ONLINE FAKE JOB ADVERTISEMENT RECOGNITION AND CLASSIFICATION USING MACHINE LEARNING

Gasim Othman Alandjani
Computer Science and Engineering Department. Yanbu University College, (Saudi Arabia).
E-mail: alandjanig@rcyci.edu.sa ORCID: https://orcid.org/0000-0003-0321-7013

Recepción: 21/10/2021 **Aceptación:** 20/12/2021 **Publicación:** 29/03/2022

Citación sugerida:
Alandjani, G. (2022). Online fake job advertisement recognition and classification using machine learning. *3C TIC. Cuadernos de desarrollo aplicados a las TIC, 11*(1), 251-267. https://doi.org/10.17993/3ctic.2022.111.251-267

ABSTRACT

Machine learning algorithms handle numerous forms of data in real-world intelligent systems. With the advancement in technology and rigorous use of social media platforms, many job seekers and recruiters are actively working online. However, due to data and privacy breaches, one can become the target of perilous activates. The agencies and fraudsters entice the job seekers by using numerous methods, sources coming from virtual job-supplying websites. We aim to reduce the quantity of such fake and fraudulent attempts by providing predictions using Machine Learning. In our proposed approach, multiple classification models are used for better detection. This paper also presents different classifiers' performance and compares results to enhance the results through various techniques for realistic results.

KEYWORDS

Machine learning, Supervised algorithms, Fake jobs detection, Classification.

1. INTRODUCTION

Every organization nowadays is the internet and social media dependent. Systems like enterprise applications, management information systems, Information systems for Human Resources, and office automation applications are pivotal for running work. Creating an effective workforce recruitment process is considered by employing online applications, as it is more convenient for applicants. The majority of the human asset specialists and associations empower the online application framework for the enlistment and choice cycle. It has many benefits. Candidates can apply without the time and transfer their educational program vitae for additional references. Managers additionally can channel the applications rapidly and make waitlists within a brief period.

In this way, electronic enrollment makes human resource capacities fast. It gives an ideal chance to online scammers to exploit their distress on these needy occasions when thousands and millions of individuals seek jobs. Over time, there is an expansion in these fake job posts where ads appear to be very ordinary, frequently these organizations will likewise have a site and will have an enlistment interaction like different firms in the area (Ward, Gbadebo, & Baruah, 2015).

Online Recruitment Fraud (ORF) is becoming a severe issue in recent times. Due to hype in social media, online job advertisements are growing rapidly, but with advantages, there are many scammers, fraud employers scam them for money or taking personal information. Deceitful jobs ads can be posted using a well-known organization for disregarding their validity (Ward *et al.*, 2015). Detecting fake job posts has taken consideration for acquiring an automatic tool, recognizing fake ads positions, and revealing them to individuals to stay away from the application for such positions.

2. RELATED WORK

All Fraud jobs advertisements can be viewed as bogus data on the web and as a type of scam. Information on the internet can be false, which is divided into misinformation and disinformation. If information is

https://doi.org/10.17993/3ctic.2022.111.251-267

falsely created by misunderstanding or misconception, disinformation is purposefully made to cheat per user (Kumar & Shah, 2018). Fake job ads are considered disinformation. Supervised and unsupervised learning solves disinformation-related problems such as fake news and reviews.

Bondielli and Marcelloni (2019) suggested two approaches in their paper. The first approach uses fact-checking websites for source information validation, it is named knowledge-based detection and the second approach uses the key attributes and extraction of essential features from source information.

Fake or bogus news datasets are created manually based on multiple resources that are:

- Creation of Fact-checking websites such as FakeNewsNet Dataset (Murtagh, 1991).
- Using document samples labeled dataset in Burfoot Satire News Dataset by Burfoot and Baldwin
- Credbank Dataset by Mitra and Gilbert (2015) approach by dataset gathering by using expert judgment.

