INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 987-991

FAKE JOB DETECTION USING MACHINE LEARNING

Mr. Prakash Hongal[1], Mrs. Rajeshwari Gamanagatti[2], Ms. Savita V Yalavigi[3]

[1,2,3]Dept of CSE SKSVMACET Laxmeshwar Laxmeshwar, India.

DOI: https://www.doi.org/10.58257/IJPREMS33369

ABSTRACT

In recent years, the rise of online job portals and platforms has provided job seekers with a convenient and efficient means of finding employment opportunities. However, this digital transformation has also given rise to a growing concern: the proliferation of fake job postings. These deceptive listings can mislead and exploit job seekers, wasting their time and potentially exposing them to fraudulent activities. This work tries to address this issue by performing a comparison between Logistic Regression, Multi-layer Perceptron, Random Forest, and Decision Trees algorithms to determine which automated system can accurately distinguish between legitimate job advertisements and fake postings. To perform our experimentation process, the Employment Scam Aegean Dataset (EMSCAD) dataset was used to train and test our models. To improve further our results, feature engineering was applied to the data set to create new features from raw data. Our results demonstrated that the Multi-layer Perceptron and Logistic Regression can accurately classify fake job posts. These models were the two that had the best results according to the accuracy, precision, recall, and f1-score which were the metrics we used to evaluate each of them. This research provides significant value to job seekers, employers, and job portals alike. By accurately detecting and filtering out fake job postings, job seekers can avoid potential scams and focus their efforts on genuine employment opportunities. Employers benefit from improved reputation and more qualified applicants, while job portals can enhance their credibility and trustworthiness.

Keywords— Fake Job, Online Recruitment, Machine Learning, Ensemble Approach, Decision Tree.

1. INTRODUCTION

Now-a-days, getting a job is difficult. Before going to any interview you have to apply for a job, get registered then further go for an interview. The first and foremost step is to apply for a job according to the requirements of a company and as per the field a user wants to get a job in it. When you explore on internet you may find several job postings, those job postings may be a phony jobs or legitimate jobs. User may not find it easy as it is hard to say, the posted job is a fake or legitimate. So, we require a software to detect which is the fake job and which isn't, helping a number of people not to disclose their personal details to anyone by being aware of the fake job postings. For this purpose, 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. 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 in the paper for identifying fake job posts from the others are described briefly. These classifiers based prediction may be broadly categorized into -Single Classifier based Prediction and Ensemble Classifiers based Prediction.

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

Naive Bayes classifier is a supervised classification tool that exploits the concept of Bayes Theorem of Conditional Probability. The decision made by this classifier is quite effective in practice even if its probability estimates are inaccurate. This classifier obtains a very promising result in the following scenario- when the features are independent or features are completely functionally dependent. The accuracy of this classifier is not related to feature dependencies rather than it is the amount of information loss of the class due to the independence assumption is needed to predict the accuracy .

b) Multi-Layer Perceptron Classifier

Multi-layer perceptron can be used as supervised classification tool by incorporating optimized training parameters. For a given problem, the number of hidden layers in a multilayer perceptron and the number of nodes in each layer can differ. The decision of choosing the parameters depends on the training data and the network architecture.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 987-991

c) K-nearest Neighbor Classifier

Nearest Neighbour Classifiers, often known as lazy learners, identifies objects based on closest proximity of training examples in the feature space. The classifier considers k number of objects as the nearest object while determining the class. The main challenge of this classification technique relies on choosing the appropriate value of k.

d) Decision Tree Classifier

A Decision Tree (DT) is a classifier that exemplifies the use of tree-like structure. It gains knowledge on classification. Each target class is denoted as a leaf node of DT and non-leaf nodes of DT are used as a decision node that indicates certain test. The outcomes of those tests are identified by either of the branches of that decision node. Starting from the beginning at the root this tree are going through it until a leaf node is reached. It is the way of obtaining classification result from a decision tree. Decision tree learning is an approach that has been applied to spam filtering. This can be useful for forecasting the goal based on some criterion by implementing and training this model.

