JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 10, ISSUE 04, 2023

MACHINE LEARNING APPROACH FOR FAKE JOB AD PREDICTION

M. Premchender[1], A. Swathi[2], B. Sneha[2], B. Pranika[2], Ch. Meghana[2]

[1,2]Department of Information Technology

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Medchal, Telangana.

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. Recruitment process is now influenced by the power of internet and social media. Since the successful completion of a recruitment process is dependent on its advertisement, the impact of social media over this is tremendous. Social media and advertisements in electronic media have created newer and newer opportunity to share job details. Instead of this, rapid growth of opportunity to share job ads 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 ads can be filtered properly predicting false job ads, this will be a great advancement for recruiting new employees.

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, 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].

Fake Job Posting

Job Scam Online job advertisements which are fake and mostly willing to steal personal and professional information of job seekers instead of giving right jobs to them is known as job scam. Sometimes fraudulent people try to gather money illegally from job seekers. A recent survey by ActionFraud from UK has shown that more than 67% people are at great risk who look for jobs through online advertisements but unaware of fake job posts or job scam. In UK, almost 700000 job seekers complained to lose over $500000 being a victim of job scam. The report showed almost 300% increase over the last two years in UK. Students, fresh graduates are being mostly targeted by the frauds as they usually try to get a secured job for which they are willing to pay extra money. Cybercrime avoidance or protection techniques fail to decrease this offence since frauds change their way of job scam very frequently [3].

Common types of Job Scam

Fraudsters who want to gain other people’s personal information like insurance details, bank details, income tax details, date of birth, national id create fake job advertisements. Advance fee scams occur when frauds ask for money showing reasons like admin charges, information security checking cost, management cost etc. Sometimes fraudsters act them- selves as employers and ask people about passport details, bank statements, driving license etc. as pre-employment check. Illegal money mulling scams occur when they convince students to pay money into their accounts and then transfer it back [4]. This ’cash in hand’ technique causes to work cash in hand without paying any tax. Scammers usually create fake company websites, clone bank websites, clone official looking documents etc. to trap job seekers. Most of the job scammers try to trap people through email rather than face to face communication. They usually target social media like Linkedln to prove themselves as recruitment agencies or headhunters. They usually try to represent their company profile or websites to the job seeker as realistic as possible. Whatever the type of job scam they use, the always target the job seeker to fall in their trap, collecting information and making benefit either earning money or any other things.

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially. A classifier maps input variable to target classes by considering training data. Classifiers addressed in the paper for identifying fake job posts from the others are described briefly. These classifiers-based prediction may be broadly categorized into -Single Classifier based Prediction and Ensemble Classifiers based Prediction.

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 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.

Columns

- Title
- Location
- Department
- salary_range
- compamy_profile
- description
- requirements
- benefits
- telecommuting
- has_company_logo

Data Preprocessing in Machine learning

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

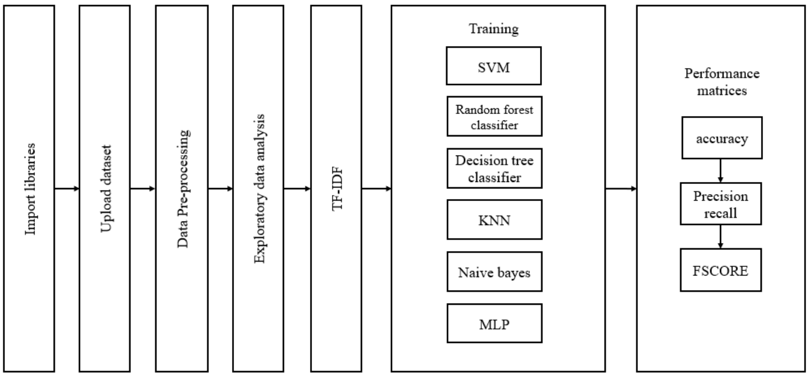


Fig. 1: Block diagram of proposed system.

