FAKE JOB DETECTION USING MACHINE LEARNING APPROACH

K. Swetha, M. Tharun Reddy, K. Sravani, B. Subramanyam

Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences (Autonomous), Rajampet, Andhra Prades , India-516126.

S. Ashok Kumar

Assistance Professor in Department of AI&DS , Annamacharya Institute of Technology and Sciences (Autonomous), Rajampet, Andhra Pradesh , India-516126.

Abstract

Advertising new job openings has recently become a very prevalent problem in the modern world as a result of advancements in social communication and modern technologies. Therefore, everyone will have a lot of reason to be concerned about bogus job postings. Fake job posing prediction presents a variety of difficulties, much as many other categorization problems. In order to determine whether a job posting is legitimate or fraudulent, this paper proposed using various data mining techniques and classification algorithms like KNN, decision tree, support vector machine, naive bayes classifier, random forest classifier, multilayer perceptron, and deep neural network. 18000 samples from the Employment Scam Aegean Dataset (EMSCAD) were used in our experiments. For this classification challenge, a deep neural network classifier excels. For this deep neural network classifier, three thick layers were employed. A bogus job advertisement may be predicted with a classification accuracy of about 98% by the trained classifier using DNN.

Index Terms—false job prediction, SVM, Logistic Regression, Random Forest, XG Boost

Introduction

Modern-day job searchers now have a wealth of fresh and varied work opportunities because to advancements in business and technology. Job searchers learn about their possibilities based on their availability, qualifications, experience, appropriateness, etc. with the aid of the advertisements for these job offers. The strength of the internet and social media now has an impact on the recruitment process. Social media has a huge influence on this since a recruiting process's ability to be successful depends on how well it is advertised. There are ever more opportunities to communicate employment details thanks to social media and electronic media marketing. Instead of this, the opportunity to share job postings quickly has increased the number of fraudulent job postings, which harass job seekers. People don't respond to new job postings because they want to keep their personal, academic, and professional information secure and consistent. The genuine goal of legitimate job advertisements via social and electronic media thus has a very difficult struggle to win over people's trust and trustworthiness. Technologies are all around us to improve and ease our lives, not to create unsafe working conditions. Recruiting new personnel will improve greatly if job postings can be correctly screened to identify fake job postings. False job postings make it difficult for job seekers to locate the positions they desire, which is a significant waste of their time. A fresh doorway for dealing with challenges in the field of human resource management is opened by an automated approach to predict fake job postings.

Background Study

A. Job Scam: Fake Job Posting

The term "job scam" refers to online job advertising that are false and frequently eager to steal the personal and professional information of job searchers in lieu of providing them with suitable positions. Fraudulent individuals occasionally attempt to steal money from job seekers. More than 67% of those who hunt for employment online without being aware of bogus job postings or job scams are at high risk, according to a recent UK poll by ActionFraud [2]. Nearly 700,000 job searchers in the UK reported losing more than $500 000 as a result of work scams. The survey indicated an almost 300% growth in the UK during the previous two years [2]. Students and recent graduates are the main targets of fraudsters since they frequently want to obtain a stable job for which they are prepared to spend more money. Techniques for avoiding or protecting against cybercrime fall short because con artists regularly alter their methods of employment fraud.

B. Common types of Job Scam

Fraudsters produce phony job adverts in order to get other people's personal information, such as insurance data, bank details, income tax details, date of birth, and national id. Advance fee scams happen when con artists demand money while using justifications such administrative fees, information security check costs, management costs, etc. Sometimes con artists pose as employers and inquire about applicants' passport information, bank account information, driving records, etc. as a pre-employment screening. When they get students to deposit money into their accounts and subsequently transfer it back, money laundering frauds take place [2]. This "cash in hand" strategy results in work with cash on hand without having to pay any taxes. In order to lure job searchers, scammers frequently develop bogus corporate websites, bank websites, official-looking papers, etc. Instead of engaging in face-to-face conversation, the majority of employment fraudsters attempt to capture victims via email. To establish themselves as headhunters or recruiting agency, they frequently use social networking sites like LinkedIn. They often work to give the job seeker the most accurate representation of their company profile or websites. Regardless of the employment scam they employ, they constantly try to lure job seekers into their traps by gathering information and using it to their advantage to either generate money or accomplish other goals. [6], [7].