For classification, supervised and unsupervised, both algorithms can work. Random forest agave learning-based approach where each classifier comprehends numerous tree-like classifiers applied to various examples, and each tree votes in favor of the most fitting class. Another helpful technique can be boosting, which can work with multiple classifiers for a single classifier to improve classification results. Extended innovation applies an algorithm for classifying the weighted adaptations of training data and chooses the grouping of the more significant voting classifier. AdaBoost illustrates a procedure of boosting, which delivers better effectiveness (Murtagh, 1991). Expanding algorithms implies tackling issues with spam filtration viably. In addition, Gradient boosting is an extra boosting procedure for a Classifier dependent on the decision tree rule (Prentzas *et al.*, 2019). It likewise limits the deficiency of accuracy.

Algorithms approaches that can distinguish fake advertisements in online media are the decision forest. Models of a quick, controlled ensemble. The decision tree can be the best model assuming the need to

https://doi.org/10.17993/3ctic.2022.111.251-267

anticipate a target for up to two tests. It is suggested to train and test different models by utilizing the Tune Model Hyperparameters system. Alghamdi and Alharby (2019) provided a model for detecting scam posts in online job ads systems. The authors had used the EMSCAD dataset on various machine-learning algorithms.

The methodology is divided into 3 steps preprocessing, selection of features, and identifying scams by the classifier:

- In step, one unwanted noise and tags are removed from the data and bringing into general text.
- To reduce extraversion features that are not in use selective features are selected using a support vector machine and random forest classifier.
- It is reported that the detect fake job posts classification accuracy showed 97.4%.

Rathi and Pareek (2013) implemented various data mining techniques to detect spam mail in conjunction with analyzing various data mining approaches on the spam dataset to search for the best classifier for email characterization. Support vector machine was utilized to classify and investigate data. A Naïve Bayes classifier was utilized to locate a specific feature of a class that was irrelevant to the existence of some other feature, analyze and clean data by breaking down the information, and eliminate immaterial and repetitive features from the data feature selection methods were used. The outcomes showed that well exactness of the classifier Random Tree is 99.715% (Rathi & Pareek, 2013).

Van Huynh *et al.* (2020) put forward a method in which authors used deep neural networks retrained models with text datasets. The classification was done on IT-related jobs. Models were text CNN, BiGRU CNN, and Bi-GRU-LSTM CNN. The TextCNN model is fully connected and contains layers of convolution and pooling (Mujtaba *et al.*, 2021; Mujtaba & Ryu, 2020). The training was done using layers (convolution and pooling). Softmax function was used in this model for classification with that ensemble classifier was used to get more accuracy. Reported accuracy was 66% from text CNN. Bi-GRU- LSTM CNN 70% accuracy

https://doi.org/10.17993/3ctic.2022.111.251-267

Zhang, Dong, and Philip (2020) presented a model, an automatic fake detector. Utilizing text processing separates good and false news, containing articles and subjects. They had gathered a custom dataset of information or articles using the Twitter account PolitiFact site. For the proposed GDU diffusive unit model custom dataset was used to train. As there are multiple sources of information simultaneously, this prepared model has worked well.

3. METHODOLOGY

3.1. RESEARCH QUESTION

What is the best suitable classification algorithm for detecting Fake job advertisements?

What are the appropriate and important features for fraudulent job detection?

This research aims at constructing a suitable model to detect fraudulent job advertisements, to protect the expatriates from falling into the trap. This research falls under the category of an empirical study that would be based on observation, testing, evaluations, and comparison of the applied algorithms.

