Machine Learning Techniques for Detecting Fake Job Listings

KOTHURU HARSHAVARDHAN1, PANDITA AJAYA KUMAR2

#1Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

#2Assistant Professor, Department of CSE, PBR Visvodaya Institute of Technology and Science, Kavali

ABSTRACT_ The project presents an application that uses categorization techniques based on machine learning to stop fake job advertising online. The results of various classifiers are compared in order to choose the most effective model for detecting job scams. These classifiers are employed to validate fake online postings. It helps in spotting fake job listings amid many other postings.

Single classifiers and ensemble classifiers are the two basic types of classifiers taken into account for the aim of detecting bogus job advertisements. However, experimental results demonstrate that ensemble classifiers outperform single classifiers at identifying fraud.

1.INTRODUCTION

One of the serious issues recently addressed in the area of online recruitment frauds (ORF) is employment scam. Nowadays, many businesses prefer to post their open positions online so that job seekers can access them quickly and easily. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. A reputable company may face criticism for false job postings if they compromise their credibility. These fraudulent job post detections attract a lot of interest in developing an automated tool for identifying fake jobs and alerting people to them so they won't apply for them. In order to accomplish this, a machine learning approach is used, which makes use of a number of classification algorithms.

A classification tool separates the fake job postings from a larger collection of job advertisements in this instance and notifies the user. Supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of identifying scams on job postings. A classifier uses training data to map input variables to target classes. In the paper, classifiers for distinguishing fake job postings from the others are discussed. These classifier-based predictions can be broadly divided into two categories: single

classifier predictions and ensemble class predictions.

One of the serious issues recently addressed in the area of online recruitment frauds (ORF) is employment scam. Nowadays, many businesses prefer to post their open positions online so that job seekers can access them quickly and easily. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. A reputable company may face criticism for false job postings if they compromise their credibility.

These fraudulent job post detections attract a lot of interest in developing an automated tool for identifying fake jobs and alerting people to them so they won't apply for them. In order to accomplish this, a machine learning approach is used, which makes use of a number of classification algorithms. A classification tool separates the fake job postings from a larger collection of job advertisements in this instance and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of identifying scams on job postings. A classifier uses training data to map input variables to target classes.

In the paper, classifiers for distinguishing fake job postings from the others are discussed. These classifier-based predictions can be broadly divided into two categories: single classifier predictions and ensemble class predictions. Ensemble class predictions, which use multiple algorithms to increase accuracy, are more accurate than single classifier predictions, which are based on a single algorithm and less accurate. The study discovered that when it came to spotting fake job postings, ensemble class predictions performed better than single classifier predictions. This emphasises how crucial it is to use ensemble techniques in machine learning to increase accuracy. Ensemble methods, which combine the advantages of multiple algorithms, can make predictions that are more accurate in a variety of applications. It has been demonstrated that ensemble methods are particularly good at spotting fraud, including phishing scams and credit card fraud.

Ensemble methods can enhance the overall effectiveness of fraud detection systems by combining the advantages of various algorithms. This is accomplished by ensemble methods, which combine the predictions of various algorithms to produce a prediction that is more trustworthy and accurate. Accurate predictions are essential for decision-making in sectors like finance and

healthcare, so this method is gaining popularity.

Today, people are eagerly waiting for jobs for their needs. Many scammers have posted fake posts with fraudulent job postings, using their situation as an advantage. It is important for job seekers to be cautious and do their research before applying to any job. They should verify the legitimacy of the company and the job posting before sharing any personal information or paying any fees. People find it a little challenging to tell the difference between authentic and fake work. It is important to look for red flags such as unrealistic job descriptions, promises of high pay for little work, and requests for sensitive information upfront. Seeking advice from trusted sources can also help in identifying potential scams. We built a model to predict fake job postings.

2.LITERATURE SURVEY

2.1 Amaar, A., Aljedaani, W., Rustam, F. et al. Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. Neural Process Lett 54, 2219–2247 (2022).

