10.48047/jocaa.2024.33.06.34

Fake Job Prediction Using Machine Learning

Mr. R. Sreedhar[1], C.Sai Nithya[2], Bodukuriwar Shivani[3], Jangam Pushpa Latha[4]
[1] Associate Professor, Department of Information Technology, Sridevi Women's Engineering College, Hyderabad
Email: rachasreedharswec@gmail.com
[2, 3, 4] Department of Information Technology, Sridevi Women's Engineering College, Hyderabad

Abstract:-Fake job listing detection is an interesting topic for computer scientists and social science. The recent growth of the online social fake job postings has great impact to the society. There is huge information from disparate sources among various users around the world. Developing a technique that can detect fake job postings from these platforms is becoming a necessary and challenging task. This project proposes a machine learning method which can identify which job posting is fraudulent and which job posting is non-fraudulent. To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed in the paper. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers.

Keywords:- NLP, Text Mining, Sentiment Analysis, Feature Extraction, Location-Based Features, Machine Learning Models, Fraud Indicators, Behavioral Analysis, Validation Techniques, Data Preprocessing, Geospatial Analysis, Web Scraping, User Feedback Analysis, Feature Engineering, Model Interpretability, Continuous Monitoring, Collaborative Filtering, Anomaly Detection, Supervised Learning, Unsupervised Learning, Classification Algorithms

I INTRODUCTION

We are living in unprecedented times due to COVID-19 pandemic hurting economies in every continent. Unemployment rates are increasing every single day with the United States reporting around 26 million people applied for unemployment benefits, which is the highest recorded in its long history, millions have been furloughed in the United Kingdom, and thousands have been laid off around the world. In these desperate times, when thousands and millions of people are on the lookout for a job, it provides a perfect opportunity for online scammers to take advantage of their desperation. We see a daily rise in these fake job postings where the posting seems pretty reasonable, often these companies will have a website as well, and they will have a recruitment process that is similar to other companies in the industry. If one looks hard enough, they can spot the differences between these fake postings and genuine ones. Most of the time these postings don't have a company logo on these postings, the initial response from the company is from an unofficial email account, or during an interview they might ask you for personal confidential information such as your credit card details by saying they need it for personnel verification. In Normal economic conditions, all these are evident hints that there something suspicious about the company, but these are not normal economic conditions. These are the worst times we all have seen in our lifetimes, and at this time,

desperate individuals just need a job, and by this, these individuals are directly playing into the hands of these scammers.

II PURPOSE

Fake job listing detection is an interesting topic for computer scientists and social science. The recent growth of the online social fake job postings has great impact to the society. There is a huge information from disparate sources among various users around the world. Developing a technique that can detect fake job postings from these platforms is becoming a necessary and challenging task. This project proposes a machine learning method which can identify which job posting is fraudulent and which job posting is non-fraudulent.

III LITERATURE SURVEY

Mykhailo Granik et. al. in their paper shows a simple approach for fake job listing detection using naive Bayes classifier. This approach was implemented as a software system and tested against a data set from Facebook. They achieved classification accuracy of approximately 74%. Classification accuracy for fake job posts is slightly worse. This may be caused by the skewness of the dataset only 4.9% of it is fake job posts.

Himank Gupta et. al. gave a framework based on different machine learning approach that deals with various problems including accuracy shortage, time lag (BotMaker) and high processing time to handle thousands of posts in 1 sec. Firstly, they have collected 400,000 posts from HSpam14 dataset. Then they further characterise the 150,000 fake posts and 250,000 non- spam posts. They also derived some lightweight features along with the Top 30 words that are providing highest information gain from Bag-of-Words model.

They were able to achieve an accuracy of 91.65% and surpassed the existing solution by approximately18%. It is a theoretical Approach which gives Illustrations of fake job listing detection by analysing the words used.

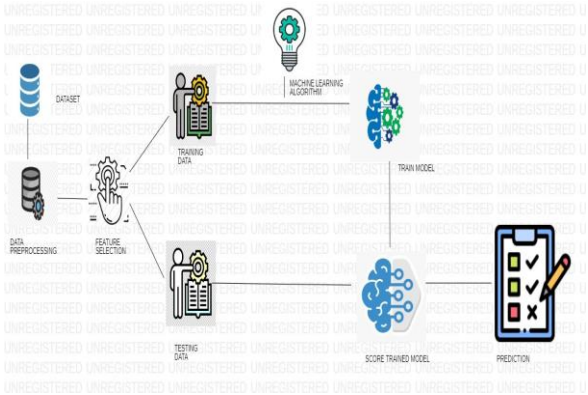
IV EXISTING SYSTEM:

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. 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. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs.

