2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) | 978-0-7381-3042-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICREST51555.2021.9331230

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques

Sultana Umme Habiba
Dept. of CSE
Green University of Bangladesh
Dhaka Bangladesh
habiba@cse.green.edu.bd

Md. Khairul Islam
Dept. of CSE
Khulna University of Engineering & Technology
Khulna, Bangladesh
mdkislam27@gmail.com

Farzana Tasnim
Dept. of CSE
International Islamic University Chittagong
Dhaka, Bangladesh
farzanatasnim34@gmail.com

***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, naïve 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.**

***Index Terms*—false job prediction, deep learning, data mining.**

I. Introduction

In modern time, the development in the field of industry and technology has opened a huge opportunity for new and diverse jobs for the job seekers. With the help of the advertise- ments of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job posts has increased the percentage of fraud job postings which causes harassment to the job seekers. So, people lacks in showing interest to new job postings due to preserve security and consistency of their personal, academic and professional information. Thus the true motive of valid job postings through social and electronic media faces an extremely hard challenge to attain people's belief and reliability. Technologies are around us to make our life easy and developed but not to create unsecured environment for professional life. If jobs posts can be filtered properly predicting false job posts, this will be a great advancement for recruiting new employees. . Fake job posts create inconsistency for the job seeker to find their preferable jobs causing a huge waste of their time. An automated system to predict false job post opens a new window to face difficulties in the field of Human Resource Management.

II. Background Study

A. Fake Job Posting: Job Scam

Online job advertisements which are fake and mostly willing to steal personal and professional information of job seekers instead of giving right jobs to them is known as job scam. Sometimes fraudulent people try to gather money illegally from job seekers. A recent survey by ActionFraud from UK has shown that more than 67% people are at great risk who look for jobs through online advertisements but unaware of fake job posts or job scam [2]. In UK, almost 700000 job seekers complained to lose over $500000 being a victim of job scam. The report showed almost 300% increase over the last two years in UK [2]. Students, fresh graduates are being mostly targeted by the frauds as they usually try to get a secured job for which they are willing to pay extra money. Cybercrime avoidance or protection techniques fail to decrease this offence since frauds change their way of job scam very frequently.

B. Common types of Job Scam

Fraudsters who want to gain other people's personal information like insurance details, bank details, income tax details, date of birth, national id create fake job advertisements. Advance fee scams occur when frauds ask for money showing reasons like admin charges, information security checking cost, management cost etc. Sometimes fraudsters act themselves as employers and ask people about passport details, bank statements, driving license etc. as pre-employment check. Illegal money mulling scams occur when they convince students to pay money into their accounts and then transfer it back [2]. This 'cash in hand' technique causes to work cash in hand without paying any tax. Scammers usually create fake company websites, clone bank websites, clone official looking documents etc. to trap job seekers. Most of the job scammers try to trap people through email rather than face to face communication. They usually target social media likeLinkedIn to prove themselves as recruitment agencies or headhunters. They usually try to represent their company profile or websites to the job seeker as realistic as possible. Whatever the type of job scam they use, the always target the job seeker to fall in their trap, collecting information and making benefit either earning money or any other things[6], [7].

C. Related Works

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.

Huynh [3] et al. proposed to use different deep neural network models like Text CNN, Bi-GRU-LSTM CNN and Bi-GRU CNN which are pre-trained with text dataset. They worked on classifying IT job dataset. They trained IT job dataset on TextCNN model consisting of convolution layer, pooling layer and fully connected layer. This model trained data through convolution and pooling layers. Then the trained weights were flattened and passed to the fully connected layer. This model used softmax function for classification technique. They also used ensemble classifier (Bi-GRU CNN, Bi-GRU-LSTM CNN) using majority voting technique to increase classification accuracy. They found 66% classification accuracy using TextCNN and 70% accuracy for Bi-GRU- LSTM CNN individually. This classification task performed best with ensemble classifier having an accuracy of 72.4%.

Zhang [4] et al. proposed an automatic fake detector model to distinguish between true and fake news (including articles, creators, subjects) using text processing. They had used a custom dataset of news or articles posted by PolitiFact website twitter account. This dataset was used to train the proposed GDU diffusive unit model. Receiving input from multiple sources simultaneously, this trained model performed well as an automatic fake detector model.

Researchers experimented a good number of classifiers and feature selection technique to achieve good performance in the field of fake job post classification. Text processing using deep learning model, feature selection using support vector machine, data pre-processing etc. were mentioned approach to apply[8], [9], [10], [11], [12]. We have proposed to use deep neural network to predict job scams. We have applied the training method only on the categorical attribute of the EMSCAD dataset instead of using text data. This approach reduces the number of trainable attribute effectively with less processing time. We have made a comparative study on the same features of EMSCAD dataset using K Nearest Neighbor, Naive Bayes classifier, fuzzy KNN, decision tree, support vector machine, random forest classifier and neural network.

