2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

ARTIFICIAL INTELLIGENCE FOR FRAUDULENT JOB ADVERTISEMENT PREDICTION FROM EMSCAD

E. Babu[1], Nampelly Naresh[2], Muda Sai Kiran[2], Vadla Ruthuja[2], Mukkapally Sai Pavan[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Information Technology

[1,2]Malla Reddy College of Engineering and Management Sciences, Medchal, Hyderabad

ABSTRACT

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 advertisements of these job offers, job seekers find out their options depending on their time, qualification, experience, suitability etc. Therefore, this project proposed to use different data mining techniques and classification algorithm like K-nearest neighbour, decision tree, support vector machine, naive bayes classifier, random forest classifier, and multi-layer perceptron to predict a job Advertisement 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 ad.

Keywords: Fake job ads, Artificial intelligence, Deep neural networks, Employment Scam Aegean Dataset.

1. 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 advertisements 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 [1]. 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 lack 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 [2].

2. LITERATURE SURVEY

Habiba et. al [6] 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.

Amaar et. al [7] used six machine learning models to analyze whether these job ads are fraudulent or legitimate. Then, we compared all models with both BoW and TF-IDF features to analyze the classifier's overall performance. One of the challenges in this study is our used dataset. The ratio of real and fake job posts samples is unequal, which caused the model over-fitting on majority class data. To overcome this limitation, we used the adaptive synthetic sampling approach (ADASYN), which help to balance the ratio between target classes by generating the number of samples for minority class artificially. We performed two experiments, one with the balanced dataset and the other with the imbalanced data. Through experimental analysis, ETC achieved 99.9% accuracy by using ADASYN as over-sampling ad TF-IDF as feature extraction. Further, this study also performs an in-depth comparative analysis of our proposed approach with state-of-the-art deep learning models and other re-sampling techniques.

Mehboob et. al [8] handles the recruitment fraud/scam detection problem. Several important features of organization, job description and type of compensation are proposed and an effective recruitment fraud detection model is constructed using extreme gradient boosting method. It develops an algorithm that extracts required features from job ads and is tested using three examples. The features are further considered for two-step feature selection strategy. The findings show that features of the type of organization are most effective as a stand-alone model. The hybrid composition of selected 13 features demonstrated 97.94% accuracy and outperformed three state-of-the-art baselines. Moreover, the study finds that the most effective indicators are "salary_range," "company_profile," "organization_type," "required education" and "has multiple jobs." The findings highlight the number of research implications and provide new insights for detecting online recruitment fraud.

Ranparia et. al [9] minimized the number of such frauds by using Machine Learning to predict the chances of a job being fake so that the candidate can stay alert and take informed decisions, if required. The model will use NLP to analyze the sentiments and pattern in the job posting. The model will be trained as a Sequential Neural Network and using very popular GloVe algorithm. To understand the accuracy in real world, we will use trained model to predict jobs posted on Linked In. Then we worked on improving the model through various methods to make it robust and realistic.

Sudhakar et. al [10] proposed a novel algorithm for classifying phony information and actual news. This study deals with logistic regression, SVM, and novel ensemble approach based on machine learning algorithms. It is divided into sample size values of 620 per group. The experiment uses a dataset of 10,000 records with binary classes (fake news, real news). The result demonstrated that the proposed novel ensemble approach obtains a better accuracy value of 95% and a loss value of 05% compared with other algorithms. Thus, the obtained results prove that the proposed algorithm is an ensemble approach that combines decision tree techniques with AdaBoost by varying parameters and can get a significantly higher accuracy value.

3. PROPOSED SYSTEM

EMSCAD Dataset

The Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset containing 17,880 real-life job ads that aims at providing a clear picture of the Employment Scam problem to the research community and can act as a valuable testbed for scientists working on the field. To train the system, this project used EMSCAD dataset, where first row represents dataset column names and remaining

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

rows contains dataset values such as Company profile, job description, salary etc. In dataset last column contains 'fraudulent' values as 'f' for Fake and 't' for "True" jobs.

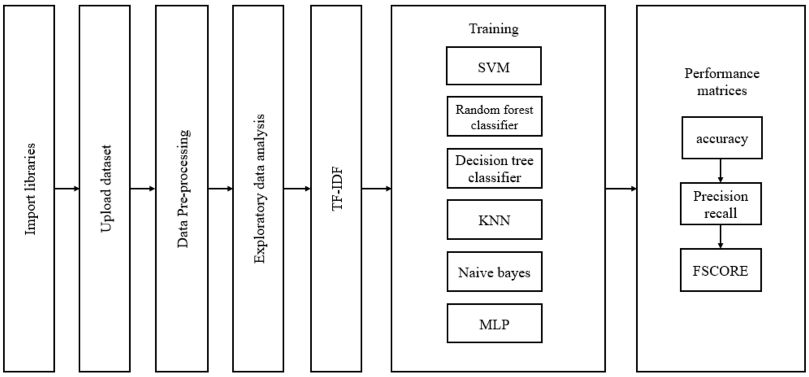


Fig. 1: Block diagram of proposed system.