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So, we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

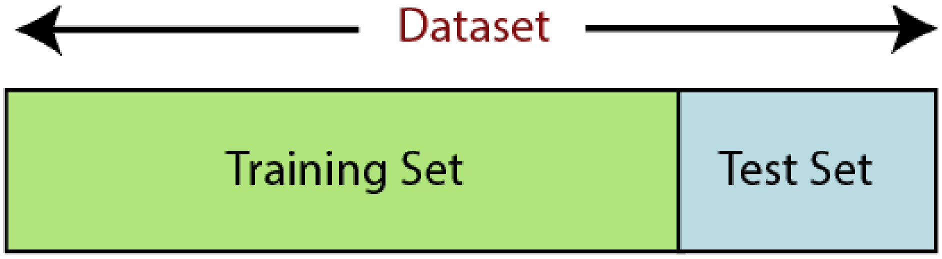


Fig. 2: Splitting of dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

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.

The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term

Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and the inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as tf * idf

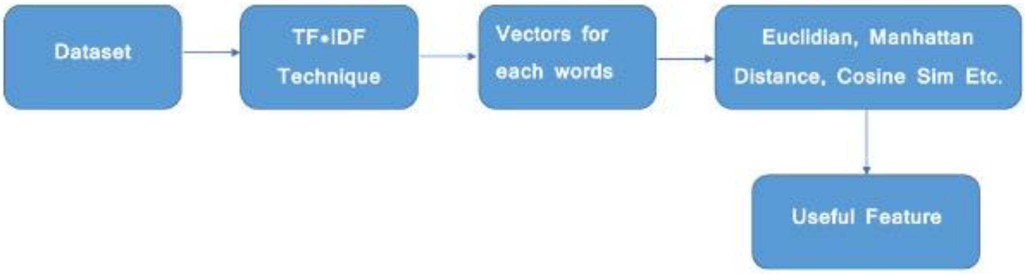


Fig. 3: TF-IDF block diagram.

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.

Terminology

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

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.

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

Implementing TF-IDF: To make TF-IDF from scratch in python, let’s imagine those two sentences from different document:

first_sentence: “Data Science is the sexiest job of the 21st century”.

second_sentence: “machine learning is the key for data science”.

First step we have to create the TF function to calculate total word frequency for all documents.

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

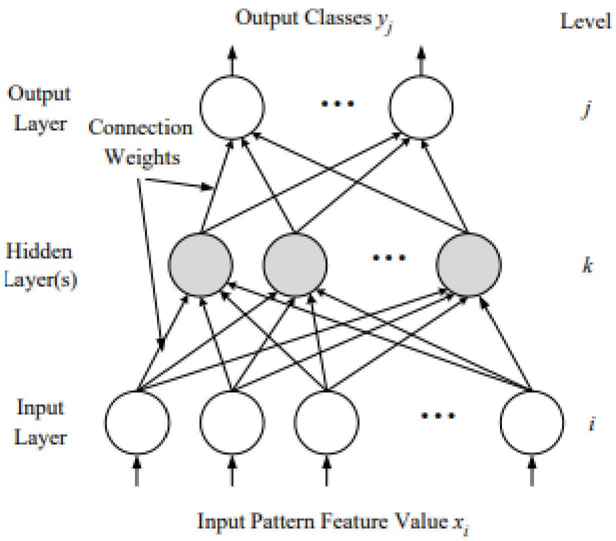


Fig. 4: Architecture of a multilayer perceptron network.

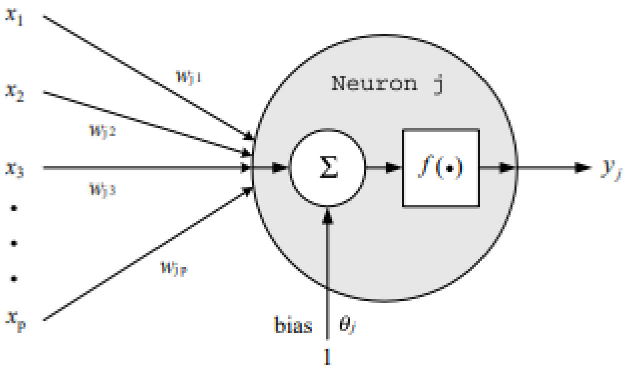


Fig. 5: One node of MLP: an artificial neuron.

