Job Listing Integrity Investigation

Motivation

As graduate students entering the job market, we spend most of our time exploring job opportunities online. It is important for us to be able to trust the platform where we share our data. This is why having the knowledge of real or fake job postings is so important.

Problem Statement

The main objective of this project is to leverage Natural Language Processing algorithms in order 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.

Dataset & Data Source

The dataset - Real / Fake Job Posting Prediction is from Kaggle, retrieved from this link[1]. 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 information and meta-information about the jobs.

Proposed Methodologies

We aim to use a combination of unsupervised and unstructured machine learning techniques to analyze the dataset holistically. We aim to incorporate the following methods:

- **NLP:** To process textual data with the help of tokenization & vectorization, and extract meaningful information from job postings, facilitating semantic understanding of the data.
- **Topic Modeling:** To uncover thematic structures in job descriptions, distinguishing between genuine and fraudulent postings based on content and language.
- **Anomaly Detection:** To identify postings deviating significantly from typical patterns, signaling potential fraud.
- **Clustering:** To group job postings without their labels and verify if they naturally separate into two distinct clusters based on their content.
- **Association Rule Mining:** To discover frequent combinations of job attributes that co-occur in postings, revealing patterns distinguishing real from fake postings.

These techniques will be complemented by supervised learning models for validation and performance enhancement. The combination of unsupervised and supervised methodologies will allow for a robust, comprehensive approach to detecting fraudulent job postings.

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.

[1] Kaggle acknowledges the University of the Aegean's Laboratory of Information & Communication Systems Security, led by the team at http://emscad.samos.aegean.gr/, for its contribution.