Amaar. A et al., proposed an approach for identifying false job postings from job portals using supervised machine learning and natural language processing. Both the Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words approaches were used to extract the features from the data (BoW). The findings of this study show that machine learning and natural language processing have the potential to increase the precision and effectiveness of job matching. Both job seekers and employers can benefit from using this technology to find more relevant job postings and better-suited candidates for open positions. In the end, a more effective job market will result from a reduction in the time and resources used in the recruitment and hiring processes. To make sure that the algorithms used do not continue to support discrimination or exclusion, it is crucial to take into account any potential biases in them. Companies should also give diversity and inclusion top priority in their hiring procedures to make sure that they are both effective and fair. Blind hiring practices and actively seeking out candidates from underrepresented groups can help with this.

2.2 Keerthana, B., Reddy, A.R., Tiwari, A. (2021). Accurate Prediction of Fake Job Offers Using Machine Learning. In: Bhattacharyya, D., Thirupathi Rao, N. (eds) Machine Intelligence and Soft Computing. Advances in Intelligent

Systems and Computing, vol 1280. Springer, Singapore.

A model for the prediction of fraudulent job advertisements was put forth by Keerthana. Btal. The authors used feature engineering techniques like one-hot encode, TFIDF Vectorizer, and count Vectorizer to improve the model's efficacy. They used classification techniques to forecast fraudulent job postings. Out of the total number of methods, the neural network, MLP Classifier, with the "Adam" model, produced an accurate result (71%). Overall, the approach taken by the authors to predict fraudulent job postings—combining feature engineering and classification techniques—proved to be effective. However, more investigation is required to enhance the model's accuracy and possibly take into account additional elements that might support phoney job postings. Using natural language processing methods to more accurately analyse the text of job postings may be one area for improvement. In order to enhance the model's capacity to identify fraudulent postings across various markets and industries, a larger and more varied dataset would be advantageous. Furthermore, incorporating machine learning algorithms that can identify patterns and anomalies in job posting data could also enhance the accuracy of detecting phony job postings. This approach would require a significant amount of data and computing power, but could ultimately lead to more effective fraud detection in the job market.

2.3 S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 543-546.

S. U. Habiba et al. suggested a number of data mining and classification algorithms, including KNN, SVM, Decision Tree, Naive Bayes classifier, Multilayer Perceptron, Random Forest classifier, and Deep Neural Network, to determine whether a job posting is real or fake.The article made use of 18000 samples from the Employment Scam Aegean Dataset (EMSCAD). A DNN classifier with three layers was employed for this classification problem.The trained classifier using DNN can predict a fake job advertisement with a classification accuracy of about 80%. The DNN classifier's high accuracy demonstrates how effective it is at spotting employment fraud.This might be helpful in creating automated systems that protect people from falling for such scams.

Additionally, the high accuracy of the DNN classifier can also save companies time and resources by quickly identifying fraudulent job postings and preventing them from being published. This can ultimately lead to a more efficient and trustworthy job market for both employers and job seekers

3.PROPOSED SYSTEM

We intend to employ a Kaggle dataset in this system that comprises details about the job, such as job id, title, location, and department. Data preprocessing involves removing things like pointless spaces, null entries, stop words, and other things from data. The data is preprocessed and cleaned to make it prediction ready before being given to the classifier for predictions.

To forecast the most accurate values, we would use the XGBoost algorithm.

3.1 IMPLEMENTATION

3.1.1 Modules and libraries

A library is a group of related functions that can be incorporated into your Python code and used as needed in the same way as other functions. Rewriting code to carry out a common task is not necessary. With libraries, you can effectively increase the functionality of your code by importing pre-existing functions. Using libraries not only reduces time and work requirements, but also guarantees that the code is dependable and has been thoroughly tested by a large developer community, reducing the likelihood of errors. Additionally, rather than wasting time on routine tasks, it enables developers to concentrate on the distinctive aspects of their project.

3.1.2 Data preprocessing

Pre-processing describes the changes made to our data before we feed it to the algorithm. A technique for turning unclean data into clean data sets is data preprocessing. In other words, data is always gathered from various sources in a raw state that precludes analysis. Data preprocessing entails a number of steps, such as data cleaning, data transformation, and data reduction, to ensure that the data is accurate, consistent, and suitable for analysis. It improves the data's quality and increases the accuracy of the inferences made from it, making it a crucial step in the analysis of data. Finding and fixing data errors, inconsistencies, and inaccuracies is known as data cleaning. While data reduction entails reducing the volume of data to a manageable size without losing crucial information, data transformation entails converting the data into a format that is suitable for analysis.

