Fake Job Posting Detection - Group 4

Members: Xuanjie Huang, Elliot Kim, Lechuan Meng, Air Singh, Catherine Win

Team Leader: Air Singh

Problem Statement:

The prevalence of fake job postings has become a significant issue in recent years. These fraudulent postings can mislead job seekers by wasting their time and resources and scamming them. In this project, we aim to develop a model that can accurately classify job postings as either real or fake, helping job seekers and job platforms identify and filter out fraudulent listings.

Dataset Description:

We will be using the "Real or Fake: Fake Job Posting Prediction" dataset available on Kaggle. This dataset consists of over 17,000 job listings, with 866 unique features such as job title, location, company profile, and job description. Each entry is labeled as either real or fake, making it suitable for a supervised learning approach.

Goals:

We aim to be able to have a final model that has high accuracy and precision in classifying whether the job postings are real or fake. In order to do this, we will be applying a variety of machine learning classification algorithms in order to find the best model for our case. We plan on using more simple algorithms such as logistic regression to much more complex models such as neural networks. We also plan on taking a look at other algorithms such as random forest, support vector machines, naive bayes, and k-nearest neighbors.

Timeline:

4/23 - Finish with exploratory data analysis in order to get a better understanding of the data. Also setup github for collaboration

4/30 - Finish with literature review to gain a better understanding of methods used with natural language processing

5/7 - Finish preprocessing data and perform feature selection

5/14 - Finish building different models and testing performance

5/21 - Finish improving accuracy through hyperparameter tuning and cross validation

5/28 - Finish writing the final report and finish developing the front-end aspect to display the model