C. Related Works

To determine if a job posting is authentic or false, several studies have been conducted. A significant amount of study is being done to identify employment fraud online. Job fraudsters were referred to be phony online job advertisers by Vidros [1] et al. They discovered statistics regarding several legitimate and well-known businesses and organizations that created false job adverts or vacancy listings with ulterior motives. They tested with a variety of classification techniques, including naive bayes classifier, random forest classifier, Zero R, One R, and others, on the EMSCAD dataset. The dataset's highest performance was displayed by the Random Forest Classifier, which had a classification accuracy of 89.5%. They discovered that the dataset had relatively low logistic regression performance. When they balanced the dataset and experimented on it, one R classifier worked well. They made an effort in their research to identify the issues with the ORF model (Online Recruitment Fraud) and to address those issues utilizing other dominant classifiers.

A methodology to identify fraud exposure in an online recruiting system was put out by Alghamdi [2] et al. On the EMSCAD dataset, they conducted experiments using a machine learning method. They worked on this dataset in three stages: feature selection, data pre-processing, and classifier-based fraud detection. In order to retain the general text pattern, they deleted noise and html tags from the data during the preparation stage. To effectively and efficiently limit the amount of characteristics, they used the feature selection approach. Support Vector Machine was used to determine the features, and a random forest ensemble classifier was utilized to identify bogus job postings from the test data. With the aid of the majority voting technique, the random forest classifier appeared to be a tree-structured classifier that operated as an ensemble classifier. With 97.4% classification accuracy, this classifier was able to identify bogus job postings.

Different deep neural network models, such as Text CNN, Bi-GRU-LSTM CNN, and BiGRU CNN, which are pre-trained using text datasets, have been suggested by Huynh [3] et al. They sought to categorize the dataset of IT jobs. They trained a TextCNN model with a convolution layer, a pooling layer, and a fully connected layer using data from IT jobs. This model used convolution and pooling layers to train data. The weights were flattened and then transferred to the layer with all connections. This model's classification method employed the softmax function. To improve classification accuracy, they also utilized an ensemble classifier (Bi-GRU CNN, Bi-GRULSTM CNN) utilizing a majority voting approach. They discovered that TextCNN had a classification accuracy of 66% and Bi-GRU-LSTM CNN had a classification accuracy of 70%. The ensemble classifier with an accuracy of delivered the best results for this classification challenge.

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 have employed a unique dataset of news or items shared on Twitter via the PolitiFact website account. To train the suggested GDU diffusive unit model, this dataset was utilized. This trained model performed well as an automated fake detecting model when input came from numerous sources at once.

To obtain high performance in the field of classifying bogus job posts, researchers experimented with a large variety of classifiers and feature selection techniques. Data pre-processing, feature selection using support vector machines, text processing using deep learning models, and other approaches were indicated as being applicable[8, [9], [10], [11], [12]. We have suggested using deep neural networks to anticipate employment frauds. Instead of using text data, we simply used the categorical characteristic of the EMSCAD dataset while applying the training approach. This method efficiently and quickly lowers the number of trainable attributes. We conducted a comparison research utilizing K Nearest Neighbor, Naive Bayes classifier, fuzzy KNN, decision tree, support vector machine, random forest classifier, and neural network on the same characteristics of the EMSCAD dataset.

Methodology

We have employed a variety of data mining techniques to determine whether a job posting is real or not. After pre-processing, we trained EMSCAD data in the classifiers. The developed classifier serves as a bogus job post detector for internet postings.

A. Neural Network

Neural networks operate on the fundamental tenets of how the human brain functions. It enables a computer to determine how much two patterns resemble or differ from one another by comparing them. A neuron is a mathematical function that extracts characteristics and categorizes particular patterns. There are several layers of interconnected nodes in a neural network. Each perceptron node functions as a multiple linear regression. The result of multiple linear regression is passed through this perceptron and converted into a non-linear activation function. Perceptrons are organized in layers that are linked to one another. To reduce mistake rates, the hidden layers change the weights of the input layers. The neural network functions as a classifier for supervised learning.