B. Ensemble Approach based Classifiers

Ensemble approach facilitates several machine learning algorithms to perform together to obtain higher accuracy of the entire system. Random forest (RF) exploits the concept of ensemble learning approach and regression technique applicable for classification based problems. This classifier assimilates several tree-like classifiers which are applied on various sub-samples of the dataset and each tree casts its vote to the most appropriate class for the input. Boosting is an efficient technique where several unstable learners are assimilated into a single learner in order to improve accuracy of classification. Boosting technique applies classification algorithm to the reweighted versions of the training data and chooses the weighted majority vote of the sequence of classifiers. AdaBoost is a good example of boosting technique that produces improved output even when the performance of the weak learners is inadequate. Boosting algorithms are quite efficient is solving spam filtration problems. Gradient boosting algorithm is another boosting technique based classifier that exploits the concept of decision tree. It also minimizes the prediction loss.

2. LITERATURE SURVEY

Some of the literature surveys are: Vidros, et.al made a significant contribution to properly identify frauds in the online process. A method known as Random Forest Classifier is used by online hiring scams. Electronic scams are distinct from frauds using online hiring.SVM is used for feature selection, while Random Forest Classifier is utilised for detection and classification. Alghamdi and Alharby, et.al made use of the EMSCAD dataset, which is openly accessible and has hundreds of data. Our final result is a 97.41% rate.The corporate logo of a corporation as well as several other crucial characteristics are the two primary points of concentration. Tin Van Huynh, et.al have proposed a model where he gave a statement that for hiring a employee one must consider his knowledge and abilities. The business companies should select a person or student who fits the position of the job.

We are using various different neural networks such as Text CNN, BI-GRU-LSTM, etc, with a pretrained data. This will produce effective output with an 72.71 percent of f1-score. Jiawei Zhang, et.al which concludes that the growth of online social networking is increasing day by day, in terms of both political and economic as well.

The fake news stories may have a wrong impact on users. It is important to know whether the news about something is fake or not. To solve the issue of fake news we use ML algorithms, to examine who are the makers of the news and the subject they have used from online social network. Our aim is to produce the good quality of news. Thin Van Dang, et.al .

Using DNN, the creation of virtual neurons takes place that have random numbers as initial value for weights. The outcome we get is between the values of 0 and 1 range, on multiplying the weight with the input. During the time of training weights are adjusted so, output are classified into different groups. The not so effective patterns are results with some extra layers causing the over fitted problem. Dense layers are employed for data training in the model.

A generic model can be created by cutting down the layers for few parameters which have to be trained. Activation function is the relu and optimizer is the adam. Adam examines the rate of learning for each trainee based on certain factors as part of the training procedure. P. Wang, et.al said in the model that tenets are the fundamentals of neural network which operate the way a brain functions of human. This allows a computer where one pattern is compared with another pattern to determine if they are similar or different. The function with some features and group categories is a neuron. Neural network is the connection of number of nodes in many layers. Jihadists about Perceptron's are arranged in layers and are connected to one another. The rate of mistake can be decreased, by changing the input layers weight through hidden layers.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 987-991

3. METHODOLOGY

D ATA AND M ETHODS

In this section we discuss the dataset utili

D ATA AND M ETHODS

In this section we discuss the dataset utili

D ATA AND M ETHODS

In this section we discuss the dataset utili

MACHINE LEARNING: Machine learning is a set of computer algorithms that, without explicit coding by a programmer, may learn from examples and improve over time. Making recommendations is a common machine learning problem. Machine learning is also utilized for a range of jobs.