3.2. PROPOSED APPROACH

The research understudy can be described as a three-tier approach starting with the dataset preprocessing, feature selection, and classifying by applying different machine learning models and evaluating them. Let us look at the research that has already been done in this field of detecting fraudulent advertisements or detection of spam emails etc., over a period. It is observed that many researchers have applied several classification algorithms, including SVM, NB, MLP, KNN, ID3, J48, decision tree, etc., among which SVM outperformed in many cases (Mitra & Gilbert, 2015). Considering this performance of SVM as a parameter to be validated, this research focuses on applying SVM, multinomial NB, decision tree, random forest, and K-nearest neighbor on the dataset and comparing their results (Islam *et al.*, 2020).

https://doi.org/10.17993/3ctic.2022.111.251-267

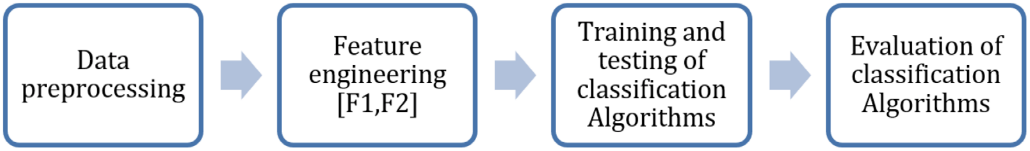

Data preprocessing
Feature engineering [F1,F2]
Training and testing of classification Algorithms
Evaluation of classification Algorithms

Figure 1. Work Methodology.
Source: own elaboration.

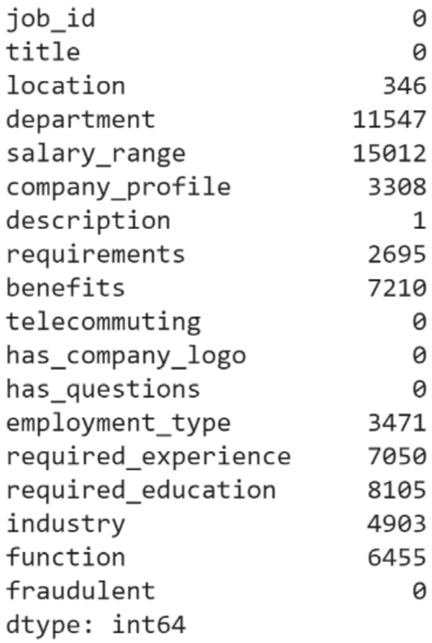
3.2.1. DATASET

This research works on a dataset from Kaggle to categorize a job advertisement as fraudulent or not based on some attributes derived from the advertisements available on different sources. The data was available in a CSV file having 17880 instances of jobs advertisements. Each advertisement is defined in terms of attributes on which we are working, that data is then preprocessed and classified through several algorithms. As the dataset has many missing values and anomalies, it needs a preprocessing step before it can be used as an input to any classification algorithm.

3.2.2. PREPROCESSING DATASET

The initial dataset had 17 attributes based on which this model would be predicting the status of an advertisement. These 17 attributes include job id, title, location, department, salary range, company profile, description, requirements, benefits, and telecommuting, has the company logo, has questions, employment type, required experience, required education, industry, and function. Each attribute contains either object or integer data. The label is binary for the specific problem domain, i.e., 0 for non-fraudulent and 1 for fraudulent.

The preprocessing phase starts after analyzing the dataset for missing values and some basic statistical operations on the integer data. Our integer fields include job id, telecommuting, has the company logo, has questions, and the final label of being fraudulent or not. Figure 1 describes the number of missing values in each field; this description justifies the deletion of job IDs and salary range containing the maximum missing values. The integer fields were ten checked for the correlation, and Figure 2 depicts the correlation heat map.



```
job_id                    0
title                     0
location                346
department            11547
salary_range          15012
company_profile        3308
description               1
requirements           2695
benefits               7210
telecommuting             0
has_company_logo          0
has_questions             0
employment_type        3471
required_experience    7050
required_education     8105
industry               4903
function               6455
fraudulent                0
dtype: int64
```

Figure 2. Key attributes.
Source: own elaboration.

After doing the exploratory data analytics, the process calls for proper preprocessing, including removing the missing values and stop words, deleting the irrelevant attributes that can be observed from the correlation heat-map, and finally removing the extra space. Now, the dataset is ideal for transforming into categorical encoding to achieve a feature vector. This feature vector would then be the final and transformed input to the classifiers.