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. The BP algorithm described above has some shortcomings. If the learning rate is set small enough to minimize the total error, the learning process will be slowed down. On the other hand, a larger learning rate may speed up learning process at the risk of potential oscillation. Another problem is that, partial minimal points or stable stages on error surface are often encountered during the learning process. Using a momentum term is the simplest method to avoid oscillation problems during the search for the minimum value on the error surface. The weight update in BP algorithm with a momentum term a is defined as follows:

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

where 0! a!1. The adaptive learning rate can also be adopted to speed up the convergence of the algorithm. For batch training strategy, the learning rate can be adjusted as follows

$$\eta(t) = \begin{cases} \beta\eta(t-1) & \mathit{if}\ E(t) < E(t-1) \\ \theta\eta(t-1) & \mathit{if}\ E(t) > kE(t-1) \\ \eta(t-1) & \mathit{otherwise} \end{cases}$$

where h(t) is the learning rate at the tth iteration, and b, q, and k are chosen as such that bO1, 0! q!1, and kO1. While for the incremental training strategy, learning rate can be updated using

$$\eta(t) = \eta_0 + \lambda E(t-1)$$

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 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:

$$\varDelta W'(t) = \varDelta W(t) - \varepsilon\ \mathrm{sgn}(W(t))$$

where 3 is the amount for the forgetting, and sgn(x) is the sign function (i.e., sgn(x)Z1 if xO0, K1 if x!0, and 0 if xZ0). The absolute value of connection weight is set to decrease by 3 due to the second term on the right-hand side. In practice, some optimization algorithms are often used to improve the network convergence, such as the steepest descent method, the Newton method, the Quasi–Newton method, and the conjugate gradients

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.

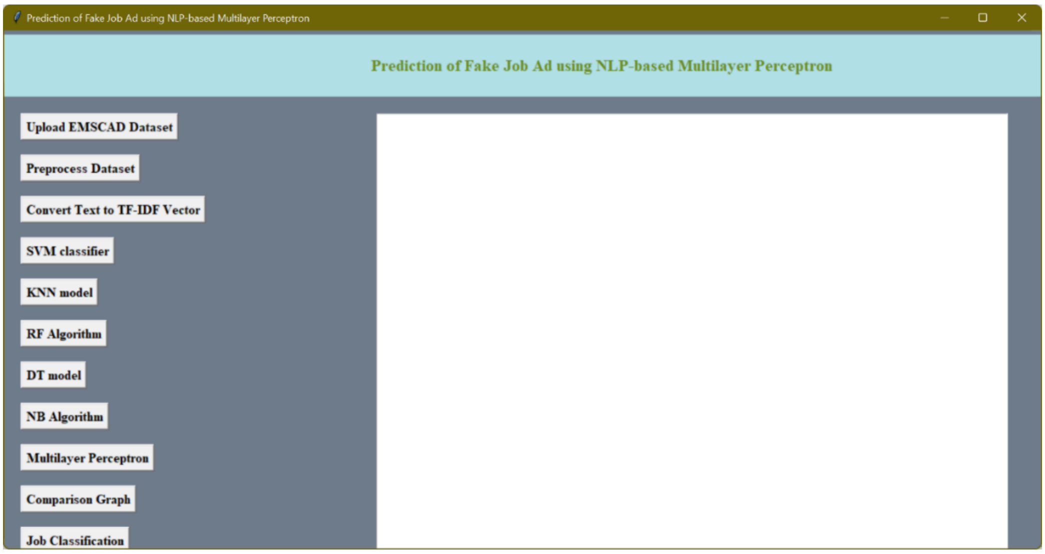
We have used above dataset to train all algorithms and evaluate their performance in terms of accuracy, precision, recall, FSCORE and confusion matrix graph