3.1.3 Data Visualisation

Data visualisation is the process of transforming sizable data sets into statistical and graphical representations. In data science and knowledge discovery techniques, it is essential to make data more understandable and accessible. Visual representation is required for charts and graphs in order to facilitate quick information absorption and make them easier to understand. Avoid hesitating on tables with large data sets if you want to keep the audience's interest for a longer period of time. Additionally, by using visualisation, it is possible to identify patterns and trends that may not be readily apparent when examining raw data, which can aid in problem-solving and better-informed decision-making. Additionally, it can help in simplifying complex information so that a larger audience can understand it. Since it offers a clear and concise way to communicate information, visualisation can improve team members' ability to collaborate and communicate. Additionally, it can make it easier to spot outliers or anomalies in the data, which is important for spotting mistakes or potential problems. Furthermore, data visualisation can aid in spotting patterns and trends in the data, which can improve decision-making. It can also be an effective tool for explaining complicated information in a way that is simple to comprehend and remember.

3.1 4 Model Building

Here are the key steps involved in the model building process:

Define the problem: Clearly understand the problem you are trying to solve and determine the type of machine learning task, such as classification, regression, clustering, or recommendation.Collect and preprocess data: Gather relevant data that is representative of the problem domain. Clean the data by handling missing values, removing outliers, and normalizing or scaling the features as necessary. Split the data into training and testing sets.Select a model: Choose an appropriate machine learning algorithm or model based on the problem type, available data, and desired outcome. Some common algorithms include linear regression, decision trees, support vector machines, random forests, and neural networks.

3.2 ALGORITHM USED

3.2.1 XGBoost Algorithm

XGBoost is a distributed gradient boosting library that has been optimised for quick and scalable machine learning model training. A number of weak models' predictions are combined using this ensemble learning technique to create a stronger prediction. Extreme Gradient

Boosting, or XGBoost, is one of the most well-known and widely used machine learning algorithms because it can handle large datasets and perform at the cutting edge in many machine learning tasks like classification and regression. Its effective handling of missing values, which enables it to handle real-world data with missing values withou requiring a lot of pre-processing, is one of the key characteristics of XGBoost. Additionally, XGBoost includes integrated support for parallel processing, enabling model training on large datasets in a reasonable amount of time.

