Final Project Report

Freddy Chen ▮▮▮▮▮

Keani Schuller ▮▮▮▮▮

Kelly Kao ▮▮▮▮▮

Ko-Jen Wang ▮▮▮▮▮

Xinran Yu ▮▮▮▮▮

▮▮▮▮▮▮▮▮

▮▮▮▮▮

▮▮▮▮▮▮

Introduction & Problem

Fake job postings are becoming a concerning issue for job seekers, employers, and job ad platforms. These deceptive listings often mislead job seekers and pose significant challenges to individuals searching for employment opportunities and legitimate businesses striving to attract top talent. Therefore, all job platforms want to filter out fake job postings. Filtering out fake job postings can prevent job seekers from wasting resources on creating applications and preparing for interviews for non-existent jobs, as well as protect them from being scammed. For job ad platforms, maintaining reliability in job postings is crucial for retaining their users and trust in their services. The project aims to identify the fake job postings through text analysis and classify them into real and fake groups. Furthermore, it aims to assist in the mitigation of this issue by providing insights and solutions to combat the occurrence of deceptive job listings, ultimately fostering a more trustworthy job search market for all stakeholders.

Analytics Approach

We carefully selected two classifiers known for their strong performance in text classification tasks: Multinomial Naive Bayes and Support Vector Classifier (SVC). These two models were selected because of their complementary nature. MultinomialNB assumes that all features (in this case, words) are independent of each other, which is a simplification of reality but works well in practice for text. SVC does not make this assumption and instead tries to find the best boundary that separates the classes, considering the data's complexity.

To rigorously test these classifiers, we conducted 12 distinct experiments. This allowed us to explore how different preprocessing techniques, like lemmatization (which reduces words to their base or root form), and variations in the text-to-numeric vectorization process might affect our models' ability to correctly identify fraudulent job postings. Our experiments were

designed to find the sweet spot for each classifier by tweaking settings, such as the range of word combinations considered (n-grams) and the specific parameters of the models (like complexity, kernel, and margin for the SVC, and smoothing parameter for MultinomialNB). This hyperparameter optimization process is crucial for enhancing our model performance.

The Expected Impact

The analysis of fake job postings has been explored before, particularly by job platforms in an effort to ensure the integrity of their listings. Such efforts use similar models who distinguish between real and fake listings based on aspects like textual features, patterns, and posting frequency. The successful filtering of fake job postings has significant positive impacts. It enhances the credibility of job platforms and the companies that rely on them for recruitment, ensuring that only genuine vacancies are advertised and reducing the risk of reputational damage associated with fake listings. It also protects job seekers from fraud, contributing to a safer and more trustworthy job search environment. This can lead to increased user confidence in online job search platforms, fostering a more reliable digital economy.

Lastly, the topic of fake job postings is important due to its impact on individuals' livelihoods and their trust in the digital job market. As mentioned above, for job seekers, fake postings represent a waste of time and resources. They also risk exposure to financial scams or identity theft. For employers and job platforms, the presence of fake postings undermines their credibility and can deter both job seekers and real companies from using their services, impacting their operational effectiveness and profitability. Overall, the integrity of the job market is a foundational element for stability and growth in industries. making fake job posting detection essential for maintaining a healthy employment ecosystem and protecting the interests of all stakeholders involved.