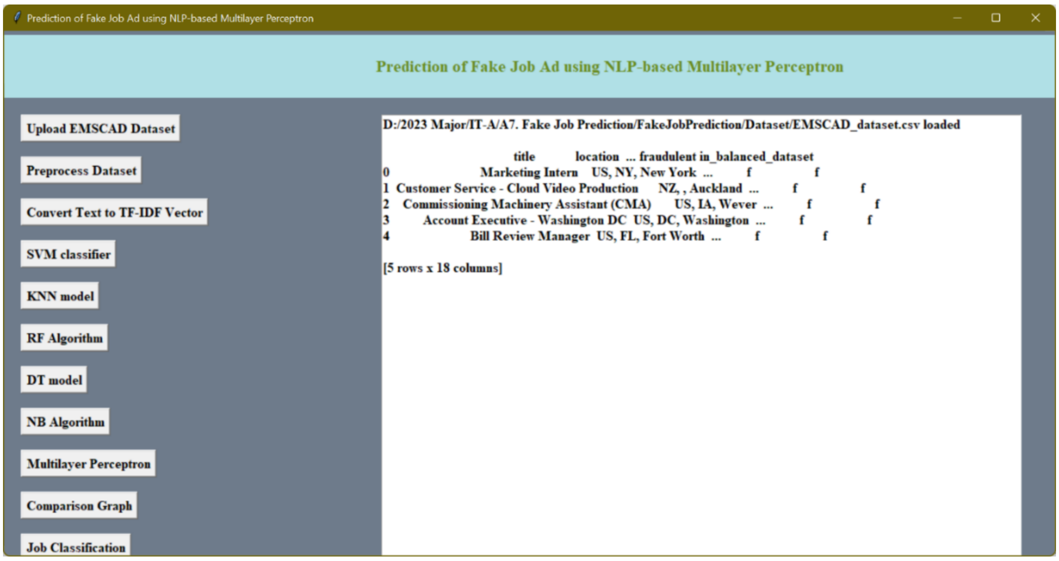
To implement this project, we have designed following modules

1) Upload EMSCAD Dataset: using this module we will upload ‘EMSCAD’ fake jobs Dataset
2) Preprocess Dataset: dataset often contains missing values and STOP WORDS so by using processing technique we will remove all missing values and stop words and make dataset values clean.
3) Convert Text to TF-IDF Vector: this dataset contains all non-numeric data and ML algorithms will not take such data so by applying TF-IDF (term frequency-Inverse document frequency) algorithm we will convert all words into numeric values by taking its average frequency of appearance in dataset
4) SVM classifier: using this module we will trained SVM algorithm
5) KNN model: using this module we will trained KNN algorithm
6) RF Algorithm: using this module we will trained Random Forest algorithm
7) DT model: using this module we will trained Decision Tree algorithm
8) NB Algorithm: using this module we will trained Naïve bayes algorithm
9) Multilayer Perceptron: using this module we will trained MLP algorithm
10) Comparison Graph: using this module we will plot accuracy comparison between all algorithms

