PREMIER UNIVERSITY, CHATTOGRAM

Department of Computer Science & Engineering

6th Semester Artificial Intelligence Laboratory Report

On

Real and Fake Job Prediction Using Artificial Intelligence

SUBMITTED BY

Name: Mohammed Shahadat Hossain Talukder

ID: 1803410201573

Section: A1

SUBMITTED TO

Avisheak Das

Lecturer

Department of Computer Science & Engineering

Premier University, Chattogram

Table of Contents

1 Introduction **1**

1.1 Background 1

1.2 Motivation 2

2 Literature Review **3**

3 Methodology **4**

3.1 Proposed Methodology 4

3.1.1 Data collection 4

3.1.2 Data Pre-processing 5

3.1.3 Model Training and Prediction 5

4 Experimental Results **7**

4.1 Results 7

4.2 Summary 8

5 Conclusion **9**

Bibliography **10**

CHAPTER 1

INTRODUCTION

1.1 Background

The expansion of fake job advertisements has become a major problem in recent years. These fraudulent job postings can lead to job seekers being scammed or exploited, resulting in loss of money, time, and sometimes even their personal information. Traditional methods of identifying fake job postings such as manual screening or keyword matching have proven to be inadequate due to the increasing sophistication of scammers. Therefore, there is a need for an automated solution that can accurately distinguish between real and fake job descriptions. In the context of job postings, machine learning models can be trained using features such as job titles, job descriptions, company information, and location. These features can be used to train a model to identify characteristics that are unique to fraudulent job postings, such as exaggerated salary, unrealistic job requirements, or requests for personal information.

In this proposal, a classification model based on artificial intelligence will be developed as a means to identify real job descriptions from fakes. The model will be trained and tested using a done using a 17K job description dataset. The proposed project aims to identify key traits and features that distinguish real job descriptions from fraudulent ones.

1.2 Motivation

The proposed project aims to address the growing problem of fake job postings in the recruitment industry. The following are some of the key motivations for this project:

- Develop an AI-based classification model to identify fake job postings in the recruitment industry.
- Protect job seekers from being scammed or exploited by fake job postings.
- Save time and resources for both job seekers and recruiters.
- Improve the recruitment industry's reputation and increase trust among job seekers.
- Provide valuable insights into the characteristics of fraudulent job postings through exploratory data analysis.
- Advance machine learning techniques and natural language processing.
- Inform future recruitment strategies and prevent fraudulent activity in the industry.

CHAPTER 2

LITERATURE REVIEW

Job description fraud detection has been the subject of numerous research. In one such study, researchers created a model that can recognize fake job postings based on the text contained in them using machine learning techniques (Bianchi Amft, 2018).[1] They discovered that certain terms, like "work from home" and "no experience required," are more likely to appear in fake job listings than in legitimate ones. This study shows the essential qualities and characteristics that set legitimate job descriptions apart from false ones. Another study examined job descriptions and business information to identify fake job listings using natural language processing algorithms (Chen Zhang, 2017).[2] They discovered that genuine job postings give more particular information on the position's criteria and the history of the company, whereas counterfeit job postings frequently employ vague and generic wording. This study emphasizes how crucial it is to evaluate both textual and metadata when creating classification models for identifying job description fraud. We will use artificial intelligence or the above research to resolve this issue.

CHAPTER 3

METHODOLOGY

3.1 Proposed Methodology

In this project, we aimed to predict whether a job posting is fraudulent using machine learning techniques. The dataset used in this project is the "Real or Fake" job postings dataset, which was obtained from Kaggle. The dataset contained several features, including job title, location, company profile, job description, and required qualifications. In this process, we will use Naïve Bayes and a Random forest classifier with Natural language to classify if the job is fake or real.

3.1.1 Data collection

The dataset contains textual information and meta-information about the jobs. It consists of approximately 800 fake job descriptions and 17,200 real job descriptions. The dataset can be used to create classification models using text data features and meta-features. This dataset is collected from kaggle.com