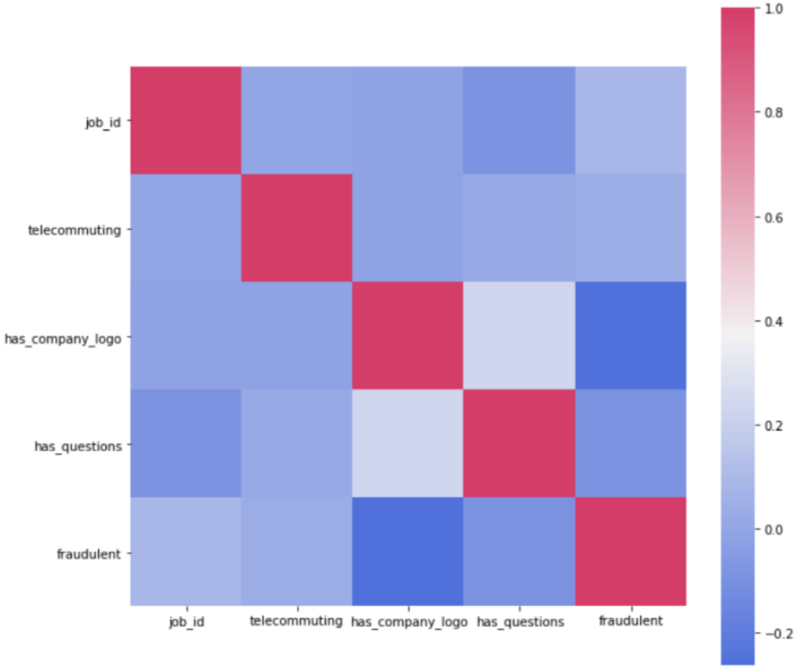


Figure 3. Attributes heat map.
Source: own elaboration.

3.2.3. IMPLEMENTATION OF CLASSIFIER

The proposed approach compares the performance of classifiers on two different feature sets. The first feature set includes the processed data discussed above and the second feature set has the integer attributes, benefits, and location. In this research, several classifiers are engaged, such as Naive Bayes Classifier, Decision Tree Classifier, K- Nearest Neighbor, and Random Forest Classifier, classifying job posts as fake. Note that 'fraudulent' is the target class for the research under discussion. Moreover, the feature sets on which the models are trained are mentioned in Table 1 below:

Table 1. Feature Sets.

Feature set 1	Feature set 2
title, location, department, company profile, description, requirements, benefits, telecommuting, has the company logo, has questions, employment type, required experience, required education, industry, and function	title, telecommuting, has the company logo, has questions, benefits

Source: own elaboration.

For both the feature sets, the classifiers are passed on to the training phase with 80 percent of the entire dataset, the remaining 20 percent would be used for the prediction phase. Training the classifiers for the proposed approach starts with choosing the right and tuned parameters as default parameters do not guarantee the best and promising results. After the prediction of the testing data, the model would be then evaluated on metrics such as Accuracy, F-measure, and Cohen- Kappa score. They are keeping the work on both the feature sets in parallel. The best classifier would be chosen to have outstanding performance among all the peer classifiers for each feature set.