TF-IDF Feature extraction

TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let's take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.

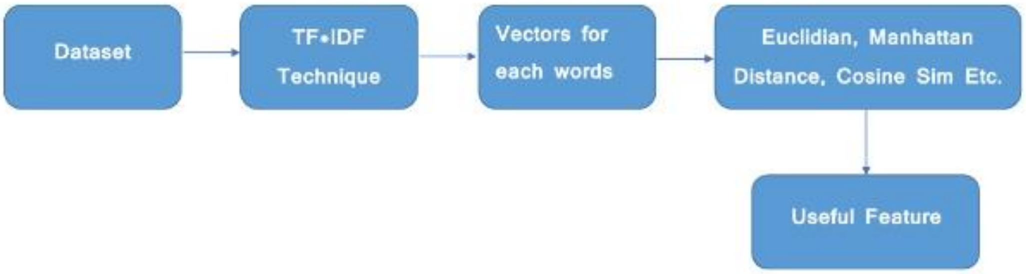


Fig. 2: TF-IDF block diagram.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

Term Frequency (TF): Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "Data Science is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Data" is", "Science", and "awesome", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. The weight of a term that occurs in a document is simply proportional to the term frequency.

$$tf(t,d) = count\ of\ t\ in\ d\ /\ number\ of\ words\ in\ d$$

Document Frequency*:* This measures the importance of document in whole set of corpora, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

$$df(t) = occurrence\ of\ t\ in\ documents$$

Inverse Document Frequency (IDF*):* While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. The IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

$$idf(t) = N/df$$

Now there are few other problems with the IDF, in case of a large corpus,say 100,000,000 , the IDF value explodes , to avoid the effect we take the log of idf . During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

$$idf(t) = log(N/(df + 1))$$

The TF-IDF now is at the right measure to evaluate how important a word is to a document in a collection or corpus. Here are many different variations of TF-IDF but for now let us concentrate on this basic version.

$$tf - idf(t,d) = tf(t,d) * log(N/(df + 1))$$

Multilayer perceptron (MLP)

MLP is one of the most frequently used neural network architectures in MDSS, and it belongs to the class of supervised neural networks. The multilayer perceptron consists of a network of nodes (processing elements) arranged in layers. A typical MLP network consists of three or more layers of processing nodes: an input layer that receives external inputs, one or more hidden layers, and an output layer which produces the classification results. Note that unlike other layers, no computation is involved in the input layer. The principle of the network is that when data are presented at the input layer, the network nodes perform calculations in the successive layers until an output value is obtained at each of the output nodes. This output signal should be able to indicate the appropriate class for the input data. That is, one can expect to have a high output value on the correct class node and low output values on all the rest. A node in MLP can be modelled as an artificial neuron (Fig. 2), which computes the weighted sum of the inputs at the presence of the bias, and passes this sum through the activation function. The whole process is defined as follows:

$$v_j = \sum_{i=1}^{p} w_{ji} x_i + \theta_j$$
$$y_j = f_j(v_j)$$

where vj is the linear combination of inputs x1; x2; xp, qj is the bias, wji is the connection weight between the input xi

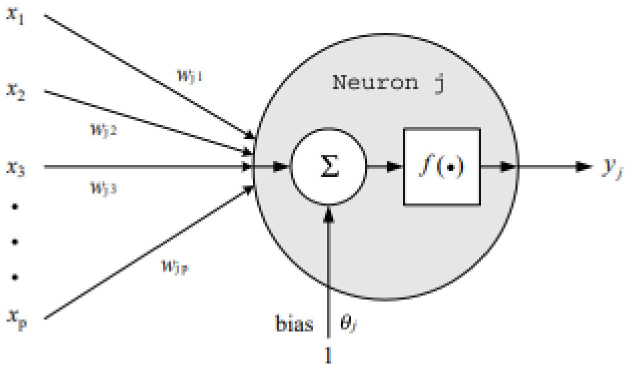


Fig. 3: One node of MLP: an artificial neuron.

and the neuron j, and fj ($) is the activation function of the jth neuron, and yj is the output.