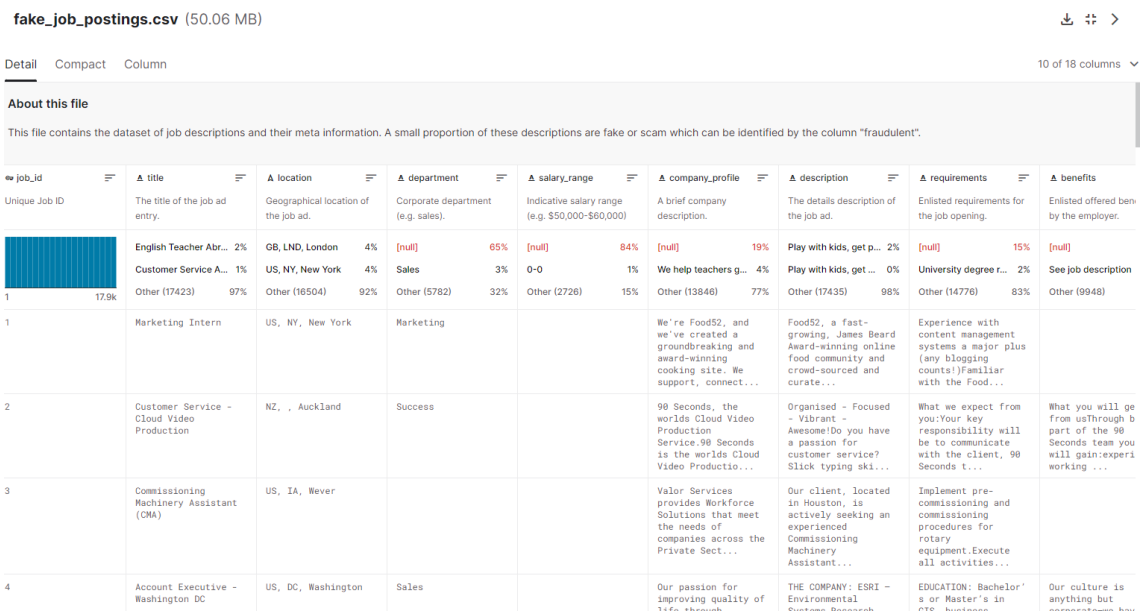

Figure 3.1. Dataset of real and fake job postings

3.1.2 Data Pre-processing

In this project, the dataset was prepared for analysis by performing several pre-processing steps. The isnull().sum() method was used to check for the presence of missing values in the dataset. The missing values were replaced with an empty string using the replace() method. We also use combining data from different columns into a single column and deleted columns that were deemed unnecessary for the analysis. The dataset was split into two subsets based on the fraudulent column. The "fraudjob" subset contained only fraudulent job postings, while the "actualjob" subset contained only actual job postings. These pre-processing steps were performed to clean and simplify the dataset and make it easier to analyze the data. Next, we split the dataset into training and testing sets using the train_test_split function from scikit-learn.[3]

3.1.3 Model Training and Prediction

In this process, we defined a pipeline that includes a CountVectorizer for feature extraction. The CountVectorizer class converts the text data into a numerical representation that can be used for machine learning. The tokenizer argument is set to split the text into individual words, the ngram_range argument specifies that we want to consider unigrams, bigrams, and trigrams, and the lowercase argument is set to True to convert all text to lowercase before tokenizing. In the pipe.predict argument applies the trained pipeline to the test data (X_test) by using the predict method to generate predictions based on the fitted model. The predictions are saved to the predicted variable. Then we calculate and print the accuracy

and recall scores of the model, as well as display the actual and predicted outcomes of the test data in a pandas DataFrame.