3.2.4. EVALUATION METRICS

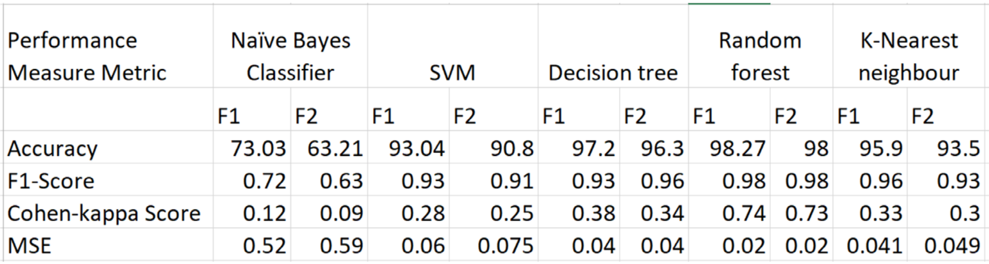
To evaluate the performance of any machine-learning model, evaluation metrics are used for this purpose. Given metrics are considered for evaluating and identifying the subtle approach for solving a problem. Accuracy metric aims to identify the true cases (predictions) from overall numbers to cases given to test. Accuracy may not be the primary metric for checking the model's performance as false cases (prediction). If a false result is taken as true, one will become problematic. It is important to consider false positive and false negative cases to requite the wrong classification. Precision checks the ratio of the right identified positive case from the total positive results given by the classifier. Recall presents the correct results of positive cases divided by the number of cases relevant. F-measure is a metric, which is involved in precision and recall, calculation is done by the harmonic mean of precision and recall.

https://doi.org/10.17993/3ctic.2022.111.251-267

4. RESULTS AND DISCUSSION

After getting the predictions from all five classifiers discussed in this research, their performance is compared based on a couple of evaluation metrics to conclude the best classifier for predicting fraudulent job advertisements. Table 1 displays the comparative study of the classifiers concerning evaluating metrics for both feature sets.

Table 2. Comparative table of classifiers performance.



Performance Measure Metric	Naïve Bayes Classifier		SVM		Decision tree		Random forest		K-Nearest neighbour	
	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
Accuracy	73.03	63.21	93.04	90.8	97.2	96.3	98.27	98	95.9	93.5
F1-Score	0.72	0.63	0.93	0.91	0.93	0.96	0.98	0.98	0.96	0.93
Cohen-kappa Score	0.12	0.09	0.28	0.25	0.38	0.34	0.74	0.73	0.33	0.3
MSE	0.52	0.59	0.06	0.075	0.04	0.04	0.02	0.02	0.041	0.049

Source: own elaboration.

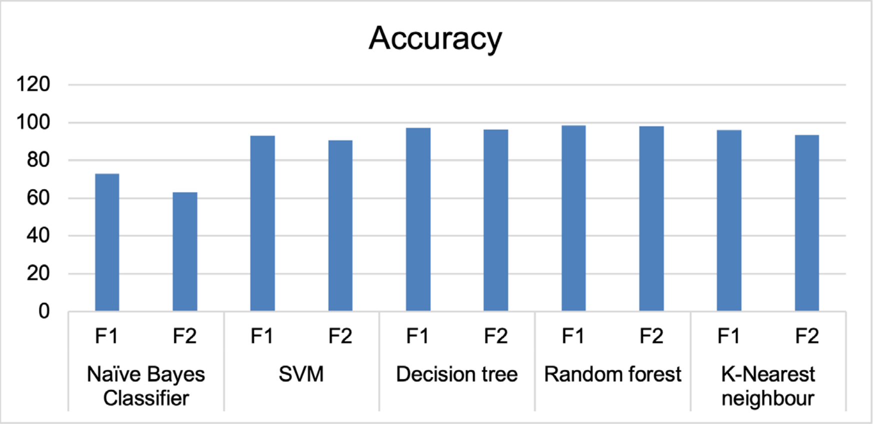


Figure 4. Accuracy Metric Comparisons of algorithms.

Source: own elaboration.

