COGS209/CSS206 Mini Project Proposal

Lead Author: Asad Tariq
Co-Authors: Isabella Mullen, Luna Bellitto, Nina Rice

Title: Detecting Scams in Job Postings

Research Question

Searching for a job is a fundamental step of a person's professional life. In recent years, online platforms, such as LinkedIn, have become the most used tools to find employment. However, the growing popularity of posting job opportunities online on various websites has led to an increase in fraudulent job postings, making it challenging for users to distinguish legitimate opportunities from scams. We believe that the employment of statistical models to successfully classify job postings will help us detect fake job postings and highlight patterns among them. We hypothesize that classification models like logistic regression, k-nearest neighbors, and support vector machines will be able to correctly classify a real or fraudulent job posting based on textual patterns and keywords. This analysis can assist in developing tools able to filter different job offers, reducing the risks of scams and increasing the safety and efficiency of the job search process on online platforms.

Data/Materials

We will use a dataset on job listings obtained from Kaggle (Real/Fake Job Postings) which contains a total of 18,000 job descriptions, of which 800 are fake. This data is also available in our project's GitHub repository. The data subfolder in the proposal folder contains the following file:

File name	Format	Description
fake_job_postings.csv	Comma Separated Values (CSV)	Contains a table of 18,000 job descriptions along with other relevant job details such as title, location, department, and salary range. Out of these 18,000 jobs, 800 are known to be fake/scams.

We will extract the features from this dataset that are relevant to our research question, such as the location of the job, the type of the job, the department, whether the hiring company has a logo or not, the required education for the job, and the required experience for the job. We will also take into account the imbalance between the categories.

Course Impact/Relevance

This project will integrate several topics from the course. We will begin with data exploration to identify patterns in job features and build a classification model using logistic regression, k-nearest neighbors (KNN) and support vector machines (SVMs). We believe that using a non-parametric method like SVM will capture complex relationships in the data and help in robust classification. The project will also make use of cross-validation for evaluating the models.

Outcome(s):

The expected outcome of this project is a classification model that can accurately predict whether a job posting is fake. By applying logistic regression, k-nearest neighbors, and support vector machines. We aim to identify which model performs best in classification and minimizing false positives. This will provide insight into which features of a job posting are most predictive of a fake post. We will evaluate the model performance by using cross validation. We will also create a confusion matrix and generate ROC and PRC curves, AUC values, and average f1 score to assess model accuracy.