4.RESULTS AND DISCUSSION

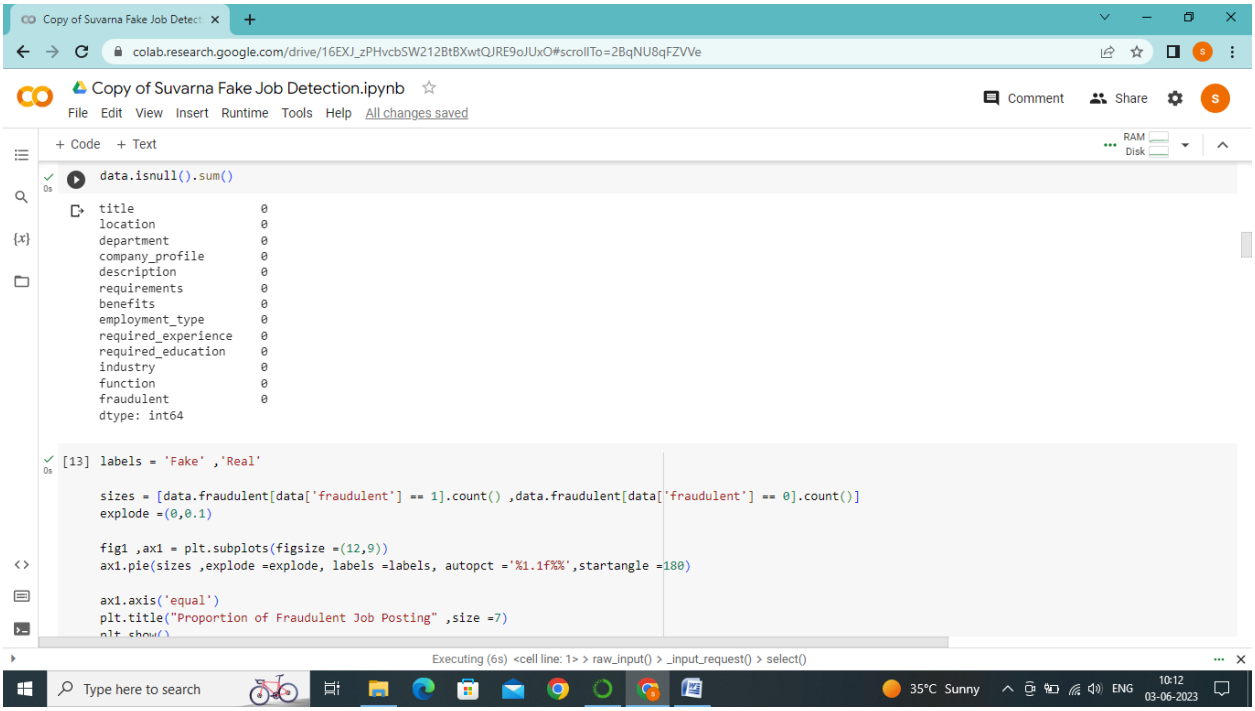

Fig 1:

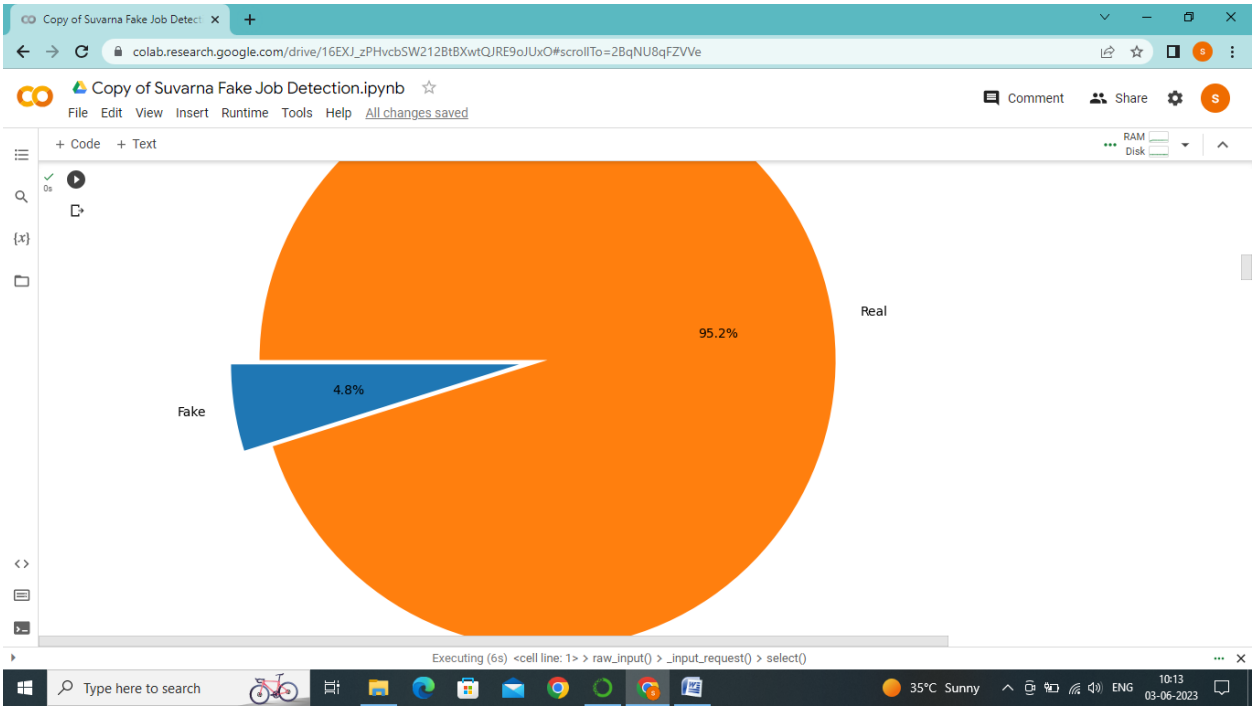

Fig 2:

The main aim of the project is to find the job posting is fake or legit. To predict it, we created a machine learning model. It aids in decision-making and prevents loss for the people. For the purpose of determining the veracity of a job posting, the machine learning model examines a number of variables, including the job description, company information, and salary range. This initiative will go a long way toward reducing employment fraud and safeguarding job seekers. Additionally, it will improve the overall job-seeking experience by giving job seekers accurate and trustworthy information so they can choose their careers with confidence. In the end, a more effective and open job market will result from this. Job seekers will be better able to make informed decisions about their careers, which will result in better job matches and greater job satisfaction, by reducing employment fraud and providing accurate information. Employers will gain from this because it will lower turnover rates and raise the caliber of their workforce.

