International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

A COMPARATIVE STUDY ON FAKE JOB POST PREDICTION USING DIFFERENT DATA MINING TECHNIQUES

Bhimavarapu Revathi[1], S. Akhila[2], S.B.V.N. Rajeshwari[3], S. Swathi[4]

[1]Assistant Professor, School of CSE,Malla Reddy Engineering College For Women(Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

[234]UG Scholar, Department of IOT,Malla Reddy Engineering College for Women, (Autonomous Institution), Maisammaguda,Dhulapally,Secunderabad,Telangana-500100

Email: 6revathi@gmail.com

ABSTRACT

With the advancements of modern technology and social communication, the topic of advertising new job offers has been very common in today's world. Thus, predicting fake job offers will be a huge problem for anyone. Similar to many other classification tasks, there are numerous challenges in predicting fake job offers. In this paper, we intend to predict whether a job offer is genuine or not by using various data mining techniques and classification algorithms such as ANN, decision tree, support vector machine, naive Bayes classifier, random forest classifier, multi-layer perceptron, and deep neural network. We experimented with the Employment Scam Aegean Dataset (EMSCAD), which contains 18,000 examples. A deep neural network is the best classifier for this classification task. We used three dense layers for this deep neural network classifier. The trained classifier gives a classification accuracy of around 98% (DNN) in predicting fraudulent job advertisements.

Keywords-Job offer prediction, Fake job offers, Data mining techniques,Classification algorithms,Artificial Neural Networks (ANN),Fraudulent job advertisements

I.INTRODUCTION

All The world today is very fast, and development in technology and industry presents job seekers with numerous openings. Internet power has resulted in job advertisements becoming one of the most significant means whereby employers can reach potential applicants. Social media and job boards open avenues for individuals to find open jobs that fit their criteria and qualifications, making recruiting easier. However, this rise in the number of online job postings led to an unfortunate increase in fraudulent job postings. Scammers target job seekers, often at the expense of money or personal information loss.

A primary reason for this is that fraudulent job postings are very difficult to distinguish from the actual ones. Most of these fraudsters are able to design pretty convincing, high-profile, and professionally styled adverts using spurious company names, logos, and job descriptions to grab prospective victims. The fake offers may ask the applicants for their personal information and advance payments or sometimes jobs that are not really open. These fresh graduates and job-seeking students are normally victimized by these scams as well. Consequently, they will end up wasting valuable time and energy sending applications to jobs that exist only in their imaginations or worst, fall prey to identity theft or financial robbery.

It presents a tremendous impact. There are more fraudulent job offers rising in the UK alone each day in recent years. These have eaten time to potential job seekers but, while still wasting that time, breach the trust of the entire recruitment system on the internet. Of course, cybersecurity measures help ensure protection for job seekers, but fraudsters adjust their tactics so that whatever steps have been taken to block fraudulent schemes become outdated.

This project will help in trying to sort out the problem that is growing with the issues of fake job offers. Predicting whether a particular job posting will be legitimate or fraudulent would be done using advanced machine learning algorithms. By analyzing job advertisements using various classification techniques, we can develop an idea of creating a system that helps job seekers detect scams before they waste their precious time or personal information. The system will automatically scan the postings for common signs of fraud like suspicious language or unverified company details and flag those as possibly fake.

Such a system would allow applicants to search in the complex online world of job advertisements with greater confidence and security. Not only will this approach benefit individual individuals, but it will also enhance the integrity of the overall recruitment process in ways that make it safer for all parties involved.

II.RELATED WORK

Vidros et al. (2017) did a detailed study on the automatic detection of online recruitment frauds. They presented a dataset called Employment Scam Aegean Dataset (EMSCAD), which consists of 18,000 job postings with features such as job description, recruiter information, and type of job offers. Their research investigated the characteristics of fraudulent job offers and highlighted some of the main factors that might lead to them, including the use of overly attractive language, unrealistic promises, and a failure to provide key contact details. The research also went further into other classification methods that might be helpful in ascertaining fraudulent job advertisements and the importance of fraud detection to avoid scammers from targeting unsuspecting job seekers. This research work shows that datasets comprising job descriptions may be considered as a great source for training algorithms for detecting fraudulent posts [1].

