FAKE JOB POSTING PREDICTION

First name	Last Name	IIT Email
Divya	Poojari	dpoojari@hawk.iit.edu
Tanmay	Patil	tpatil3@hawk.iit.edu

Introduction

Fake job posts are everywhere, wasting time of dozens or hundreds of aspiring applicants. A single fake job posting can lead to deceitful crowd gaining a wealth of information. Some of them lead to nothing more than adding email addresses to the spam distribution lists and some of the worst cases involve actual thieves trying to cull sensitive information for deviant activities.

Through this project, we are trying to build classification models that can help predict such fraudulent job postings.

Data Sets

We have chosen the dataset provided by Kaggle to predict Fake Job Description Prediction based on various parameters such as job_id, title, location, department,salary_range, company_profile, etc.

The Dataset has been taken from the given link below:

https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction

The Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset containing 17,880 real-life job ads that aims at providing a clear picture of the Employment Scam problem to the research community and can act as a valuable testbed for scientists working on the field.

EMSCAD records were manually annotated and classified into two categories. More specifically, the dataset contains 17,014 legitimate and 866 fraudulent job ads published between 2012 to 2014.

Research Problems

Below are the problems we are looking forward to resolve:

1. Identification of key traits or features like words/phrases/entities of job descriptions which are fraudulent in nature.

2. Usage of text and metadata features and predict which job descriptions are real and fake.

Summary: With the help of exploratory data analysis and classification models, we would be identifying interesting insights and run a contextual embedding model from this dataset.

Potential Solutions

Solution 01: Various text pre-processing approaches to clean the text and search for the most frequent words and feature extraction functions are used to identify key traits of job postings which are fraudulent in nature and this analysis will set the fraudulent flag to 0(in case of fake) and 1(in case of real) accordingly.

Solution 02: We would be using ML algorithms like KNN algorithm to find significant variables, build classification models using the text features in the job description column to predict fake job postings and will then be compared to each other on the basis of its RMSE value, simulation run-time, accuracy, sensitivity and specificity.

Solution 03: We would be using two-sample hypothesis testing to prove if the job posting is fake or not based on the dependent attributes in the dataset.

Evaluations

We will be using hold-out evaluation where the dataset will be divided into training dataset(90% of data) and validation dataset(10% of data). Entire classification modelling will be done on the training dataset and will be later used on the validation dataset to predict its accuracy.

Expected Outcomes

Outcome 01: Classification of frequent words from job description of genuine job posting and then on the fake job postings followed by comparison and extraction of frequent words present only in the fake ones.

Outcome 02: The dependent variable ‘fraudulent’ has already been added to the dataset to use it as outcome which will be set as 0 for fake postings and 1 for real postings.