The sigmoid function is a common choice of the activation function, as defined

$$f(a) = \frac{1}{1 + e^{-a}}$$

The bias term qj contributes to the left or right shift of the sigmoid activation function, depending on whether qj takes a positive or negative value. Once the architecture of MLP has been determined, the connection weights of the network have to be computed through a training procedure based on the training patterns and the desired output. BP is one of the simplest and most general methods for the supervised training of MLP. The basic BP algorithm works as follows:

- Initialize all the connection weights W with small random values from a pseudorandom sequence generator.
- Repeat until convergence (either when the error E is below a preset value or until the gradient vE(t)/vW is smaller than a preset value).
  - Compute the update using DWðtÞZKh vEðtÞ vW
  - Update the weights with WðtC1ÞZWðtÞC DWðtÞ
  - Compute the error E(tC1).

where t is the iteration number, is the connection weight, and h is the learning rate. The error E can be chosen as the mean square error (MSE) function between the actual output yj and the desired output dj:

$$E = \frac{1}{2}\sum_{j=1}^{n_j}(d_j - y_j)^2$$

There are two common training strategies: the incremental training strategy and the batch training strategy. Usually, an incremental strategy is more efficient and also faster for systems with large training samples, as random disturbances can be induced to help the system to escape from a local minimum point.

$$\varDelta W(t) = -\eta\frac{\partial E(t)}{\partial W} + \alpha\varDelta W(t-1)$$

where h0 is a preset learning rate, and lO0. The learning algorithm with forgetting mechanics is an algorithm that can ‘forget’ unused connections. With this forgetting mechanism, the weights that are not reinforced by learning will disappear. The obtained network, thus, has a skeletal structure that

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

reflects the regularity contained in the data, useful to improve the convergence and the network accuracy. In general, the updating of connection weights with forgetting mechanics term is given by:

$$\Delta W'(t) = \Delta W(t) - \varepsilon \operatorname{sgn}(W(t))$$

method. In this study, the conjugate gradients method is adopted, as it has a low computation cost and exhibits good results. The connection weights thus can be expressed by:

$$W(t+1) = W(t) + \eta(t)d(t)$$

$$d(t) = -\nabla E[W(t)] + \beta(t)d(t-1)$$

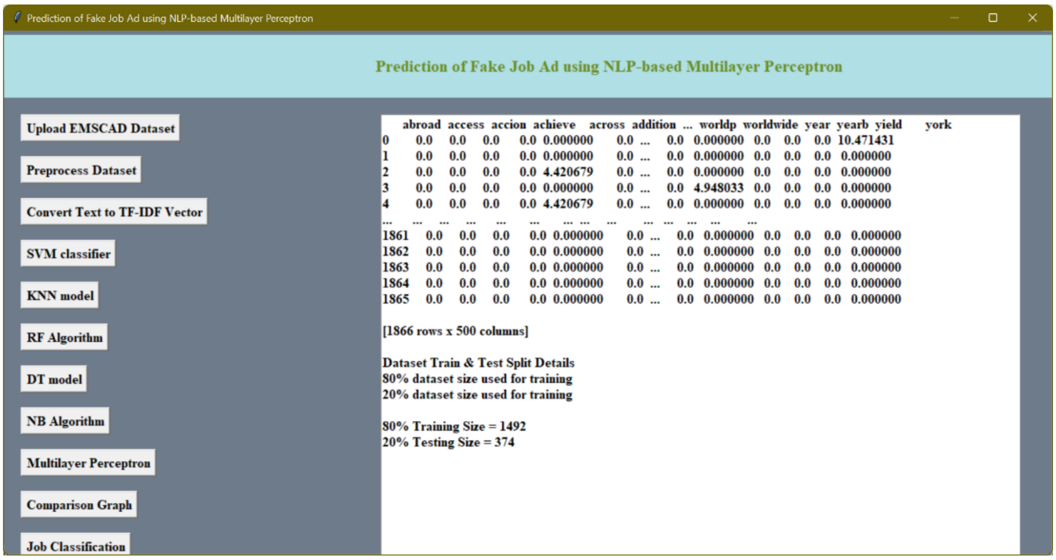
$$d(0) = -\nabla E[W(0)]$$

where PE is the gradient, d(t) is conjugate gradient, h(t) is the step wide, b(t) is determined given by Polak–Ribiere function

$$\beta(t) = \frac{[\nabla E(W(t)) - \nabla E(W(t-1))]^T \nabla E[W(t)]}{\nabla E[W(t-1)]^T \nabla E[W(t-1)]}$$