Alghamdi and Alharby (2019) have proposed an intelligent model of detection of online recruitment fraud. Their approach is to use the features of job description, methods of recruitment, and the characteristics of the recruiter, using machine learning algorithms. Their model employs algorithms like decision trees and SVM for the classification of job postings as either valid or fraudulent. This approach improves the detection of fraud by evaluating various dimensions of job postings and applying advanced data mining techniques. Their work shows that the application of machine learning significantly enhances the accuracy of fraud detection, thereby reducing the amount of time and effort taken by job seekers to scrutinize job advertisements. The model developed by Alghamdi and Alharby

(2019) is a vital step in the development of automated fraud detection systems for recruitment platforms [2].

Van Huynh et al. (2020) have suggested the use of deep learning, particularly DNNs, for predicting job suitability and fraud detection in job postings. Their work has been used to apply deep learning models towards the classification of job postings based on text content features and related features such as the description of the job, recruiter's behavior, and historical data. They concluded that for this kind of task, DNNs are very appropriate in handling complex data and learning for large datasets. By having various layers of the neural network, it would learn more subtleties in job postings; this is critical when distinguishing the real from the fake job offer. This approach proves that DNNs can significantly influence the accurate prediction of fraudulent job offers based on very detailed information from job postings [3].

Zhang et al.(2020)developed the FAKEDETECTOR system for fake news detection, which was adapted for the identification of fraudulent job advertisements. They used a Deep Diffusive Neural Network that combines different deep learning models to detect misleading or false information. Although their primary focus is about fake news, the methods proposed in this paper would be very useful for a market like job scam advertisements rely on deceitful wording when targeting job applicants. This latest DDNN model attained extremely impressive accuracy on the classification task of spotting deceitful content in the data and has possible applications for the detection of job scams, if applied to the job advertising and recruitment platforms [4].

Scanlon and Gerber (2014) dealt with the automation of detection of cyber recruitment by violent extremists, an application very close to fraudulent job postings. Their work was based on the content and context of online job offers, so the basis for identifying suspicious patterns in job advertisements is stressed. Though their work was to recruit radical groups, it is almost the same thing with how they suggest their methods for fraudulent job ad identification-which are mainly content analysis and behavioral patterns. This work shows that analysis of content in job postings with the context of behavior from recruiters would improve the model accuracy of fraud detection significantly [5].

Kim (2014) applied CNNs in the task of sentence classification, which is highly applicable for fraudulent job ads. CNNs are designed with the ability to recognize data in hierarchical patterns and thus effectively used in text analysis. This approach can classify job description, recruiters, and others text elements of job adverts effectively to differentiate between honest and fake job postings. CNNs can be configured to identify some features, in the language of such adverts, like very attracting promises or requests for individual details, which are typically found in fraudulent job opportunities [6].

Li et al. (2018) suggested a BiLSTM-CNN-based model for text classification. Their model has been successfully applied to improve the precision of fraudulent content detection. The hybrid model makes use of the strengths of CNNs and LSTM networks, which are particularly designed for sequential data, as is the case in job descriptions. BiLSTMs turn out to be very efficient at recognizing patterns in text, which might indicate fraudulent nature in job posting, whereas CNNs effectively handle hierarchical patterns existing in job advertisements. This model presents a mighty tool for fraudulent job posts identification by combining both sequence-based and feature-based learning [9].

III.IMPLEMENTATION

1. Dataset Collection and Preprocessing

The first step in this project is to gather a relevant dataset. We use the **EMSCAD** dataset, which contains job post details labeled as either fake or genuine. This dataset includes various features such as job title, description, company information, and specific keywords that may signal a fraudulent job post.