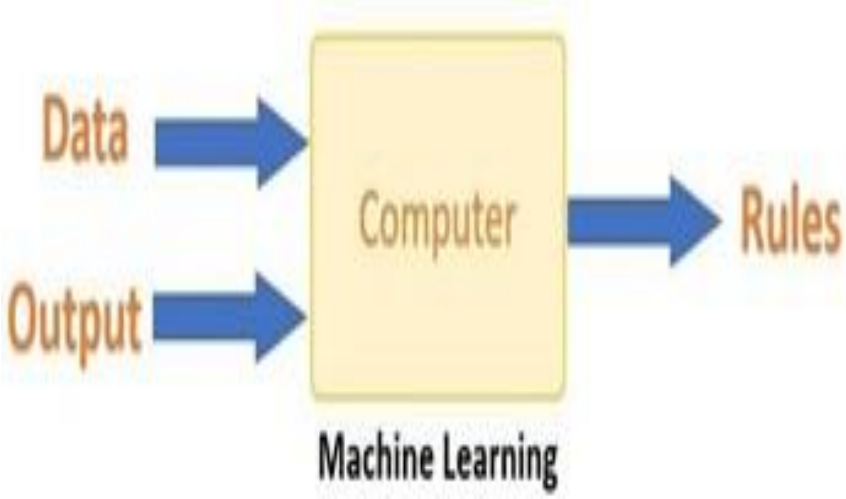

Figure 1: Machine Learning

All the learning occurs in the brain of a machine. The learning of a machine is comparable to how a person learns. Experience is how people learn. Our chances of success are lower than they would be in a known situation when we encounter one. Machines receive the same training.

To get the result more accurately by prediction, the system looks for an example. The machine can predict the result when we provide a similar case. The primary purpose of ML is the learning and then the inference. From the discoveries, the machine learns first.

The data allowed for this finding to be made. The data scientist's ability to carefully select the data to give the computer is one of their most important skills. A feature vector is a collection of attributes that are used to solve an issue. A feature vector can be thought of as a part of data that is utilized to solve a problem.

The machine simplifies reality using some sophisticated algorithms, turning this discovery into a model. As a result, the data are described and condensed into a model during the learning step.

Machine Learning is of two types 1.Supervised Learning 2. Unsupervised Learning

1. **Supervised Learning :** We train the machine with some data that is feed into the computer. The data feed is in the form of input to produce results. It has various different types of classifiers and algorithms in it.
2. **Unsupervised learning :** Without being assigned a specific output variable, an algorithm investigates input data in unsupervised learning. It can be used when we don't know how to classify the data and need of algorithm to look for trends and do it for us.

Random Forest Classifier: The group of decision tree classifiers is called as random forest classifier.

We get the results on majority which is based on voting procedure. The steps here are:

1. From the dataset given, select a random sample.
2. A decision tree is constructed for every sample present over there and produce a result of prediction for each sample.
3. Each prediction result has been voted.
4. Choose the predicted result, with the highest number of voting.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 987-991

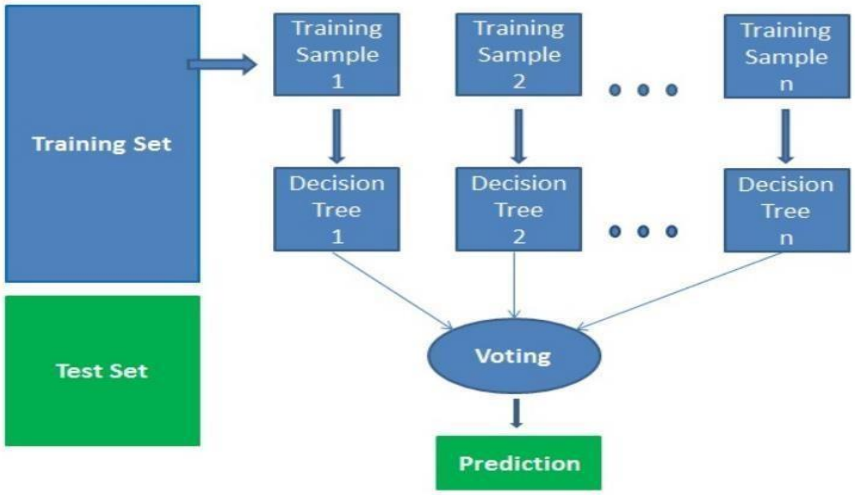

Figure 2: Random Forest Classifier

4. MODELING AND ANALYSIS

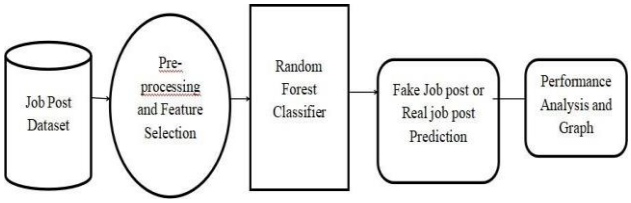

figure 3: System Architecture

There are certain tasks performed such as passing the input and going for preprocessing, then you have to train the data and apply classifier. It will result in prediction. The outcome will be a fake job or legitimate.