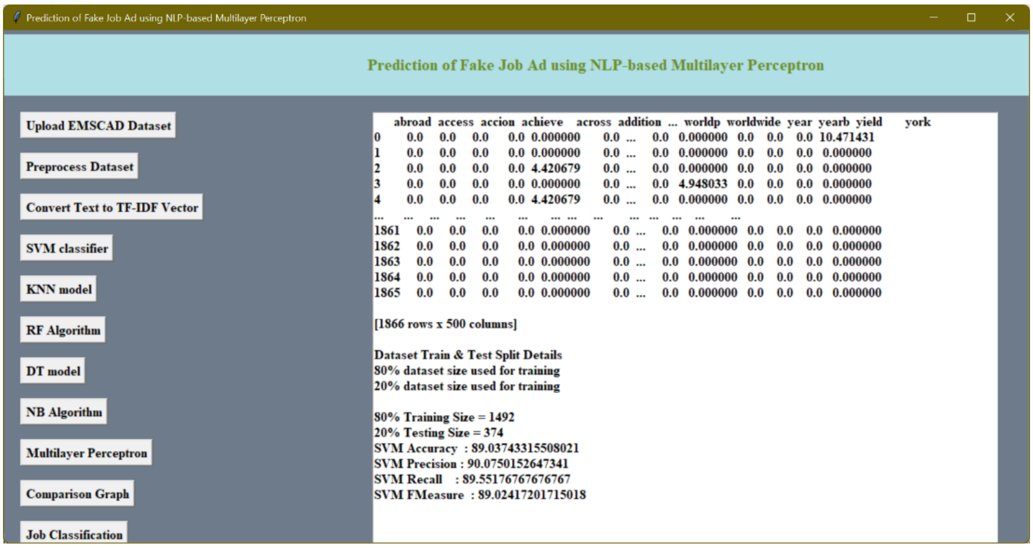
4. RESULTS AND DISCUSSION

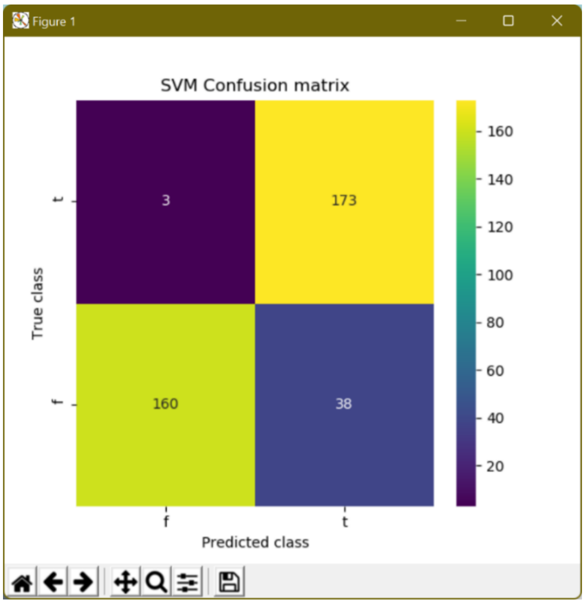
To train the existing and proposed models, this project has used 'Employment Scam Aegean Dataset (EMSCAD)' dataset, where first row represents dataset column names and remaining rows contains dataset values such as Company profile, job description, salary etc. In dataset last column contains 'fraudulent' values as 'f' for Fake and 't' for "True" jobs.



In above screen we can see the values are converted to vector where first row contains WORDS and remaining rows contains average frequency of those words. In above screen in last lines we can see dataset train and test split details.

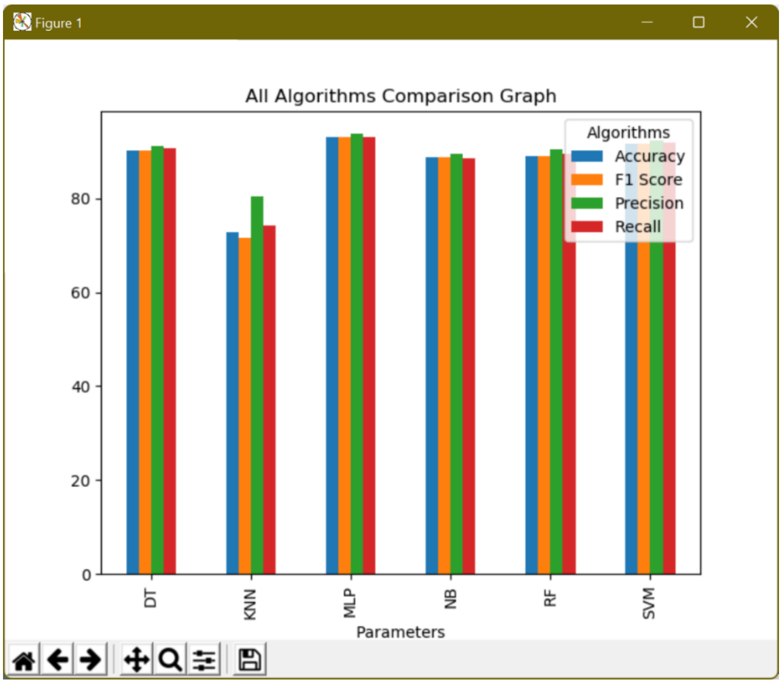
2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal



