ZKG INTERNATIONAL ISSN: 2366-1313

FAKE JOB POST PERDITION USING MACHINE LEARNING MECHANISM

[1]KORUKONDA VENKATA RATHNAM, [2]SARABU VENKATABALAJI,
[3]B.VENKATESWARLU, [4]SYED KALEEM, [5]NIMMAKAYALA MAHESH REDDY

[1]Assistant Professor, Dept. Of AI, ABR College of Engineering and Technology, Kanigiri

[2, 3, 4, 5,] BTech Student, Dept. Of AI, ABR College of Engineering and Technology, Kanigiri

Abstract: *In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.*

Keywords: *Machine learning, fake job post prediction, EMSCAD Dataset*

I. 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. The companies post about the job to make the hiring process easier and more immediate. We are using different data mining techniques to solve the issue of fake job posting. On applying Random Forest Classifier, it gives the best results, in identifying the

ISSN: 2366-1313

fake job postings which is better than the previously used. This helps them to avoid financial losses like they may ask you to pay application fee, for getting registered or they may ask money in different forms, as a part of procedure in recruitment or others. All companies go for the online process of hiring employees, by posting the job details, if the information entered by the student or user matches the job details then they are hired by the company. The need for job by the people, exploring on internet may blindly have trust on anyone and disclose their information to any fake job postings, which can be misused like bank information, etc. The person seeking for a job should be careful while applying for job as they may get into the trap of fake people posting fake jobs, which can be misused for some other purpose. The classifier we are using is random forest which gives much improved result than the previously used algorithms. The developed project gives better outcomes in terms of accuracy, efficiency, cost and time. The online procedure of hiring people for employment has moved towards failure because of such frauds and scams taking place that make misuse of personal information, and harming the reputation of a company.

II. LITERATURE SURVEY

Many researches occurred to predict if a job post is real or fake. A good number of research works are to check online fraud job advertiser.

Vidros [1] et al. identified job scammers as fake online job advertiser. They found statistics about many real and renowned companies and enterprises who produced fake job advertisements or vacancy posts with ill-motive. They experimented on EMSCAD dataset using several classification algorithms like naive bayes classifier, random forest classifier, Zero R, One R etc. Random Forest Classifier showed the best performance on the dataset with 89.5% classification accuracy. They found logistic regression performing very poor on the dataset. One R classifier performed well when they balanced the dataset and experimented on that. They tried in their work to find out the problems in ORF model (Online Recruitment Fraud) and to solve those problems using various dominant classifiers. Alghamdi [2] et al. proposed a model to detect fraud exposure in an online recruitment system. They experimented on EMSCAD dataset using machine learning algorithm. They worked on this dataset in three steps- data pre-processing, feature selection and fraud detection using classifier. In the pre-processing step, they removed noise and html tags from the data so that the general

text pattern remained preserved. They applied feature selection technique to reduce the number of attributes effectively and efficiently. Support Vector Machine was used for feature selection and ensemble classifier using random forest was used to detect fake job posts from the test data. Random forest classifier seemed a tree structured classifier which worked as ensemble classifier with the help of majority voting technique. This classifier showed 97.4% classification accuracy to detect fake job posts

Some of the literature surveys are: Vidros, et.al [3] 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 [4] 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 [5] have proposed a model where he gave a statement that for hiring an 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 a 72.71 percent of f1-score.

Jiawei Zhang, et.al [6] 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 [7]. 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 is 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 [8] 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 [9] 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.

III. METHODOLOGY

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

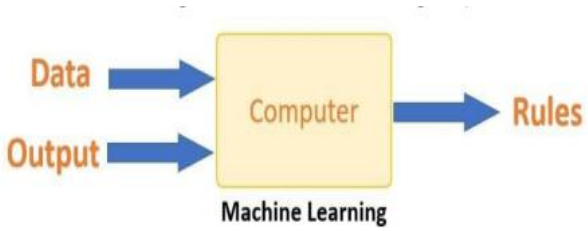

Fig.1 Machine learning

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

ISSN: 2366-1313

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.

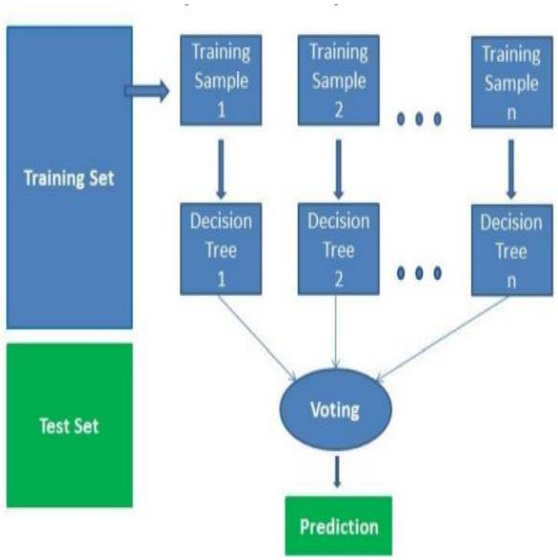

Fig.2 Random Forest Classifier

MODELING AND ANALYSIS

The project is to find the phoney jobs to avoid users getting into the scams. This makes assurance that the data they provide at the time of recruitment will not be misused. We are working on a EMSCAD dataset to find better results using different algorithms. The dataset for fake job post is collected and pre-processed. The feature selection is the process of selecting some important features from the data required for analysing and getting a proper output. We are applying the Random Forest Classifier to detect whether the job posted is a fake or a legitimate one.

