

Real / Fake Job Posting Prediction

Information and Network Security



Dinesh Marupudi Veera – 999901286

Vivek Kiran Kataram – 999902375

Uditha ReddyKattegummula - 999902201

**Abstract:**

Employment fraud is becoming more prevalent. According to CNBC, there were twice as many employment frauds in 2018 than there were in 2017. The state of the market today has resulted in substantial unemployment. Numerous people have experienced much less job loss and economic stress because of the coronavirus. Such a situation offers con artists the ideal opening. Due to a rare incidence, many people are becoming victims of scammers who feed on their despair. Most con artists use this technique to obtain personal information from their victims. Addresses, bank account information, social security numbers, and other personal information are examples. As a student, I have encountered several of these fraudulent emails. The con artists offer their victims incredibly lucrative career opportunities and then demand payment in exchange. Or they demand money from the job seeker in exchange for the promise of employment. Natural Language Processing and machine learning approaches can be used to solve this severe dilemma (NLP).

This project makes use of data from [Kaggle](https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction). The characteristics of a job ad are contained in this data. These job listings are either labeled as genuine or bogus. A very small portion of this collection consists of fake job listings. That is acceptable. We don't anticipate seeing many fake job postings. This project is divided into five stages. The five steps of developing our machine learning model include:

Diagram

Description automatically generated **Figure 1: Fake jobs machine learning stages**

**Research Problem:**

The goal of this project is to develop a classifier that can distinguish between phony and legitimate jobs. Two distinct models are used to evaluate the outcome. One model will be used on the text data and another on the numeric data since the submitted data comprises both text and numeric properties. The result will be a synthesis of the two. The final model will incorporate all pertinent information from job postings and generate a conclusion on whether the position is legitimate or not.

1. What are the most important features in predicting whether a job posting is real or fake?
2. How accurate are different machine learning algorithms in predicting whether a job posting is real or fake?
3. Are there any significant differences in the characteristics of real job postings compared to fake job postings?
4. Can natural language processing techniques improve the accuracy of predicting whether a job posting is real or fake?
5. Is there a correlation between the geographical location of the job posting and its likelihood of being fake?

**Introduction:**

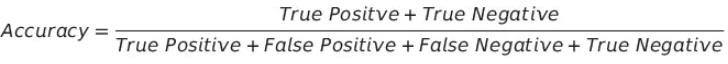
The development of the internet has significantly streamlined the hiring process. Additionally, the ongoing pandemic has significantly contributed to the present shift in the pattern of job recruitment. Online hiring has made it possible to find more prospects and streamline the hiring process, and it has been very helpful in bridging the gap between recruiters and potential candidates. With only the press of a mouse, candidates may now apply online to a wide number of positions based on their area of expertise. E-recruitment assists businesses in utilizing a variety of internet-based options. Users can broaden their employment searches and find the best applicants by using online recruitment where they can converse with competent prospects from around the world. When a client relies on internet recruitment, the best candidate is ultimately hired. Candidates' online personas can be discovered through social media sites like Facebook and LinkedIn. Companies can choose competent individuals and increase efficiency by using tools like pre-employment screening, personality assessments, and tests for candidate screening.

The process of online recruitment involves minimal human interaction. Communication costs are lower, making it more cost-effective. However, some advertised job openings are fraudulent and serve as bait to obtain personal data, instead of genuine job opportunities. When candidates apply for these jobs, their potential information is stolen, or hackers gain access to their computers to steal important data. Cybercriminals may combine victim data and sell it on the dark web or continue to use it for years. Research has been conducted on detecting fake job postings using machine learning algorithms and ensemble classification modelling to enhance accuracy. By utilizing an appropriate dataset, sufficient analysis, cleaning, and machine learning techniques, the identification of false job postings is possible.

**Dataset exploration Metrics:**

The models are evaluated based on two metrics:

Accuracy: This formula defines this metric –



As the formula suggests, this metric produces a ratio of all correctly categorized data points to all data points. This is particularly useful since we are trying to identify both real and fake jobs, unlike a scenario where only one category is essential. There is, however, one drawback to this metric. Machine learning algorithms tend to favor dominant classes. Since our classes are highly unbalanced, a high accuracy would represent how well our model categorizes the negative category (real jobs).

Text

Description automatically generated

F1-score is used because, in this scenario, both false negatives and false positives are crucial. This model needs to identify both categories with the highest possible score since both have high costs.

https://[www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction](http://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction) has the data for this experiment. 17,880 observations and 18 characteristics make up the dataset. The information consists of three different data types: integer, boolean and text. Below is a quick definition of the variables:

A black screen with white text

Description automatically generated with low confidence

# Figure 2: Model feature description

A summary statistic is not required in this case because most of the datatypes are either booleans or text. Job\_id is the lone integer, which is irrelevant for our investigation. To find null values, the dataset is further examined.

Table

Description automatically generated

# Figure 2: The null value dataset count

The values for many variables, including department and salary range, are missing. The remaining analysis skips these columns.

A first analysis of the dataset revealed that the job posts were in various languages because e they were taken from various nations. Nearly 60% of the dataset used for this research comes from US-based sources. All the data is in English for simple interpretation thanks to the data from US- based locales. For additional study, the location is divided into the state and the city. The final dataset has 20 features and 10593 observations.

Chart

Description automatically generated

# Figure 3: dataset correlation matrix

With only 725 or 7% of the jobs in the sample being fraudulent, only 9868 or 93% of the positions in the dataset are actual. a chart of counts can very clearly demonstrate the discrepancy.

**Future Work: Algorithm and Techniques**

It is clear from the preliminary study that text and numerical data must both be employed for the final modeling. A final dataset is selected prior to data modeling. For the final analysis, this project will employ a dataset with the following characteristics:

1. Telecommuting
2. Fraudulent
3. ratio: fake to real job ratio based on location
4. text: a combination of title, location, company\_profile, description, requirements, benefits, required\_experience, required\_education, industry, and function
5. character\_count: Count of words in the textual data Word count histogram
6. Further pre-processing is required before textual data is used for any data modeling.

The algorithms and techniques that we expected to use in the project are:

* Naïve Bayes Algorithm
* SGD Classifier
* Natural Language Processing

**Sample Code- Processing:**

Below is the list of sample data pre-processing and processing steps that were followed.

1. Screenshot below shows the removal of undesired columns from dataset and clearing out the null and nan values in text and categories columns. The last step shows the result of total real vs fake jobs – which is about 5% of total postings.

Graphical user interface, text, application

Description automatically generated

1. Screenshot below shows the pie-chart plotting for the said ratio.

Chart

Description automatically generated

1. Screenshot below shows a part of the step where the difference of real and fake postings in various scenarios are defined and plotted accordingly.

Scatter chart

Description automatically generated

1. Screenshot below shows various scenarios of plotting between real and fake jobs.

A picture containing timeline

Description automatically generated

1. Screenshot below shows the plotting of top 3 list of categories.

Chart, waterfall chart

Description automatically generated

**References:**

1. Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In International conference on intelligent, secure, and dependable systems in distributed and cloud environments, 127-138.
2. Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-36.
3. <https://www.kaggle.com/code/shivamburnwal/nlp-98-acc-eda-with-model-using-spacy-pipeline>
4. <https://www.kaggle.com/code/madz2000/text-classification-using-keras-nb-97-accuracy>