B. Deep Neural Network

Deep neural networks (DNNs) are Artificial Neural Networks (ANNs) that include numerous layers between the input and output layers. The feed forward algorithm powers DNN. Data is moved from the input layer to the output layer [13]. DNN generates a large number of virtual neurons that have their connection weights initialized with random numbers. This weight is multiplied by the input, and the result is an output that ranges from 0 to 1. To efficiently categorize the output, the training process modifies the weights. The model overfits as a result of learning unusual patterns from additional layers. Dropout layers allow for a generalization of the model by reducing the amount of trainable parameters.

C. Other classifiers

Decision Tree, Naive Bayes Classifier, K Nearest Neighbor, Random Forest Classifier, and Support Vector Machine .

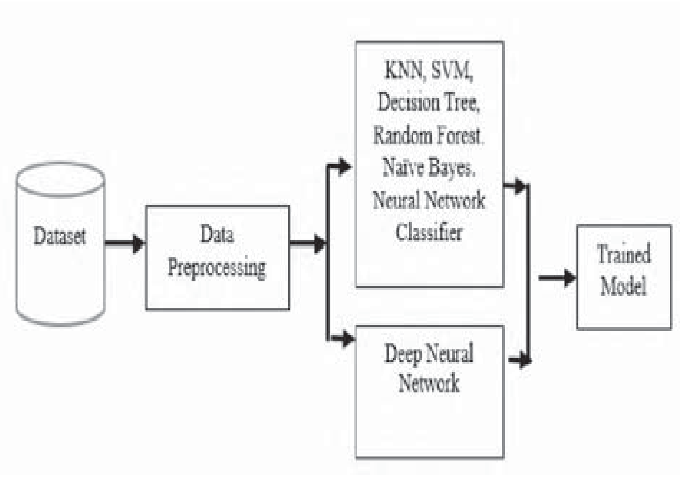

Dataset
Data Preprocessing
KNN, SVM, Decision Tree, Random Forest, Naïve Bayes, Neural Network Classifier
Deep Neural Network
Trained Model

Fig. 1. Proposed Methodology

Our research dataset is trained for the classifiers Multilayer Perceptron (MLP).

D. Dataset

EMSCAD has been used to identify bogus job postings. Each row of the data in this dataset has 18 characteristics, including the class label, and there are 18000 samples in total. The properties include employment type, required experience, required education, industry, function, salary range, company profile, description, requirements, benefits, telecommunication, has company logo, has questions, job id, title, location, department, and fraudulent. Only 7 of these 18 traits, which are transformed into category attributes, have been used.Telecommuting, has company logo, has questions, employment type, necessary education, required experience, and fraudulent are converted from text value to categorical value. As an illustration, the values for "employment type" are changed to 0 for "none," 1 for "full-time," 2 for "part-time," 3 for "others," 4 for "contract," and 5 for "temporary."

EXPERIMENTAL RESULT ANALYSIS

EMSCAD has been used to identify bogus job postings. Each row of the data in this dataset has 18 characteristics, including the class label, and there are 18000 samples in total. The properties include job id, title, department, location, pay range, company profile, requirements, benefits, and telecommunication.We used the EMSCAD dataset to implement the task in Google Colab. We have employed hold out cross validation for traditional machine learning algorithms like KNN, Random forest, SVM, etc. 20% of the total data was utilized for testing and evaluating the model's performance, while the remaining 80% was used for training. We used K values ranging from 1 to 40 in the KNN model, and the value of 13 produced the least amount of inaccuracy. During training, the average error rate was under 0.05. In SVM, the RBF kernel and gamma value = 0.001 are both employed.

