

Comparitive Study On Counterfeit Job Post Prediction

P.Reshma,S.Mahesh,R.Prasad Kumar,G.Harsha,Ms.A.Surekha

¹Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivasala, Visakhapatnam, Andhra Pradesh, India.

²Associate professor, Department of Information Technology, Anil Neerukonda Institute of Technology and Sciences, Sangivasala, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

With the development of social media and modern technologies, advertising new job openings has recently become a very prevalent problem in the current world. Therefore, everyone will have a lot of reason to be concerned about bogus job postings. Fake job posting prediction presents a variety of difficulties, just like many other classification tasks.

In order to determine whether a job posting is legitimate or fake, this article suggested using various data mining methods and classification algorithms like logistic regression, support vector machine, and random forest classifier. 18000 data from the Employment Scam Aegean Dataset (EMSCAD) were used in our experiments. The trained classifier shows approximately 98% classification accuracy (logistic regression) to predict a fraudulent job post.

KEYWORDS:

Fake Job, Online Recruitment, Machine Learning, Ensemble Approach, False job prediction, Data mining.

I. INTRODUCTION:

Modern-day job seekers now have a wealth of new and varied employment options because to advancements in industry and technology. Job seekers are able to learn about their options based on their availability, qualifications, experience, suitability, and other factors with the aid of the advertisements for these job offers.

The employment process has now been impacted by the force of the internet and social media. Because the effectiveness of a recruitment process depends on how it is portrayed, social media has a huge impact on this. Social media and electronic media marketing have made it simpler to share information about one's job. Instead, the ability to quickly spread job listings has increased the number of fraudulent job postings, which annoys job searchers. People are therefore less likely to show interest in new job listings because they want to keep their personal, scholastic, and professional information secure and consistent.

Thus, gaining the public's confidence and dependability is a very difficult task that genuine employment listings through social and electronic media confront. Although technologies are all around us to make our lives easier and better, they shouldn't be utilised to establish hazardous working circumstances. If job postings can be correctly filtered to anticipate false job postings, it will significantly improve the process of hiring new employees. False job postings make it challenging for job seekers to regularly locate employment. False job postings squander a lot of time by making it difficult for job searchers to locate their perfect jobs. A New Window Is Opened to Face Difficulties In The Field Of Human Resource Management Due To An Automated System To Predict False Job Posts.

A. Fake Job Posting:

"Job scams" are online job postings that are dishonest and typically prepared to collect applicants' personal and professional information rather than matching them with pertinent employment. Sometimes dishonest individuals try to take money from job seekers.

More than 67% of people who look for work online but are unaware of fake job listings or employment scams are at major risk, according to a recent report by ActionFraud in the UK. Around \$5,00,000 in losses were reported by over 7,00,000 job seekers in the UK due to recruitment scams. According to the survey, the UK experienced an almost 300% increase over the preceding two years.

Because they typically try to acquire a stable job for which they are willing to spend more money, students and new graduates are the major targets of fraudsters. Because con artists continuously change their techniques of work theft, strategies for preventing or combating cybercrime fall short.

B. Common types of Job Scam:

In order to obtain other people's private information, including insurance information, bank account information, income tax information, date of birth, and a national identification number, fraudsters fabricate fake job advertising. When con artists seek money while using justifications like administrative costs, information security testing fees, management costs, etc., they are committing advance fee fraud.

As part of a pre-employment screening, con artists would occasionally assume the identity of employers and ask about candidates' driving records, bank account details, and visa status. Money mulling schemes fail when they get students to deposit money into their accounts and then move it back. By employing the "cash in hand" technique, one can find work that pays cash wages without having to pay taxes. Scammers regularly create bogus business websites, bank websites, official-looking papers, etc. to entice job seekers. The majority of employment scammers try to catch victims via email rather than face-to-face interaction.

They commonly promote their services as headhunters or job agencies on social media platforms like LinkedIn. Usually, they strive to provide the most realistic portrayal of their company or webpages to potential candidates. Regardless of the work scam they use, their goal is to obtain information about the job candidate and use it for their own gain, whether it be money or otherwise.