https://doi.org/10.17993/3ctic.2022.111.251-267

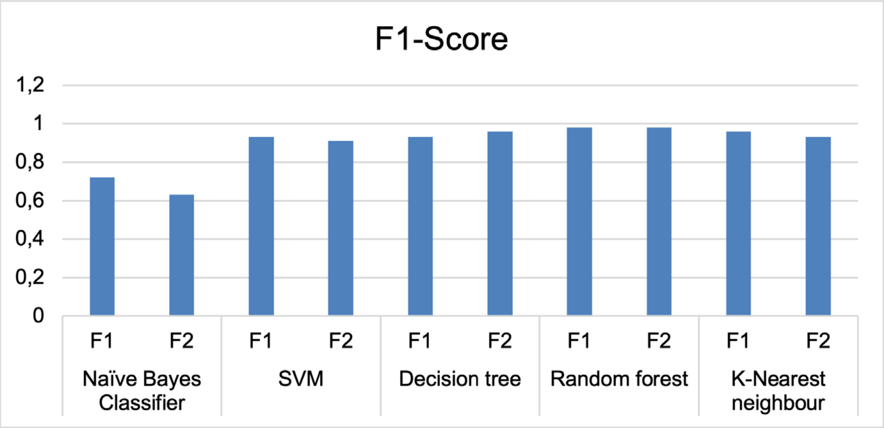


Figure 5. F1 Metric Comparison of algorithms.

Source: own elaboration.

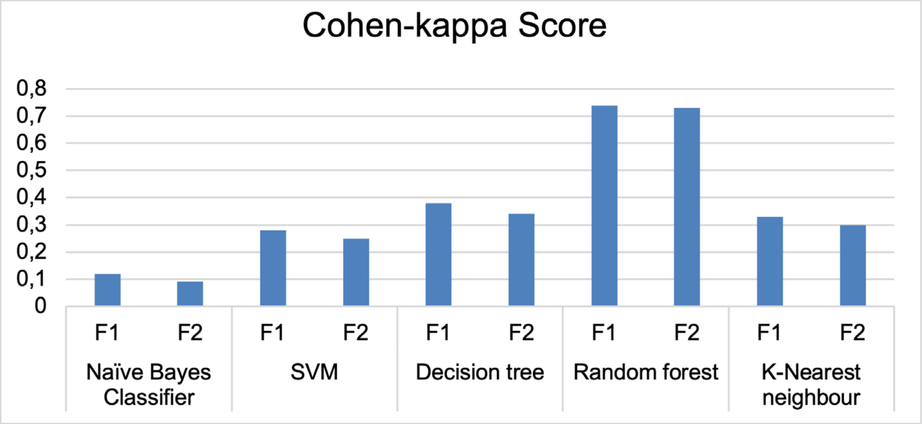


Figure 6. Cohen-Kappa Score of algorithms.

Source: own elaboration.

https://doi.org/10.17993/3ctic.2022.111.251-267

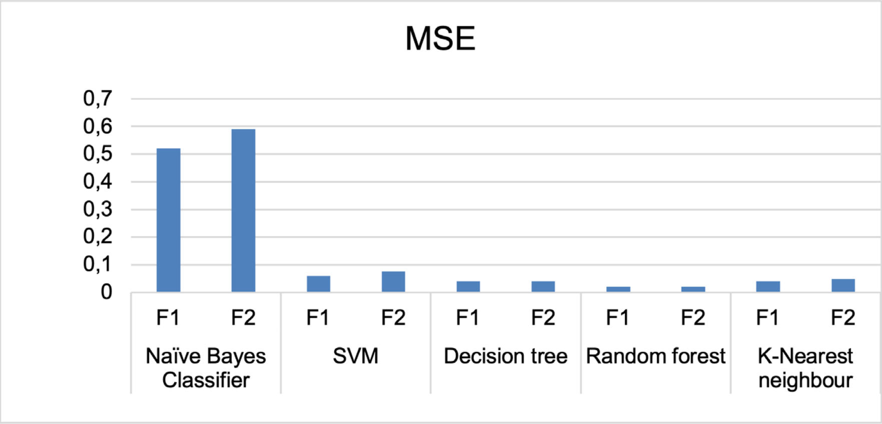


Figure 7. MSE calculations.

Source: own elaboration.