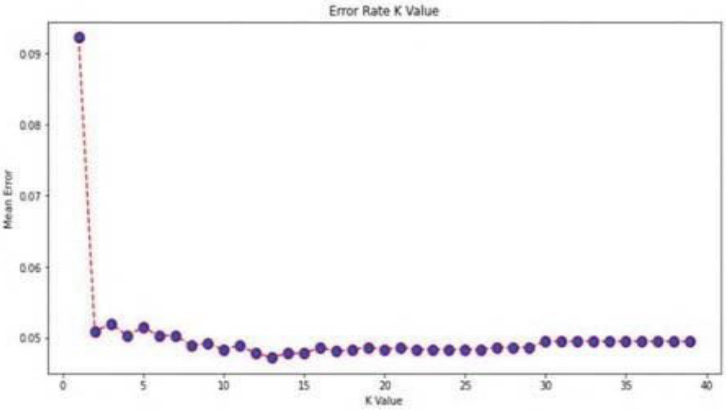
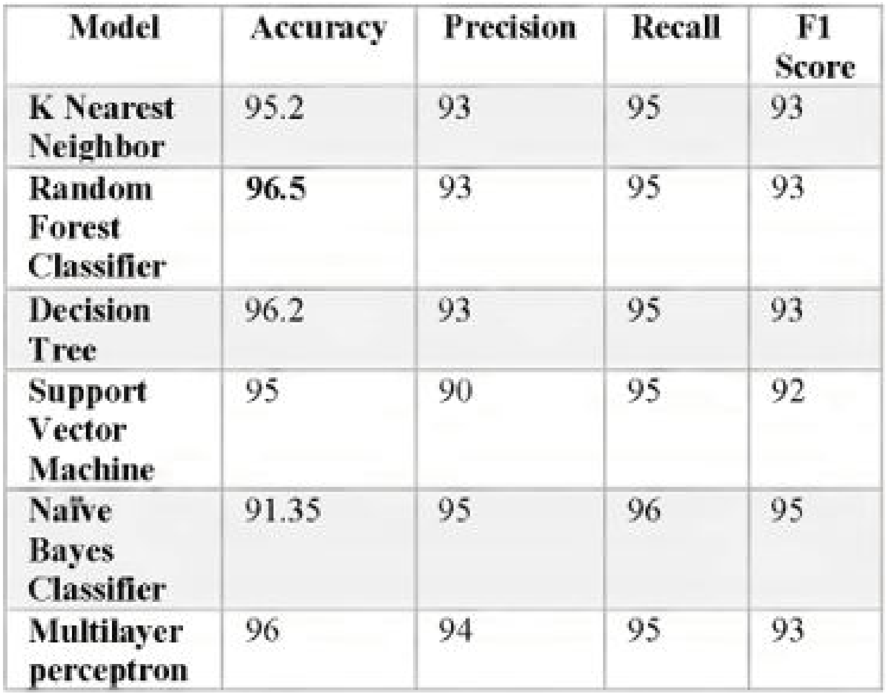


Fig 2: Relation between mean error and k value in KNN



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

TABLE 1 COMPARISON AMONG THE CLASSIFIERS

All of these classifiers' classification accuracy, precision, recall, and f1 score are displayed in Table I. With the Random Forest classifier, we have a classification accuracy of 97% or higher. In order to determine if the model performs well with both false positive and false negative samples, we additionally examined the f1 score. Below are the formulae for the measured parameters:

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)

In order to train the data for the deep neural network model, 10 fold cross validation is employed. A total of 60% of the data was utilized for training, 20% for determining validation accuracy, and the remaining 20% for testing the model's effectiveness. The validity accuracy reveals the model's level of performance with respect to unobserved data.In each training period, we have seen a positive correlation between validation and training accuracy. It is possible to identify the trained model as a generalized one if the validation accuracy is higher than the training accuracy. We utilized a dropout layer to lessen the model's overfitting. In order for the model to function effectively outside of the training dataset, this layer decreases the trainable parameters at each round of training.