III. METHODOLOGY

We have used different data mining techniques to predict if a job post is fake or not. We have trained EMSCAD data in the classifiers after a pre-processing step. The trained classifier act as an online fake job post detector.

A. Neural Network

Neural Network works on the core principle of human brain's function. It makes a computer capable of analyzing a specific pattern with another pattern to define how much the two are similar or different. The mathematical function extracting features and classifying specific pattern is known as a neuron. Neural Network consists of many layers of inter- connected nodes. Each node which is called a perceptron acts as a multiple linear regression. This perceptron pass through the output of multiple linear regression into a non-linear activation function. Layers in which perceptrons are arranged, are interconnected. The hidden layers adjust the weights of the input layers to optimize the error rate. Neural Network works as a supervised learning classifier.

B. Deep Neural Network

An Artificial Neural Network (ANN) which contains multiple layers between the input and output layer is called Deep Neural Network. DNN works on feed forward algorithm. Data flow is directed from input to output layer [13]. DNN creates a number of virtual neurons initialized with a random numerical value as connection weights. This weight is multiplied with the input and produce an output between 0 and 1. The training process adjust the weights to classify the output efficiently. Added layers make the model to learn rare patterns which leads the model to overfitting. Dropout layers reduce the number of trainable parameters to make the model generalized. In this paper, we have used a sequential model of dense layers for training the data, relu as activation function and adam as optimizer. During training process, adam calculates individual learning rates on different parameters as this is an adaptive learning method.

C. Other classifiers

K Nearest Neighbor, Random Forest Classifier, Decision Tree, Naïve Bayes Classifier, Support Vector Machine (RBF kernel)

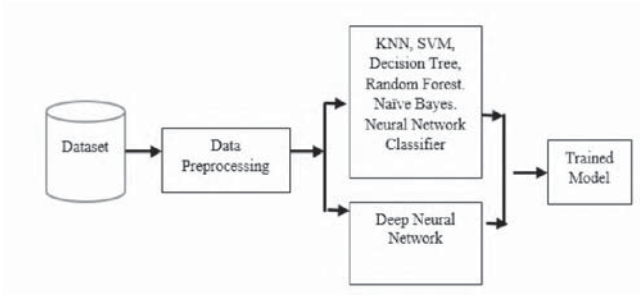

Fig. 1. Proposed Methodology

and Multilayer Perceptron (MLP) are the classifiers where our work dataset is trained.

D. 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 attribute, we have used only 7 attributes which are converted into categorical attribute. Telecommuting, 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. EXPERIMENTAL RESULT ANALYSIS

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 used.

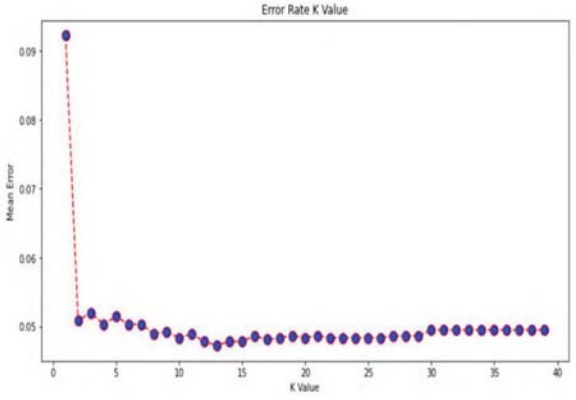

Fig. 2. Relation between mean error and K value in KNN

TABLE I COMPARISON AMONG THE CLASSIFIERS

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

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 analyzed 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:

Accuracy = TP+TN/TP+FP+FN+TN
Precision = TP/TP+FP
Recall = TP/TP+FN
F1 Score = 2*(Recall * Precision) / (Recall + Precision)
(TP= True Positive, TN= True Negative, FP= False Positive, FN= False Negative)

10 fold cross validation is used to train the data in deep neural network model. 60% data was used for training, 20% was used for measuring validation accuracy and remaining 20% was used to test the performance of the model. Validation accuracy indicates the level of performance of the model on unseen data. We have noticed a good relation between the validation and training accuracy in each epoch of training. If the validation accuracy is higher than the training accuracy, we can find the trained model as a generalized one. To reduce overfitting of the model, we have used a dropout layer. This layer reduces trainable parameters at each step of training so that the model will perform well beyond the training dataset.