V PROPOSED SYSTEM:

The target of this study is to detect whether a job post is fraudulent or not. Identifying and eliminating these fake job advertisements will help the job seekers to concentrate on legitimate job posts only. In this context, a dataset from Kaggle [13] is employed that provides information regarding a job that may or may not be suspicious. The dataset has the schema as shown in Fig. 1. This dataset is used in the proposed methods for testing the overall performance of the approach. For better understanding of the target as a baseline, a multistep procedure is followed for obtaining a balanced dataset. Before fitting this data to any classifier, some pre-processing techniques are applied to this dataset. Pre-processing techniques include missing values removal, stop-words elimination, irrelevant attribute elimination and extra space removal. This prepares the dataset to be transformed into categorical encoding in order to obtain a feature vector. This feature vectors are fitted to several classifiers.

a) System Architecture



Proposed Architecture

VI MODULES USED

i) NumPy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using NumPy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

ii) Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyse. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

iii) Matplotlib

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and I Python shells, the Jupyter Notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample plots and thumbnail gallery.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object-oriented interface or via a set of functions familiar to MATLAB users.

iv) Scikit – learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

vi) Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Steps Involved in Data Preprocessing:
Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

- **Ignore the tuples:**

  This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

- **Fill the Missing values:**

  There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

(b). Noisy Data:

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

Regression:

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Clustering:

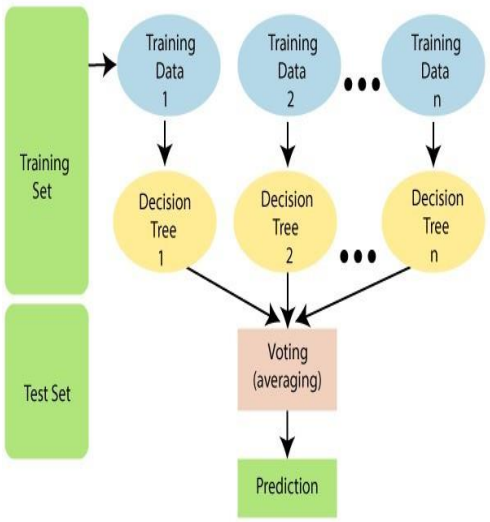
This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

VII ALGORITHMS USED

a) RANDOM FOREST ALGORITHM

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. The below diagram explains the working of the Random Forest algorithm:



b) NAÏVE BAYES ALGORITHM

Naïve Bayes algorithms are a classification technique based on applying Bayes' theorem with a strong

assumption that all the predictors are independent to each other.

In simple words, the assumption is that the presence of a feature in a class is independent to the presence of any other feature in the same class.

For example, a phone may be considered as smart if it is having touch screen, internet facility, good camera etc.

Though all these features are dependent on each other, they contribute independently to the probability of that the phone is a smart phone. In Bayesian classification, the main interest is to find the posterior probabilities i.e. the probability of a label given some observed features, $(L \mid features)$. With the help of Bayes theorem, we can express this in quantitative form as follows:

P(L|features)=P(L)P(features|L)P(features)

Here, $(L \mid features)$ is the posterior probability of class.

(L) is the prior probability of class.

$(features|L)$ is the likelihood which is the probability of predictor given class.

$(features)$ is the prior probability of predictor.

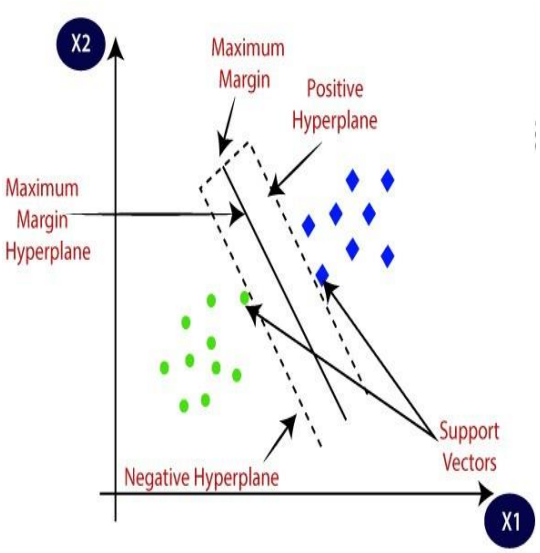
Building model using Naïve Bayes in Python, Python library, Scikit learn is the most useful library that helps us to build a Naïve Bayes model in Python.