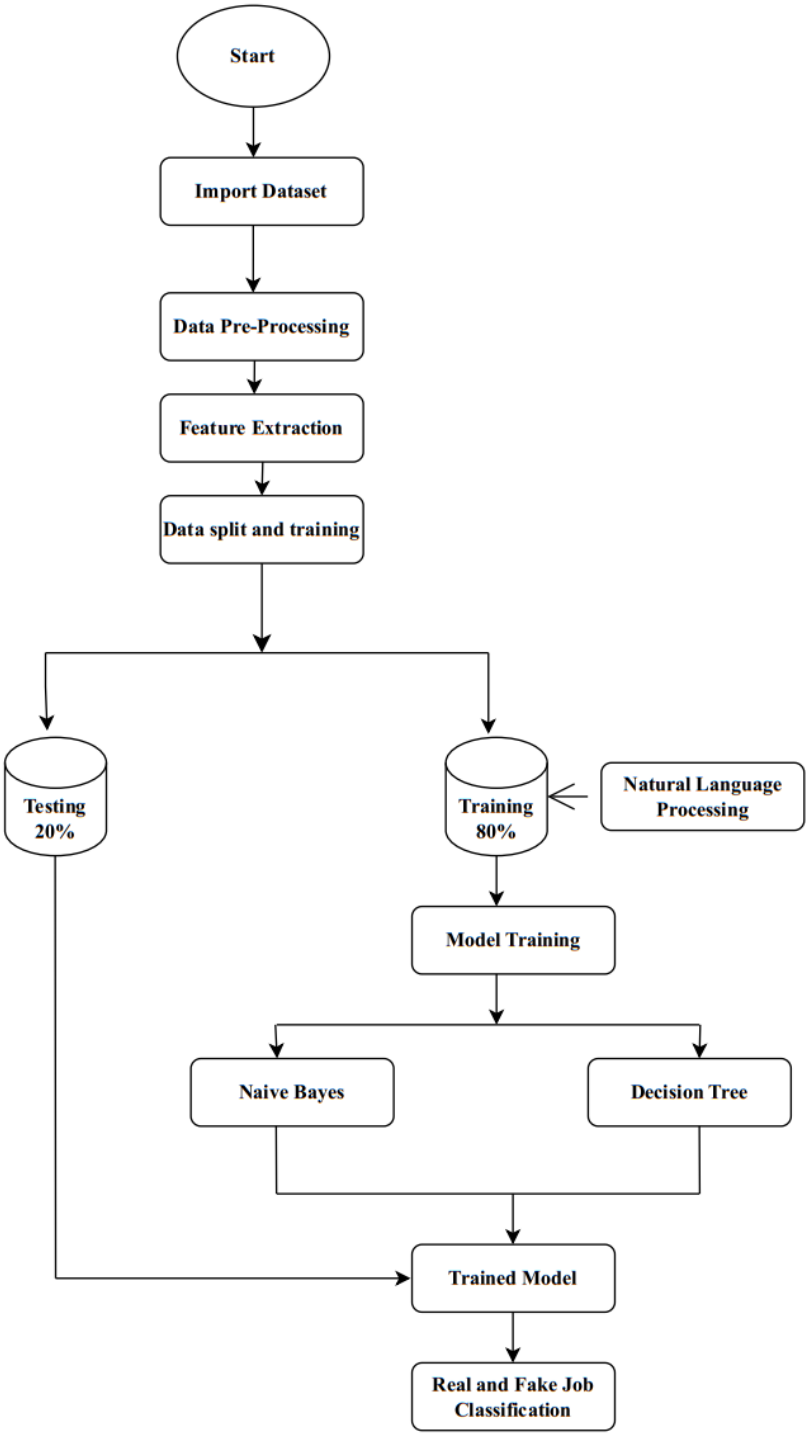

Figure 3.2. Proposed methodology

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Results

The two models used for the classification of real and fake job descriptions were Random Forest and Naive Bayes. The performance metrics used to evaluate the models were accuracy and recall.

Random Forest achieved an accuracy of 0.9837, which means that 98.37% of the job descriptions were classified correctly. It also achieved a recall of 0.6519, which means that out of all the fake job descriptions, 65.19% were correctly classified as fake.

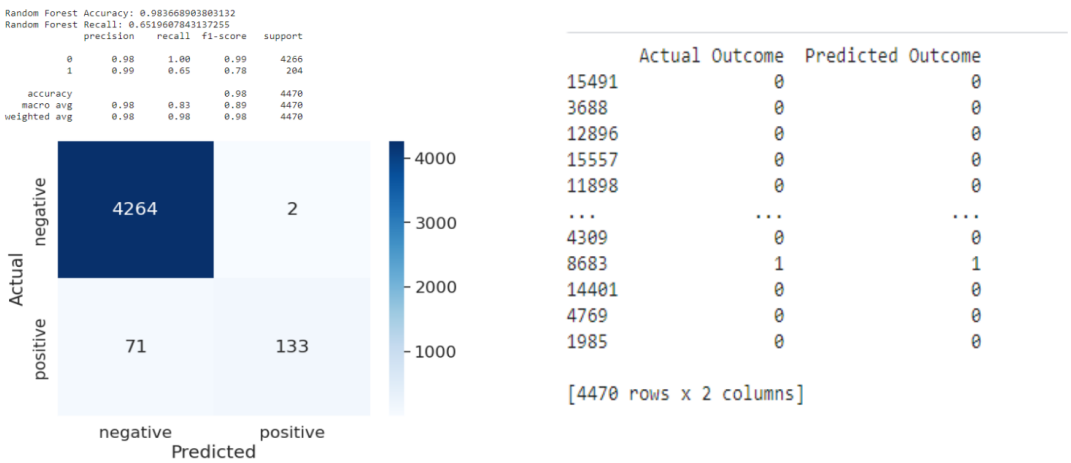

Figure 4.1. Random Forest result prediction

On the other hand, Naive Bayes achieved an accuracy of 0.9792, which means that 97.92%

of the job descriptions were classified correctly. It achieved a recall of 0.5539, which means that out of all the fake job descriptions, 55.39% were correctly classified as fake.