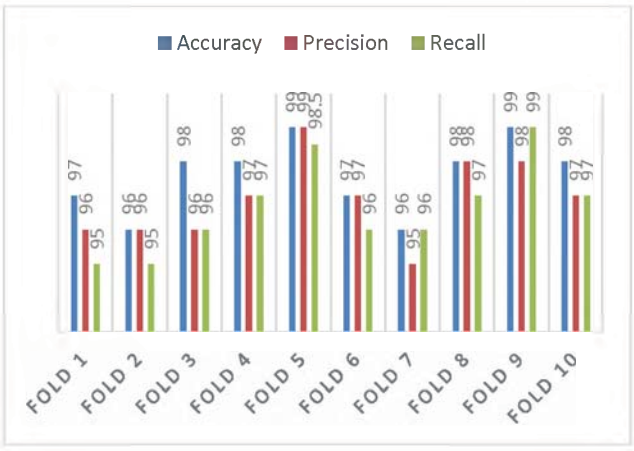

Fig. 3. Accuracy, Precision and Recall for 10 Folds in DNN model

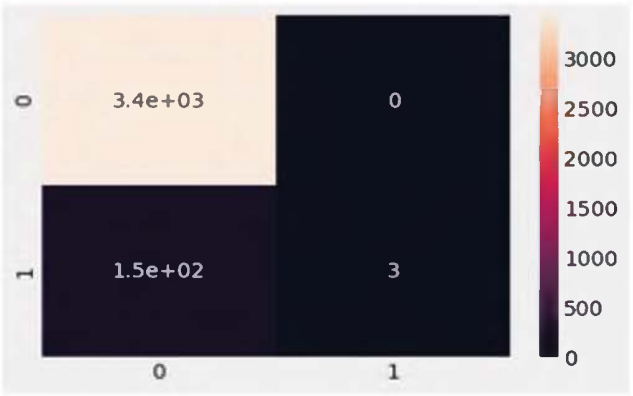

Fig. 4. Confusion matrix for DNN Model (Fold 2)

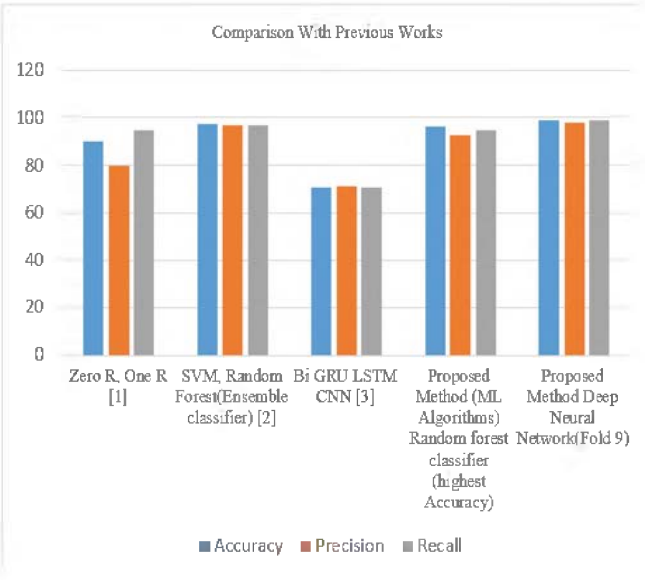

Fig. 5. Comparison of our proposed method with the previous works

From fig. 3. We can notice the accuracy, precision and recall of each fold of deep neural network model. Fold 2 and 7 have shown 96% classification accuracy and fold 5 and 9 have shown the highest accuracy which is **99%**. On average the trained deep neural network model shows 97.7% classification accuracy. Since we have worked using a class unbalanced dataset, only accuracy can't measure the performance of a generalized model. The values of both precision and recall are also good for the trained model. In fig. 4. The confusion matrix for the DNN model (fold 2) is given. Most of the test data are positioned diagonally. In fig. 5. We have made a comparison among the previous works and our proposed methodology. We have implemented both conventional machine learning algorithms and deep learning model. In first case, we have achieved highest classification accuracy in random forest classifier **(96.7%)** and in deep learning model (DNN), we have achieved **99%** accuracy for fold 9 where fold 9 was used as test data.. The average classification accuracy (10 fold) for DNN model is 97.7%.

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, Naïve 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.

Reference

[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/jis.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] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55-72. https://doi.org/10.5121/ijnsa.2016.8405

[12] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen."Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.

[13] 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.

[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14–17 December 2014; pp. 899–904.

[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. InProceedings of the 21st international conference on World Wide Web, Lyon, France, 16–20 April 2012; ACM: New York, NY, USA, 2012; pp. 201–210.

[16] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174. https://doi.org/10.1016/j.eij.2014.07.002