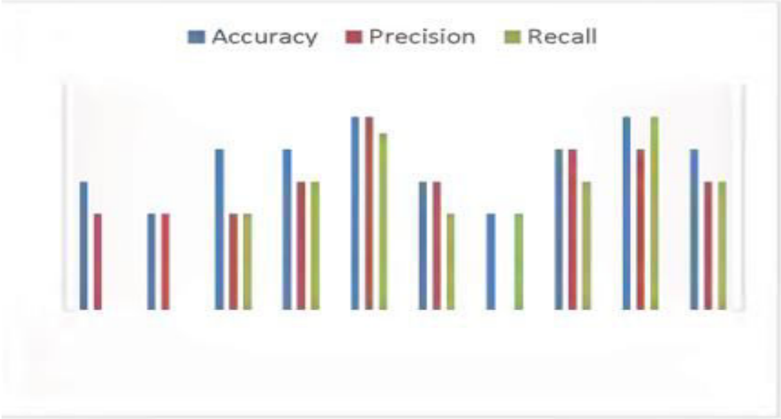


Fig. 3. Accuracy, Precision and Recall for 10 Folds in DNN model

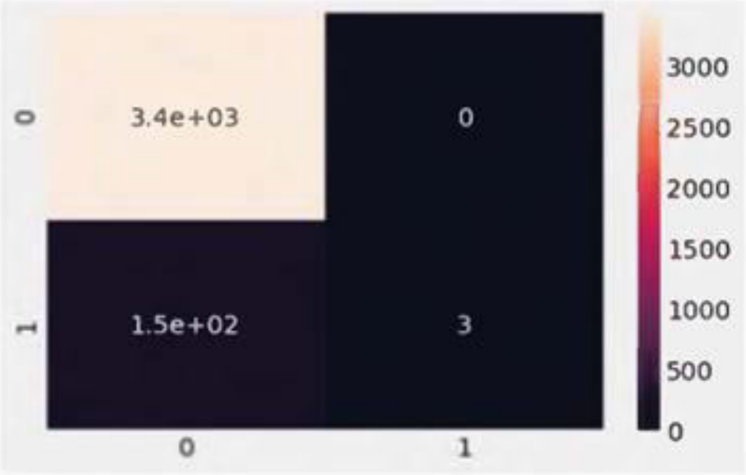


Fig. 4. Confusion matrix for DNN Model (Fold 2)

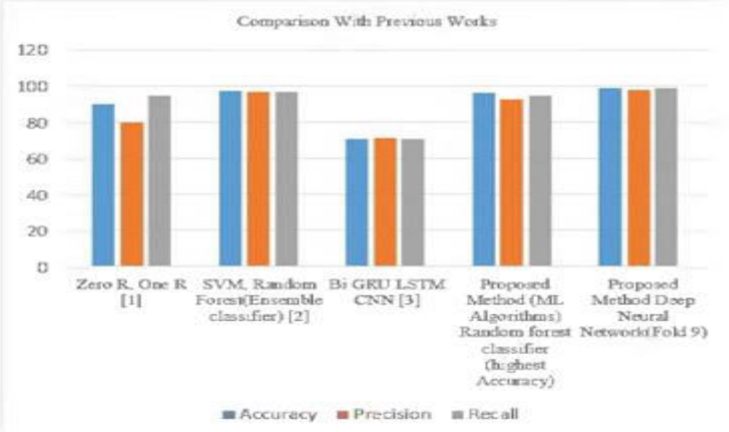


Fig. 5. Comparison of our proposed method with the previous works

Figure 3 demonstrates the recall, accuracy, and precision ofFigure 3 shows the recall, accuracy, and precision of each fold of the deep neural network model. The categorization accuracy for folds 2 and 7 was 96%, and folds 5 and 9 had the best accuracy at 99%. The trained deep neural network model has an average classification accuracy of 97.7%. Only accuracy can't gauge a generalized model's performance because we used a class- unbalanced dataset.The trained model's accuracy and recall metrics are also favorable. The confusion matrix for the DNN model (fold 2) is shown in Figure 4. The vast majority of the test data are arranged diagonally. Figure 5 shows a comparison between our suggested technique and earlier efforts. Both traditional machine learning methods and a deep learning model have been implemented. In the first instance, we were able to get the maximum classification accuracy (96.7%) using a random forest classifier, and we were able to achieve 99% accuracy for fold 9 using a deep learning model (DNN). The DNN model's average classification accuracy (10 fold) is 97.7%.

Conclusion

The identification of job scams has recently become a major problem worldwide. We have examined the effects of employment scams in this paper since they might be a very lucrative topic of study and make it difficult to identify fake job postings. We experimented with the EMSCAD dataset, which contains actual fake job postings. In this research, we experiment with both deep learning (Deep Neural Network) and machine learning (SVM, KNN, Naive Bayes, Random Forest, and MLP). This article presents a comparison study on the assessment of classifiers based on deep learning and conventional machine learning. In comparison to other conventional machine learning methods, Random Forest Classifier has the greatest classification accuracy. DNN (fold 9) and Deep Neural Network have the highest classification accuracy on average.

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] 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/imsa.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 LuuThuy 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 LuuThuy 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