Project Scope and Direction:

Using various datamining techniques, such as categorization algorithms, the primary objective of this project is to forecast the false job listings and identify the individuals who are most willing to take the personal and professional information of job searchers instead of providing them with appropriate employment.

Impact, Significance, and Contribution:

Fake job post's prediction impact is increasing daily. False job postings make it difficult for job seekers to discover the positions they favor, which is a huge waste of their time. An automatic system to forecast false job postings opens a new door to challenges in the area of human resource management and aids in the identification of false postings among a huge number of postings.

Goals and objectives:

Fake job post prediction has gained popularity in recent years. The purpose of predicting fake job listings is to locate those individuals using various data mining techniques, such as classification algorithms, who are most eager to steal the professional and confidential information of job searchers.

False job listings make it challenging for job searchers to find the jobs they want, which is a significant time waster. An automated system to predict fake job postings offers up new opportunities for HR management problems and helps to identify false postings among a large number of postings.

II. RELATED RESEARCH:

To determine whether an employment posting is genuine or fake, numerous studies have been conducted. A significant amount of study is being done to identify employment fraud online.

Videos et al. recognised fraudulent online job advertisers as work fraudsters. They discovered data about numerous legitimate and well-known businesses and organisations that created false employment advertisements or vacancy listings with ulterior motives. On the EMSCAD dataset, they conducted experiments using a variety of categorization methods, including the naive bayes classifier, random forest classifier, Zero R, and One R.

The Random Forest Classifier performed the best on the sample, with a classifying accuracy of 89.5%. They found that the sample's logistic regression performance was appalling. One R algorithm performed well after testing with a modified dataset. In their study, they made an attempt to pinpoint the problems with the ORF model (Online Recruitment Fraud) and to use various well-known models to solve those problems.

Alghamdi [2] et al. proposed a model to assess fraud risk in an online job system. They ran trials utilising a machine learning technique on the EMSCAD dataset. Feature selection, data pre-processing, and classifier-based scam identification were the three processes they took when working with this dataset. They eliminated noise and html tags from the data during the preprocessing stage to preserve the overall text structure. They applied the feature selection method to reduce the number of attributes effectively and successfully. A random forest ensemble classifier was used to extract the bogus job ads from the test data, and SupportVectorMachine was utilised to choose the features. The random forest classifier appeared to be a tree-structured classifier that worked as an ensemble classifier with the help of the majority voting technique.

Huynh [3] et al. have proposed a number of deep neural network models that are pre-trained on text datasets, including Text CNN, Bi-GRU-LSTM CNN, and BiGRU CNN. They tried to categorise the list of IT jobs. Using information from the IT employment market, they developed a Text CNN model with three layers: a convolution layer, a pooling layer, and a fully connected layer. To learn the data, this algorithm used convolution and pooling layers. After being compressed, the weights were moved to the stratum along with all of their connections. The softmax function was employed as the categorization strategy in this model. They also utilised an ensemble classifier (Bi-GRU CNN, Bi-GRULSTM CNN) with a majority voting mechanism to increase classification accuracy. They found that the classification accuracy of TextCNN was 66% and that of Bi-GRU-LSTM CNN was 70%. The ensemble classifier, which had an accuracy of 72.4%, performed the classification task the best.

In order to differentiate between real and false news (including articles, authors, and topics) using text processing, Zhang [4] et al. suggested an automatic fake detector model. They had used a unique collection of news or stories shared on Twitter by the PolitiFact website account. This dataset was used to create the recommended GDU diffusive unit model. When data arrived from several sources at once, our trained model performed effectively as an automated false detecting model.