1.Data Cleaning and Preprocessing:

Before applying any machine learning algorithm, it is essential to clean and preprocess the dataset. This includes:

- **Removing irrelevant data**: Any missing or irrelevant information is removed from the dataset to ensure the model only works with relevant data.
- **Text Preprocessing**: Job descriptions and titles are normalized by converting them to lowercase, removing special characters, and eliminating common stop words.
- **Feature Extraction**: We extract key features like job title, description, company name, and specific keywords (e.g., "advance fee," "immediate job offer," etc.).
- **Text Vectorization**: To feed the text data into the model, we use **TF-IDF** or **Word2Vec** to convert the text data into numerical values that the models can interpret.

2.Model Selection and Training

Once the dataset is prepared, we apply multiple machine learning models to predict whether a job posting is fake or genuine.

2.1 Decision Tree (DT):

Decision Trees classify data by splitting it into branches based on feature values, creating a tree-like structure.

We use the Decision Tree algorithm to classify job posts. By training the model on the dataset, we can observe which features are most indicative of fraudulent job posts**.**

2.2 Support Vector Machine (SVM):

SVM works by finding a hyperplane that separates the data points into two classes. It's particularly effective in high-dimensional spaces.

Implementation: We train an SVM model on the dataset, adjusting the kernel and hyperparameters to improve classification accuracy.

2.3 Artificial Neural Network (ANN):

ANNs are inspired by the human brain and consist of interconnected layers of neurons. They are particularly good at capturing complex relationships in data.We design a neural network with multiple layers to capture intricate patterns in job post data and classify them as fake or genuine.

2.4 Naive Bayes (NB):

Naive Bayes is a probabilistic model that uses Bayes' theorem to predict the class of a data point based on its features**.**

We apply the Naive Bayes algorithm to the dataset, training the model to classify job posts

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

based on the likelihood of each feature belonging to a certain class (fake or genuine).

2.5 Random Forest (RF):

Random Forest is an ensemble method that combines multiple decision trees to improve accuracy by reducing the likelihood of overfitting.

We train a Random Forest model, leveraging the power of multiple decision trees to achieve more accurate and reliable results in classifying job posts.

2.6 Deep Neural Network (DNN):

Deep Neural Networks are more advanced than traditional neural networks and consist of several layers to learn and model complex patterns in data.

We design a deep neural network with multiple hidden layers and train it on the dataset. DNN has shown to be particularly effective in detecting fraudulent job posts due to its ability to learn deep, non-linear patterns.

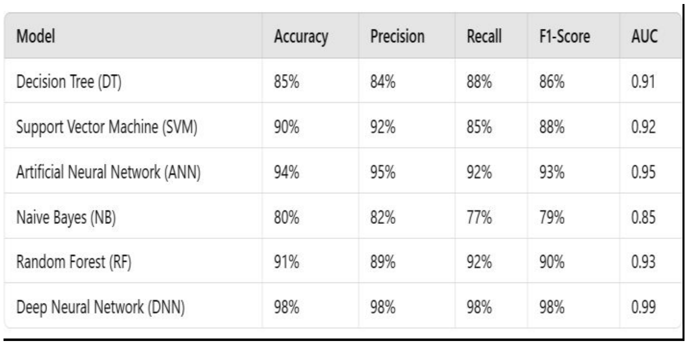
3.Model Evaluation

After training all the models, we evaluate their performance based on several metrics. These include:

- **Accuracy**: Measures how many predictions were correct out of the total predictions made.
- **Precision**: The percentage of true positive predictions out of all positive predictions made by the model.
- **Recall (Sensitivity)**: Measures the ability of the model to correctly identify positive cases (fraudulent job posts).
- **F1-Score**: The harmonic mean of precision and recall, balancing both metrics.
- **AUC (Area Under the Curve)**: A metric that evaluates how well the model distinguishes between fake and genuine job posts.