5. CONCLUSIONS

Platforms such as online job portals or social media for job advertisements are an exciting way of attracting potential candidates on which many enterprise companies are dependent on the hiring process. Fake jobs scam detection at an early stage can save a job seeker and make them only apply for legitimate companies. For this purpose, various machine learning techniques were utilized in this paper. Specifically, supervised learning algorithms classifiers were used for scam detection. This paper experimented with different algorithms such as naïve Bayes, SVM, decision tree, random forest, and K-Nearest Neighbor. It is reported that the K-NN classifier gives a promising result for the value k=5 considering all the evaluating metrics. On the other hand, Random Forest is built based on 500 estimators on which the boosting is terminated. In the future, the proposed method can be used for mobile devices using energy-efficient techniques (Mujtaba, Tahir, & Soomro, 2019; Mujtaba & Ryu, 2021).

https://doi.org/10.17993/3ctic.2022.111.251-267

REFERENCES

Alghamdi, B., & Alharby, F. (2019). An intelligent model for online recruitment fraud detection. *Journal of Information Security, 10*(03), 155. https://www.scirp.org/journal/paperinformation.aspx?paperid=93637

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences, 497*, 38-55. https://app.dimensions.ai/details/publication/pub.1114201506

Islam, M. R., Liu, S., Wang, X., & Xu, G. (2020). Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Social Network Analysis and Mining, 10*(1), 1-20. https://doi.org/10.1007/s13278-020-00696-x

Kumar, S., & Shah, N. (2018). *False information on web and social media: A survey*. arXiv preprint arXiv:1804.08559.

Mitra, T., & Gilbert, E. (2015). Credbank: A large-scale social media corpus with associated credibility annotations. In *Ninth international AAAI conference on web and social media*. https://ojs.aaai.org/index.php/ICWSM/article/view/14625

Mujtaba, G., & Ryu, E. S. (2020). Client-driven personalized trailer framework using thumbnail containers. *IEEE Access, 8,* 60417-60427. https://ieeexplore.ieee.org/document/9046852

Mujtaba, G., & Ryu, E. S. (2021). Human Character-oriented Animated GIF Generation Framework. In *2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC)* (pp. 1-6). IEEE.

Mujtaba, G., Lee, S., Kim, J., & Ryu, E. S. (2021). Client-driven animated GIF generation framework using an acoustic feature. *Multimedia Tools and Applications*, 1-18. https://link.springer.com/article/10.1007/s11042-020-10236-6

Mujtaba, G., Tahir, M., & Soomro, M. H. (2019). Energy efficient data encryption techniques in smartphones. *Wireless Personal Communications, 106*(4), 2023-2035. https://link.springer.com/article/10.1007/s11277-018-5920-1

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing, 2*(5-6), 183-197. https://pure.hud.ac.uk/en/publications/multilayer-perceptrons-for-classification-and-regression

Prentzas, N., Nicolaides, A., Kyriacou, E., Kakas, A., & Pattichis, C. (2019). Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction. In *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 817-821). IEEE. https://ieeexplore.ieee.org/document/8941679

Rathi, M., & Pareek, V. (2013). Spam mail detection through data mining-A comparative performance analysis. *International Journal of Modern Education and Computer Science, 5*(12), 31. https://www.mecs-press.org/ijmecs/ijmecs-v5-n12/v5n12-5.html

Van Huynh, T., Van Nguyen, K., Nguyen, N. L. T., & Nguyen, A. G. T. (2020). Job prediction: From deep neural network models to applications. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)* (pp. 1-6). IEEE. https://ieeexplore.ieee.org/document/9140760

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(1), 6. https://www.mdpi.com/1999-5903/9/1/6

Ward, A., Gbadebo, A., & Baruah, B. (2015). Using job advertisements to inform curricula design for the key global technical challenges. In *2015 International Conference on Information Technology Based Higher Education and Training (ITHET)* (pp. 1-6). IEEE. https://ieeexplore.ieee.org/abstract/document/7218042

Zhang, J., Dong, B., & Philip, S. Y. (2020). Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1826-1829). IEEE. https://arxiv.org/abs/1805.08751

https://doi.org/10.17993/3ctic.2022.111.251-267