OBJUCETIVES

Protecting Job Seekers: One of the primary objectives is to safeguard job seekers from falling victim to fraudulent job postings. By accurately identifying fake job listings, individuals can avoid potential scams that could lead to financial loss or identity theft.

Maintaining Trust in Online Platforms: For online job platforms or recruitment websites, maintaining trust is crucial for their reputation and user retention. Detecting and removing fake job postings can help in ensuring the credibility of the platform and fostering a trustworthy environment for both job seekers and employers.

Reducing Economic Losses: Fake job postings can have significant economic implications, both for individuals and organizations. By using machine learning to detect and remove such postings, economic losses associated with fraudulent activities can be minimized.

Enhancing Efficiency in Recruitment Processes: For companies and recruiters, identifying fake job postings can be time-consuming and resource-intensive. Implementing machine learning algorithms can automate the process of detecting fraudulent listings, thereby streamlining the recruitment process and allowing recruiters to focus on genuine job opportunities.

Improving Data Quality: Fake job postings can distort data analytics and insights derived from job platforms. By filtering out fake listings, machine learning can help improve the quality and accuracy of data used for various analyses, such as market trends, salary insights, and demand for specific skills

5. RESULTS

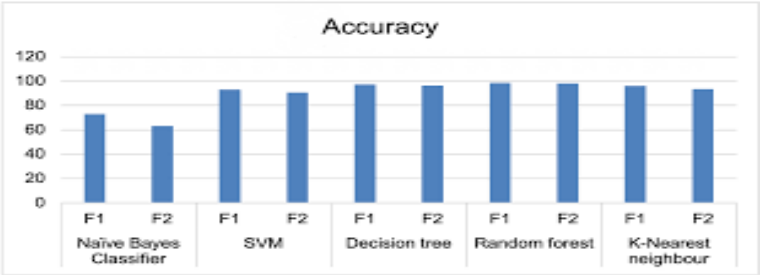

Figure 4: F1 score and accuracy

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 04, April 2024, pp: 987-991

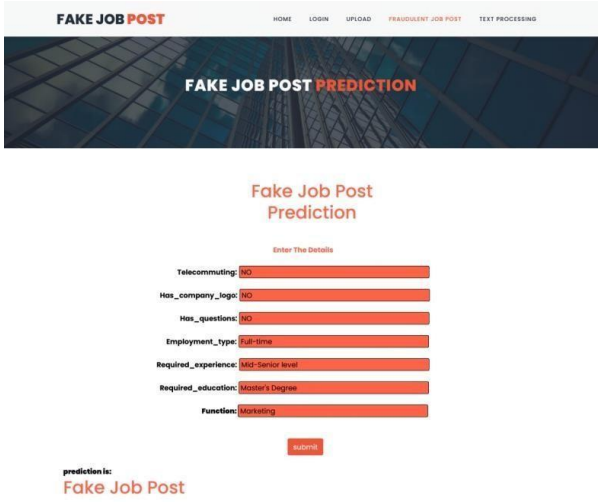

Figure 5: Fake Job Post

6. CONCLUSIONS

The detection of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this project since they might be a very lucrative topic of study and make it difficult to identify the posts of fake job. We made use of EMSCAD dataset, which includes real-time job postings. Random forest classifier gives 98 percent of accuracy then the previously used algorithms like SVM, Decision tree classifier, etc which gives 90 percent of accuracy. We are making the hiring procedure through online safer, by avoiding frauds and scams in the job. Therefore, you can go for applying the jobs through online process. Therefore, avoiding the financial losses of a person and protecting the personal information of a person.

7. REFERENCES

[1] S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, "Automatic Detection of Online Recruitment Frauds:

[2] Characteristics, Methods, and a Public Dataset", Future Internet 2017, 9, 6; doi:10.3390/fi9010006

[3] B. Alghamdi, F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection Journal of Information Security, 2019, Vol 10, pp. 155176, https://doi.org/10.4236/iis.2019.103009

[4] 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), 2020T.

[5] Jiawei Zhang, Bowen Dong, Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[6] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.-T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model," arXiv Prepr. arXiv1911.03644, 2019

[7] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan Luu-Thuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806814, 2016.

[9] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s1338-014-0005-5