SYSTEM ARCHITECTURE

ISSN: 2366-1313

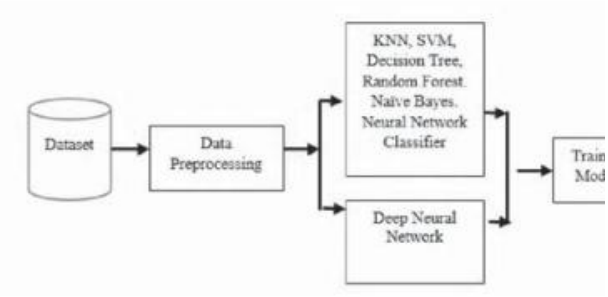

Fig.3 Proposed system architecture

Dataset

We have used EMSCAD to detect fake job post. This dataset contains 18000 samples and each row of the data has 18 attributes including the class label. The attributes are job_id, title, location, department, salary_range, company_profile, description, requirements, benefits, telecommunication, has_company_logo, has_questions, employment_type, required_experience, required_education, industry, function, fraudulent (class label). Among these 18 attributes, we have used only 7 attributes which are converted into categorical attribute. T elecommuting, has_company_logo, has_questions, employment_type, required experience, required_education and fraudulent are changed into categorical value from text value. For example, "employment_type" values are replaced like this- 0 for "none", 1 for 'full-time", 2 for "part-time" and 3 for "others", 4 for "contract' and 5 for "temporary". The main goal to convert these attributes into categorical form is to classify fraudulent job advertisements without doing any text processing and natural language processing. In this work, we have used only those categorical attributes.

IV. RESULTS AND DISCUSSIONS

We have implemented the work using EMSCAD dataset in google colab. In case of conventional machine learning algorithms like KNN, Random Forest, SVM etc. we have used hold out cross validation. 80% of the total data was used for training and 20% was used for testing and checking the model performance. In KNN model, we have applied K value from 1 to 40 and minimum error is found when k= 13. Mean error rate was less than 0.05 during the training process (Fig.2). RBF kernel is used in SVM and gamma value = 0.001 is also use

Table.1 Comparison among the Classifiers

In Table I, the classification accuracy, precision, recall and f1 score of all these classifiers are shown. We have achieved approximately 97% classification accuracy (highest) for Random Forest classifier. We have analysed f1 score also to check if the model works well at both false positive and false negative samples. The equations

of the measured parameters are given below

Model	Accuracy	Precision	Recall	F1 Score
K Nearest Neighbor	95.2	93	95	93
Random Forest Classifier	**96.5**	93	95	93
Decision Tree	96.2	93	95	93
Support Vector Machine	95	90	95	92
Naïve Bayes Classifier	91.35	95	96	95
Multilayer perceptron	96	94	95	93

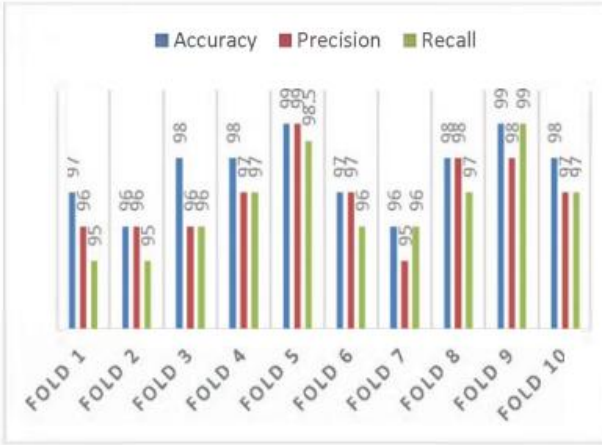

Fig.4 F1-score and Accuracy

V. CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, KNN, Naive Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). This work shows a comparative study on the evaluation of traditional machine learning and deep learning-based classifiers. We have found highest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99 % accuracy for DNN (fold 9) and 97.7% classification accuracy on average for Deep Neural Network.

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] 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.

[4] 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.

[5] 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/s13388-014-0005-5

[6] Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882, 2014.

[7] 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.

[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. 806 814, 2016.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209

[11] Prasadu Peddi (2015) "A review of the academic achievement of students utilising large-scale data analysis", ISSN: 2057-5688, Vol 7, Issue 1, pp: 28-35.