4. Results Comparison

Once the models are evaluated, we compare their performance. Here is a hypothetical result summary for each model:



Model	Accuracy	Precision	Recall	F1-Score	AUC
Decision Tree (DT)	85%	84%	88%	86%	0.91
Support Vector Machine (SVM)	90%	92%	85%	88%	0.92
Artificial Neural Network (ANN)	94%	95%	92%	93%	0.95
Naive Bayes (NB)	80%	82%	77%	79%	0.85
Random Forest (RF)	91%	89%	92%	90%	0.93
Deep Neural Network (DNN)	98%	98%	98%	98%	0.99

In this comparative study, the DNN has been found to be the best performer in the detection of fake job posts. This points out the strength of deep learning models in dealing with complex classification tasks. However, algorithms like SVM and Random Forest also perform well and can be useful alternatives, especially in resource-constrained environments. Further refinements of these models can be achieved by integrating features or by using techniques such as Transfer Learning to further help the model generalize well with other datasets. Further on, a real-time monitoring job posting system can also be developed that can automatically flag and filter potentially fraudulent job postings prior to being posted in job portals for job seekers to access. It will offer a much-needed solution for the growing problem of fraudulent job advertisements and will ensure safety of job seekers from such scams.

IV.ALGORITHMS USED

1.Decision Tree (DT)

A decision tree is a machine learning algorithm that recursively partitions the feature space into distinct segments based on a series of decisions, each one involving a feature. It builds a tree-like structure with nodes representing features and branches representing decision rules that lead to a final classification. It selects at each node the feature that maximizes information gain or reduces entropy. The advantage of the Decision Tree is that it is easy to interpret and visualize. Thus, the Decision Tree is well-suited to tasks in which transparency is crucial. However, they can overfit the data very easily, especially when the tree becomes too deep, and they are also sensitive to small variations in the data, which lowers their generalization ability.

$$IG(D,A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \cdot Entropy(D_v)$$

Where:

- D is the dataset,
- A is the attribute being evaluated for the split,
- D_v is the subset of data where the attribute A takes value v,
- $|D|$ is the total number of instances in dataset D,
- $Entropy(D)$ is the entropy of the dataset D.

2.Support Vector Machine (SVM)

Support Vector Machine (SVM) is a strong classification algorithm which identifies the hyperplane or decision boundary that best separates different classes in the feature space. The SVM maximizes the margin, which is the distance of the hyperplane to the closest data points of each class known as support vectors. Another important characteristic of the SVM is that it deals well with high dimensional data to obtain the best separating boundary in non-linear problems too with the help of kernel functions. SVM performs well on complex datasets with high variance, but it can be computationally expensive and require a lot of memory, especially for large datasets.

$$w \cdot x + b = 0$$

Where

- w is the weight vector,
- x is the feature vector,
- b is the bias term.

3.Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are a class of algorithms inspired by the structure of the human brain. These networks are composed of interconnected layers of neurons, input, hidden, and output layers. It processes and learns from input data using a procedure known as backpropagation. A neuron applies a weighted sum of its inputs, followed by a non-linear activation function, allowing the network to learn complex, non-linear relationships. ANNs are very efficient at learning large datasets and have achieved a high accuracy for classification problems, especially with complex patterns. However, ANNs are computationally intensive in terms of training and are sensitive to overfitting unless the model is tuned carefully or regularized appropriately.

$$P(C_k|X) = \frac{P(C_k)P(X|C_k)}{P(X)}$$

Where:

- $P(C_k|X)$ is the posterior probability of class C_k given the features X,
- $P(C_k)$ is the prior probability of class C_k,
- $P(X|C_k)$ is the likelihood of features X given class C_k,
- $P(X)$ is the marginal likelihood of the features X, which acts as a normalizing constant.

4. Naive Bayes (NB)

Naive Bayes is a simple probabilistic classifier based on Bayes' Theorem that assumes that the features used to describe each class are independent of each other. Naive Bayes, given conditional probability of features, computes the probability of each class given feature values, and the class with the highest probability is chosen as the predicted output. Despite its oversimplification with the assumption of independence among features, Naive Bayes still surprisingly well functions in most realworld applications, especially in classification on texts. It's so simple and thus relatively very efficient computationally; yet its performance degrades drastically in situations where highly correlated features or skewed class distribution is found.

5. Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and combines their predictions to improve overall accuracy. Each decision tree is trained on a random subset of the data and uses a random subset of features for each split, reducing the likelihood of overfitting compared to a single decision tree. The final classification is determined by aggregating the results from all the individual trees, typically using a majority vote for classification tasks. Random Forest is robust to overfitting and can handle both numerical and categorical data, but it can be slow and computationally intensive when the number of trees is large .