We have the following three types of Naïve Bayes model under Scikit learn Python library:

c) SUPPORT VECTOR MACHINE ALGORITHM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



d) LOGISTIC REGRESSION:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

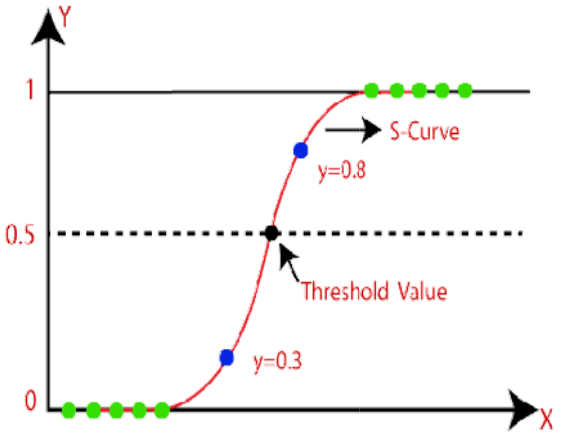
Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which

predicts two maximum values (0 or 1).The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significa t machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Logistic Regression can be us ɛ d to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below im ɛge is showing the logistic function:



e) K-NEAREST NEIGHBOR ALGORITHM:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

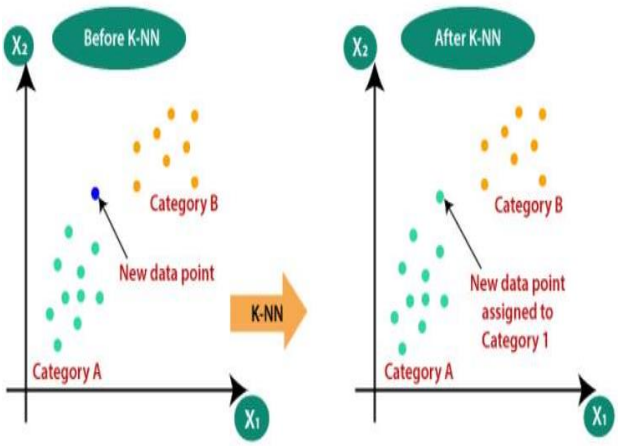
K-NN algorithm stores all the available data and classifies a new data point based ɑr the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



f) DECISION TREE

Decision Tree is a supervised learning technique that can be used for both cla ɛsification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches,

whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

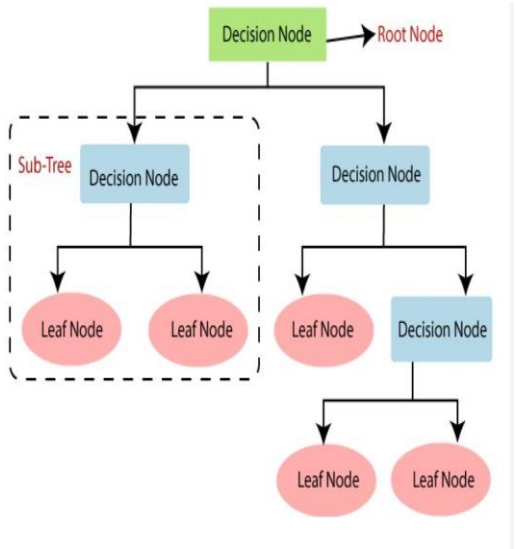
It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Below diagram explains the general structure of a decision tree:



- **Information gain**
  Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:
  Information Gain= Entropy(S)- [(Weighted Avg) *Entropy (each feature)
- **Entropy:** Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

  Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)
- **Gini index**

  Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

  Gini Index= $1- \sum_j P_j^2$

Advantage:

i. Machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user.

ii. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed.

Disadvantages:

i. In recent days, many companies prefer to post their vacancies online so that these can be

accessed easily and timely by the job-seekers. 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.

VIII CONCLUSION AND FUTURE SCOPE

Currently, we live in times which none of us ever expected. Coronavirus has not only brought health emergencies to nations but also accelerated an impending recession. Many employees are being laid off every day, and the demand for jobs is way higher than the number of posts available in the market. Turmoil and chaos are the perfect proponents for scammers, and currently, cyber scam attacks are on the rise. Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

For the future works, we intend to expand dataset and enrich the ruleset by focusing on user behaviour, company and network data as well as user-content-IP collision patterns. Moreover, we would like to employ graph modelling and explore connections between fraudulent job ads, companies, and users. Ultimately, our goal is to propose an applicable employment fake job listing detection tool for commercial purposes

IX REFERENCES

[1] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection,” J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2] I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier,‖ no. January 2001, pp. 41–46, 2014.

[3] D. E. Walters, ―Bayes’s Theorem and the Analysis of Binomial Random Variables,‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4] F. Murtagh, ―Multilayer perceptrons for classification and regression,‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90023-5.

[5] P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers,‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6