4.1 Result

Our project has achieved a good accuracy percentage of prediction. This level of accuracy can be attributed to the use of advanced machine learning algorithms and extensive data analysis.By including extra features and optimising the model parameters, more advancements can be made. The XGboost algorithm was used to build the model.By including extra features and optimising the model

parameters, more advancements can be made. The XGboost algorithm was used to build the model. A thorough data analysis was then performed on the model to assess its performance and pinpoint any areas that needed improvement. The outcomes demonstrated the model's high accuracy rate and its potential for further optimization to produce even better outcomes.

4.2 CONFUSION MATRIX

An aid to visualising a classification problem's results is a confusion matrix, which presents a table layout of the various outcomes of the prediction and results. It plots a table of all the predicted and actual values of a classifier.The matrix's cells each represent a unique combination of expected and observed values, making it simple to distinguish between true positives, true negatives, false positives, and false negatives. This data is essential for assessing a classification model's effectiveness. The matrix is commonly referred to as a confusion matrix, and it provides a comprehensive overview of the model's performance. It is often used to calculate various performance metrics, such as accuracy, precision, recall, and F1 core.

Train Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	12772
1	1.00	1.00	1.00	638
accuracy			1.00	13410
macro avg	1.00	1.00	1.00	13410
weighted avg	1.00	1.00	1.00	13410

Test Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	4242
1	1.00	1.00	1.00	228
accuracy			1.00	4470
macro avg	1.00	1.00	1.00	4470
weighted avg	1.00	1.00	1.00	4470

5.CONCLUSION

Future research should concentrate on attempting to use a sophisticated or recently acquired dataset in order to compare the efficacy of machine learning and deep learning methodologies. Furthermore, given that the dataset being used is unbalanced, it is suggested to use a different dataset. To create a much more universal method of feature selection and focus, additional evaluation should be conducted to assess feature relevance across various datasets that may or may not have similar characteristics. Since all features were used in this study, we will use feature selection to determine the difference between the total and chosen features in the future

REFERENCES

- Amaar, A., Aljedaani, W., Rustam, F. et al. Detection of Fake Job Postings by

Utilising Machine Learning and Natural Language Processing Approaches. Neural Process Lett 54, 2219–2247 (2022).

- Keerthana, B., Reddy, A.R., Tiwari, A. (2021). Accurate Prediction of Fake Job Offers Using Machine Learning. In: Bhattacharyya, D., Thirupathi Rao, N. (eds) Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, vol 1280. Springer, Singapore.
- S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 543-546.
- Chiraratanasopha, B., & Chay-intr, T. (2022). Detecting Fraud Job Recruitment Using Features Reflecting from Real-world Knowledge of Fraud. CURRENT APPLIED SCIENCE AND TECHNOLOGY, 12-pages.
- Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.

AUTHOR PROFILES

Ms.M.ANITHA completed her Master of Computer Applications and Masters of Technology. Currently working as an Assistant professor in the Department of Masters of Computer Applications in the SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Mr. Y. NAGAMALLESWARA RAO Completed his Bachelor of Computer Applications at Acharya Nagarjuna University. He completed his Master of Science at Acharya Nagarjuna University.He completed MTECH from Jawaharlal Nehru Technological University Kakinada. Currently working as an Assistant professor in the Department of Computer Applications SRK Institute of Technology, Enikepadu, Vijayawada, NTR(DT). His areas of interest include Networks, Machine Learning & Artificial Intelligence.

Ms.T.SUVARNA is an MCA student in the Department of Computer Applications at SRK Institute of Technology, Enikepadu, Vijayawada, NTR District. She has a Completed Degree in B.Sc.(computers) from Svl kranthi Degree college Avanigadda. Her areas of interest are DBMS, Java Script, and Machine Learning with Python.