Precision 0.96 Recall 0.96 F1 Score 0.96	Accuracy 0.95 Precision 0.95 Recall 0.95 F1 Score 0.95
Word2vec	Accuracy N/A Precision N/A Recall N/A F1 Score N/A	Accuracy 0.94 Precision 0.94 Recall 0.94 F1 Score 0.94	Accuracy 0.96 Precision 0.96 Recall 0.96 F1 Score 0.96	Accuracy 0.99 Precision 0.99 Recall 0.99 F1 Score 0.99	Accuracy 0.99 Precision 0.99 Recall 0.99 F1 Score 0.99

07 Conclusion

In our project to classify real and fake jobs, we employed various machine learning models. All of the models achieved satisfactory performance on the classification task, with Naive Bayes, Random Forest, and SVM achieving accuracies between 97% and 98%. However, the classification reports indicated that these models were overfitting the data, as they exhibited high precision and recall scores on the training data but lower scores on the validation data. To address this overfitting issue, our group switched to neural network models, namely RNN and LSTM. These models demonstrated superior performance on the validation data, achieving accuracies of 99%. This suggests that RNN and LSTM models possess a better ability to generalize to unseen data and are less susceptible to overfitting.

08 References

1. https://www.ijraset.com/best-journal/fake-job-detection-using-machine-learning
2. https://www.researchgate.net/publication/371508732_Fake_Job_Detection_with_Machine_Learning_A_Comparison
3. https://turcomat.org/index.php/turkbilmat/article/view/13533
4. https://www.researchgate.net/publication/360849325_A_machine_learning_approach_to_detecting_fraudulent_job_types
5. https://www.hindawi.com/journals/js/2022/4583512/
6. https://www.trendytechjournals.com/files/issues/volume6/issue1-3.pdf
7. https://link.springer.com/article/10.1007/s11063-021-10727-z
8. Dataset Link - http://emscad.samos.aegean.gr/
9. Synthetic Data Generation - https://gretel.ai/