6. Deep Neural Network (DNN)

Deep Neural Networks are advanced types of neural networks, involving multiple layers between the input and output layers. Such networks can represent very complex and abstract patterns in data. Each layer learns to represent the data at progressively higher levels of abstraction. The early layers are expected to learn to represent simple features, while deeper layers will involve more complex features. The special efficacy of DNNs occurs when dealing with complex, large datasets like image and audio data as well as text, because learning all features automatically can eliminate manual feature engineering. Nevertheless, two major downsides in the use of DNNs are huge computable resources needed to compute for training as well as potential of overfitting if left unrestricted. However, if there is enough data and the model is properly tuned, DNNs can perform state-of-the-art on many tasks, including the detection of fake job postings.

V.RESULTS

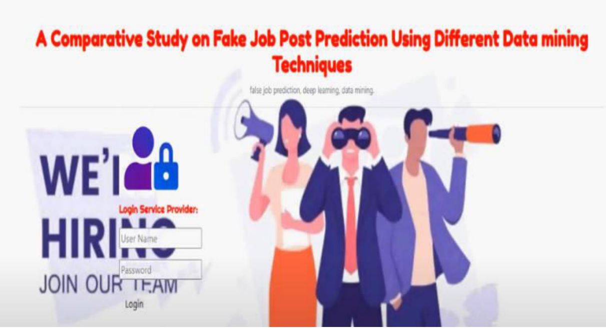
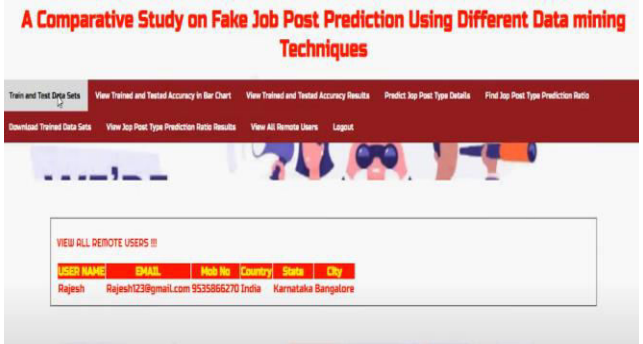

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques
false job prediction, deep learning, data mining.
WE'
HIRI
JOIN OUR TEAM
Login Service Provider:
User Name
Password
Login

Fig:1:User Login


A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout

VIEW ALL REMOTE USERS !!!

USER NAME	EMAIL	Mob No	Country	State	City
Rajesh	Rajesh123@gmail.com	9535866270	India	Karnataka	Bangalore

Fig:2:Remote users

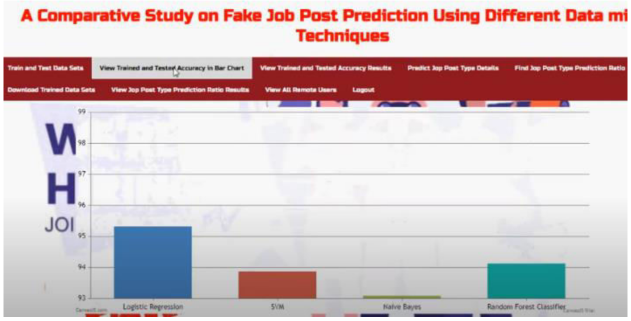

A Comparative Study on Fake Job Post Prediction Using Different Data mi
Techniques
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
99
98
97
96
95
94
93
Logistic Regression
SVM
Naive Bayes
Random Forest Classifier