Researchers experimented with a wide range of algorithms and feature selection strategies in order to produce great results in the classification of fake job postings. Useful methods included feature selection using a support vector machine, data pre-processing, and text processing using a deep learning model. We have put out the idea of predicting employment plans using deep neural networks. Instead of using text data, we just used the categorical features of the EMSCAD dataset while using the training technique. With less processing time, this method effectively lowers the number of trainable attributes. Using the identical properties of the EMSCAD dataset, we performed a comparison analysis using K Nearest Neighbor, Naive Bayes classifier, fuzzy KNN, decision tree, support vector machine, random forest classifier, and neural network.

III. METHODOLOGY:

Types of Algorithms:

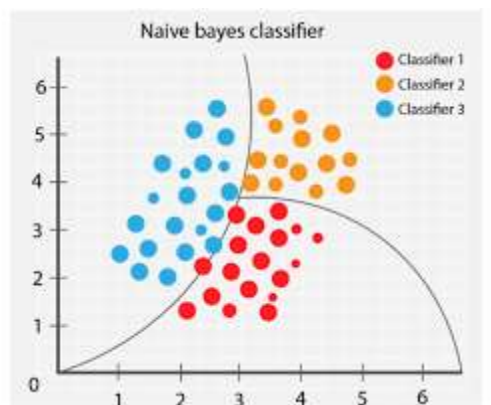
A. Single Classifier based Prediction:

Classifiers are trained for predicting the unknown test cases. The following classifiers are used while detecting fake job posts-

a) Naive Bayes Classifier:

The too simplistic concept that a class's presence (or lack) of one particular characteristic has no influence on that class's presence (or absence) of any other feature is the foundation of the naive bayes approach, a supervised learning technique. But despite this, it seems strong and effective. Its effectiveness is on par with that of other supervised learning methods. The literature has suggested a number of explanations. We emphasise a representation bias-based explanation in this tutorial. Linear classifiers include the naive bayes classifier, linear

discriminant analysis, logistic regression, and linear support vector machines (support vector machine). The approach used to estimate the classifier's parameters accounts for the disparity (the learning bias).

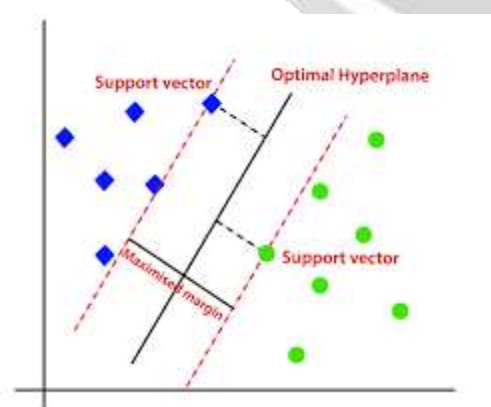


Although the Naive Bayes classifier is frequently used in research, it is not frequently employed by practitioners who wish to get useful results. On the one hand, the researchers discovered that it is particularly simple to programme and apply, that its parameters are simple to estimate, that learning happens quickly even on very big databases, and that its accuracy is reasonable compared to other systems. But, the final users do not receive a model that is simple to use and interpret, and they are unable to see the value of such a method.

b) SVM:

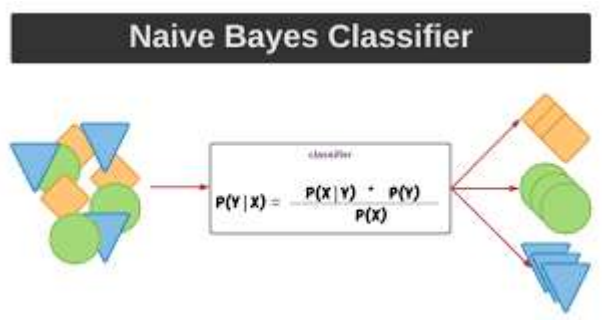
A discriminant machine learning technique for classification tasks tries to discover a discriminant function that can accurately predict labels for recently acquired examples using an independent and identically distributed training dataset.