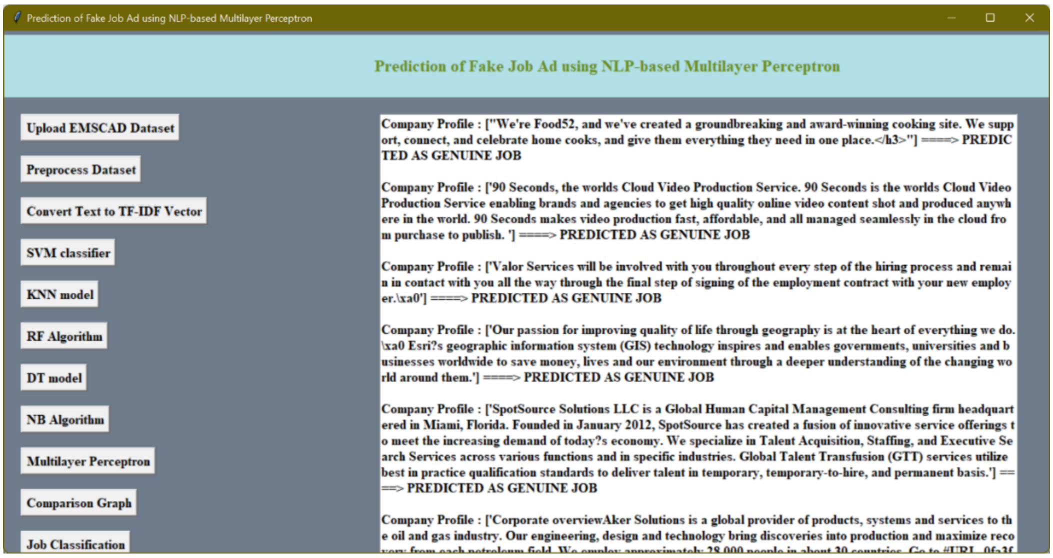


In above screen with SVM we got 89% accuracy, and we can see its confusion matrix graph. Similarly, we can run all the algorithms to get below output.

International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal
2581-4575



In above graph x-axis represents algorithms names and y-axis represents accuracy, precision, recall and FSCORE in different colour bars and in above graph we can see in all algorithms MLP got high accuracy and other values.



In above screen in square bracket, we are displaying JOB profile details and after square bracket and arrow symbol ==➔ we are displaying predicted output as GENUINE or FAKE.

5. CONCLUSION

Job scam detection has become a great concern all over the world at present. In this project, we have analyzed the impacts of job scam which can be a very prosperous area in research filed creating a lot of challenges to detect fraudulent job posts. We have experimented with EMSCAD dataset which contains real life fake job posts. In this paper, we have experimented both machine learning algorithms SVM, KNN, Naive Bayes, Random Forest and a neural network concept called MLP. This work shown the evaluation of machine learning and MLP-based classifiers.

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] B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection,” J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009

[6] S. U. Habiba, M. K. Islam and F. Tasnim, "A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 543-546, doi: 10.1109/ICREST51555.2021.9331230.

[7] Amaar, A., Aljedaani, W., Rustam, F. et al. Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches. Neural Process Lett 54, 2219–2247 (2022). https://doi.org/10.1007/s11063-021-10727-z

[8] Mehboob, A., Malik, M.S.I. Smart Fraud Detection Framework for Job Recruitments. Arab J Sci Eng 46, 3067–3078 (2021). https://doi.org/10.1007/s13369-020-04998-2

[9] D. Ranparia, S. Kumari and A. Sahani, "Fake Job Prediction using Sequential Network," 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS), 2020, pp. 339-343, doi: 10.1109/ICIIS51140.2020.9342738.

[10] Sudhakar, M., Kaliyamurthie, K.P. (2023). Efficient Prediction of Fake News Using Novel Ensemble Technique Based on Machine Learning Algorithm. In: Kaiser, M.S., Xie, J., Rathore, V.S. (eds) Information and Communication Technology for Competitive Strategies (ICTCS 2021). Lecture Notes in Networks and Systems, vol 401. Springer, Singapore. https://doi.org/10.1007/978-981-19-0098-3_1