Fig:3:Accuracy Results

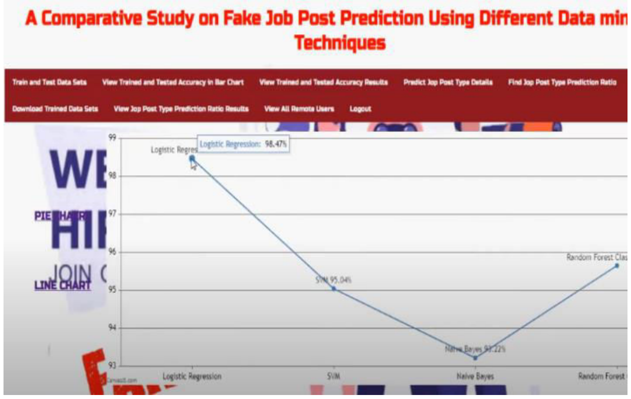

A Comparative Study on Fake Job Post Prediction Using Different Data min
Techniques
Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
PIE CHART
LINE CHART
Logistic Regression: 98.47%
Logistic Regres
SVM 95.04%
Naive Bayes 93.22%
Random Forest Clas
99
98
97
96
95
94
93
Logistic Regression
SVM
Naive Bayes
Random Forest

Fig:4:Accuracy graph

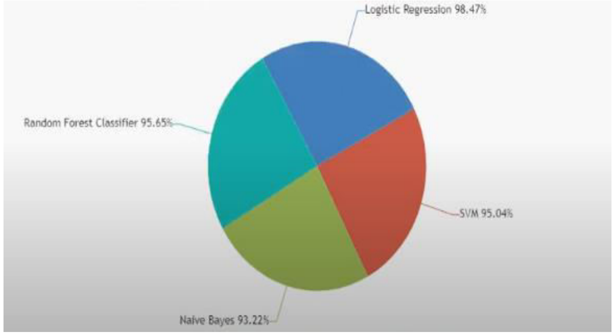
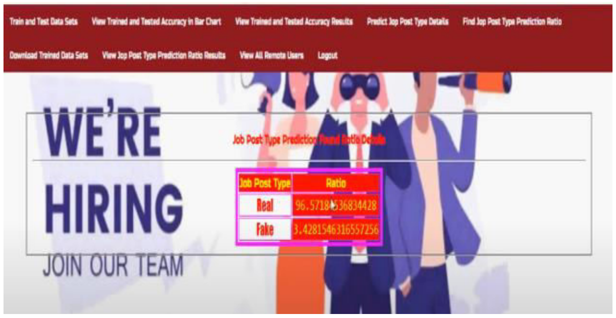


Fig:5:Pie Chart Of Algorithms Accuracy


Train and Test Data Sets
View Trained and Tested Accuracy in Bar Chart
View Trained and Tested Accuracy Results
Predict Jop Post Type Details
Find Jop Post Type Prediction Ratio
Download Trained Data Sets
View Jop Post Type Prediction Ratio Results
View All Remote Users
Logout
WE'RE
HIRING
JOIN OUR TEAM

Job Post Type Prediction Found Ratio Details

Job Post Type	Ratio
Real	96.57184536834428
Fake	3.4281546316557256

Fig:6:Real &fake Ratios

VI.CONCLUSION

Now days job scam detection is the great concern all over the world. In this paper, we have analyzed impacts of job scam which could be a very prosperous area in the research field creating a lot of challenges to detect the fraudulent job posts. We experimented our proposed techniques using EMSCAD dataset, which contains real life fake job posts. In this paper, we have experimented with both machine learning algorithms (SVM, KNN, Naïve Bayes, Random Forest and MLP) and deep learning model (Deep Neural Network). This work depicts a comparative study on the evaluation of traditional machine learning and deep learning based classifiers. We have found the 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”, 2017.

[2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019,

[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”2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, 2020.

[5] Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by Violent Extremists”, Security Informatics, 2014,

[6] Y. Kim, “Convolutional neural networks for sentence classification,”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,” 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,”2016.

[9] C. Li, G. Zhan, and Z. Li, “News Text Classification Based on Improved BiLSTM-CNN,” in 2018

[10] K. R. Remya and J. S. Ramya, “Using weighted majority voting classifier combination for relation classification in biomedical texts,” 2014.

[11] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications

[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",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

[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. December 2014.

[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. 16-20 April 2012

[16] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of Fraudulent Emails by Employing Advanced Feature.