A discriminant classification function takes a data point x and assigns it to one of the various classes that are a part of the classification task, in contrast to generative machine learning approaches that call for calculations of conditional probability distributions. Discriminant procedures are less efficient than generative approaches, which are typically utilised when outlier detection is part of the prediction process. This is particularly true for multidimensional feature spaces and when just posterior probabilities are required. Finding the equation for a multidimensional surface that best divides the various classes in the feature space is analogous to learning a classifier in terms of geometry.



c) Logistic regression Classifiers:

The association between a set of independent (explanatory) variables and a categorical dependent variable is investigated using logistic regression analysis. When the dependant variable simply has two values, such as 0 and 1 or Yes and No, the term logistic regression is employed. In situations where the dependant variable contains three or more distinct values, such as Married, Single, Divorced, or Widowed, the term multinomial logistic regression is typically reserved. Despite using a different set of data for the dependant variable than multiple regression, the method has a similar practical use.



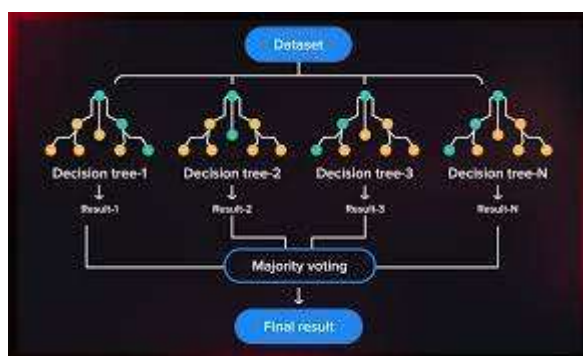
When it comes to evaluating categorical-response variables, discriminant analysis and logistic regression are competitors. Several statisticians believe that logistic regression, as opposed to discriminant analysis, is more adaptable and more suitable for modelling the majority of scenarios. This is so because, unlike discriminant analysis, logistic regression does not presuppose that the independent variables are regularly distributed.

B. Ensemble Approach based Classifiers:

The ensemble method enables several machine learning algorithms to work together to improve the overall system's precision. The concept of ensemble learning approach and regression technique are both used by random forest (RF) to solve classification-based problems. This classifier combines a number of classifiers that resemble trees and are used on different subsamples of the dataset. Each tree sends a vote for the class that best fits the input.

a) Random Forest:

The random forests or random decision forests ensemble learning strategy, which is used for classification, regression, and other tasks, builds a lot of decision trees during the training phase. The result of the random forest for classification issues is the class that the majority of the trees choose. The mean or average prediction of each individual tree is returned for regression tasks. Random choice forests, which have a tendency to overfit to their training set, are appropriate for decision trees. Although they frequently outperform decision trees, gradient boosted trees are more accurate than random forests. Yet, their effectiveness may be impacted by data peculiarities.



IV. PROPOSED APPROACH:

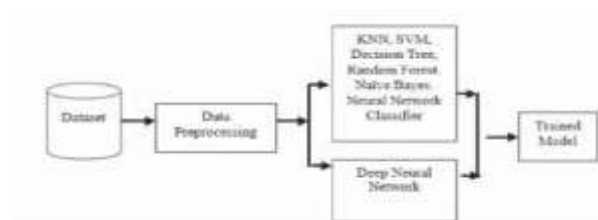


Fig. 1. Proposed Methodology

In order to determine whether a job posting is legitimate or fake, this article suggested using various data mining methods and classification algorithms, such as support vector machines, naive bayes classifiers, random forest classifiers, and support vector machines. This research aims to identify whether or not an employment posting is fraudulent. The removal of these false job postings will enable job searchers to focus solely on genuine job postings. A dataset from Kaggle is used in this situation to provide details about a task that may or may not be suspect.

The dataset's structure is as follows:

job_id	int64
title	object
location	object
department	object
salary_range	object
company_profile	object
description	object
requirements	object
benefits	object
telecommuting	int64
has_company_logo	int64
has_questions	int64
employment_type	object
required_experience	object
required_education	object
industry	object
function	object
fraudulent	int64

Fig. 1. Schema structure of the dataset

This collection includes 17,880 employment postings. The proposed methods test the overall effectiveness of the strategy using this dataset. The removal is obtained using a multistep process for better baseline understanding of the target. This prepares the information to be converted into categorical encoding in order to acquire a feature vector.