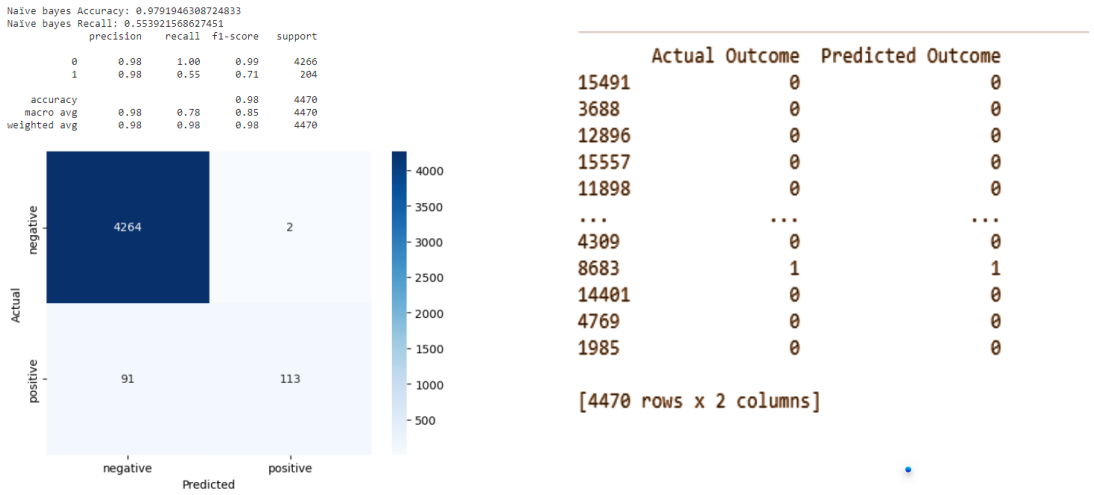

Figure 4.2. Naïve Bayes result prediction

4.2 Summary

In summary, we can say Random Forest achieved near-perfect accuracy, with only 30 out of the 4,470 job descriptions misclassified. This is a very low error rate and indicates that the model is highly reliable in identifying fake job descriptions. The recall metric is also important in this context, as it measures the model's ability to correctly identify all positive cases (i.e., fake job descriptions). While the recall score of 65.19% for Random Forest is not as high as the accuracy score, it is still a reasonable level of performance, considering that identifying fake job descriptions can be a challenging task.

Both Random Forest and Naive Bayes performed well in classifying real and fake job descriptions with Random Forest performing slightly better than Naive Bayes. Random Forest achieved higher accuracy and recall than Naive Bayes, which means it was able to correctly classify more fake job descriptions. Overall, the choice of model depends on the specific needs and constraints of the application. If the goal is to achieve high accuracy and recall in identifying fake job descriptions, Random Forest appears to be a suitable choice. If speed and simplicity are more important, Naive Bayes may be a better option.

CHAPTER 5

CONCLUSION

In conclusion, we can say that the development of a classification model that predicts whether job descriptions are real or fake using artificial intelligence can have significant benefits for organizations and job seekers alike. With the rise of online recruitment and job platforms, the risk of fraudulent job postings has also increased, making it important to develop effective methods for identifying such postings. The proposed project aims to address this challenge by utilizing machine learning algorithms such as Random Forest and Naive Bayes with Natural Language Processing to analyze the text of job descriptions and identify key features that distinguish real job postings from fraudulent ones. The project is expected to achieve high accuracy in predicting fraudulent job postings, which can help organizations avoid costly and damaging outcomes associated with hiring fraudulent candidates. All around, the successful implementation of this project can contribute to the development of a safer and more secure job market by providing valuable insights into fraudulent job postings and enabling organizations to take aggressive measures to prevent them.

BIBLIOGRAPHY

[1] C. Bianchi Amft, "Fake job postings detection using machine learning techniques," *Procedia Computer Science*, vol. 138, pp. 303–310, 2018.

[2] W. Chen Zhang, S. Wu, and J. Zhang, "Identifying fake job advertisements through analysis of online job descriptions and business information," *Journal of Organizational Computing and Electronic Commerce*, vol. 27, no. 2, pp. 156–172, 2017.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.