V. RESULTS:

Website Screenshots:

1.The below image is showing the user login page i.e., using this page user can login into the web application.



2. The below image is showing the user Register page i.e., using this page user can Register with the web application.



3. The service provider login screen, which allows service providers to access the web application, is displayed in the below-mentioned picture. The person who offers services to end consumers is known as a service provider.



4.Related Formulae:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

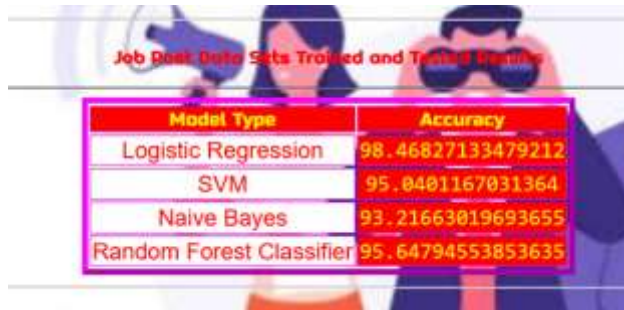
$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

(TP= True Positive, TN= True Negative, FP= False Positive,FN= False Negative).

5.Accuracy:



Job Post Data Sets Trained and Tested Results

Model Type	Accuracy
Logistic Regression	98.46827133479212
SVM	95.0401167031364
Naive Bayes	93.21663019693655
Random Forest Classifier	95.64794553853635

The trained classifier shows approximately 98.46% classification accuracy (logistic regression) to predict a fraudulent job post.

6.Barchat:



To determine whether a job posting is legitimate or fraudulent, we use a variety of machine learning methods and categorization algorithms, including logistic regression, support vector machines, and random forest classifiers.. This bar chat comparatively shows the accuracy of all algorithms and finally shows that Logistic Regression is having highest accuracy.

VI. CONCLUSION:

In conclusion, the discovery of employment scams has recently sparked grave concerns all around the world. We've examined the effects of work scams in this article since they can be quite lucrative and make it difficult to recognise fake job postings. The EMSCAD dataset, which contains fictional real-world job ads, was used in our research.

In this project, a number of machine learning techniques are proposed as defences against online fake job postings. Supervised Mechanism Is Used As An Example To Show How To Use Several Classifiers To Identify Employment Scams. According to experimental results, logistic regression works better than its peer classification tool when compared to all four algorithms. The accuracy of the proposed approach was 98.47%, which is much higher than the methods currently in use.

VII. REFERENCES:

[1] S. Vidros, C.Kolias , G. Kambourakis and L.Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

- [2] B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", *Journal of Information Security*, 2019, Vol 10, pp. 155–176, <https://doi.org/10.4236/iis.2019.103009>
- [3] "Hate Speech Detection on Vietnamese Social Media Text Using the Bi-GRU-LSTM-CNN Model" by T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen, arXiv prepr. arXiv1911.03644, 2019.
- [4] "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," *Neurocomputing*, vol. 174, pp. 806–814, 2016. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao.
- [5] "News Text Classification Based on Improved BiLSTM-CNN "by C. Li, G. Zhan, and Z. Li was published in the 2018 9th International Conference on Information Technology in Medicine and Education (ITME), pp. 890–893.
- [6] Tin Van Huynh¹, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen¹, and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.
- [7] "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network," by Jiawei Zhang, Bowen Dong, and Philip S. Yu, IEEE 36th International Conference on Data Engineering (ICDE), 2020
- [8] "Automatic Detection of Cyber Recruitment by Violent Extremists", *Security Informatics*, 3, 5, 2014, doi:10.1186/s13388-014-0005-5. Scanlon, J.R. and Gerber, M.S.
- [9] "Convolutional neural networks for sentence categorization" by Y. Kim was published in 2014 on arXiv Prepr as arXiv1408.5882.