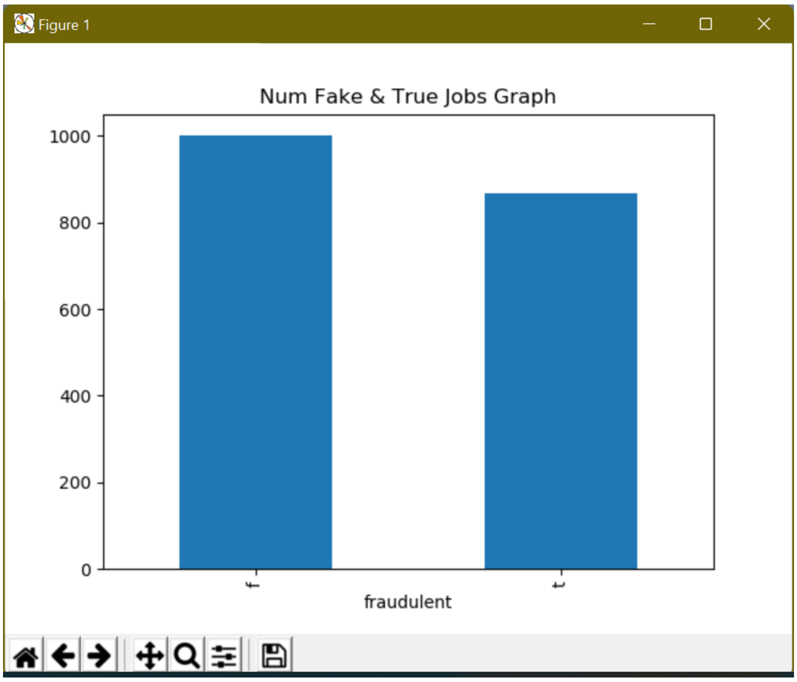
SCREEN SHOTS

In below screen click on ‘Upload EMSCAD Dataset’ button to upload dataset to application and then click on ‘Open’ button to load dataset and to get below screen

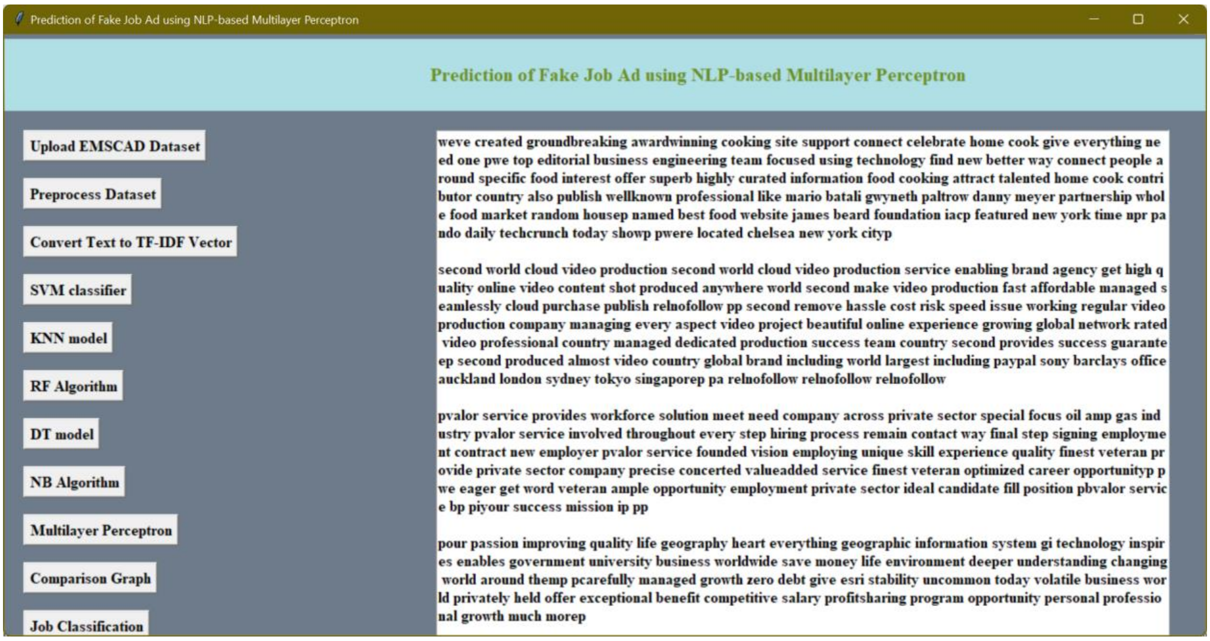

Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification


Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification

```
D:/2023 Major/IT-A/A7. Fake Job Prediction/FakeJobPrediction/Dataset/EMSCAD_dataset.csv loaded

                              title          location  ... fraudulent in_balanced_dataset
0                      Marketing Intern   US, NY, New York  ...          f                 f
1  Customer Service - Cloud Video Production     NZ, , Auckland  ...          f                 f
2   Commissioning Machinery Assistant (CMA)      US, IA, Wever  ...          f                 f
3         Account Executive - Washington DC  US, DC, Washington  ...          f                 f
4                   Bill Review Manager  US, FL, Fort Worth  ...          f                 f

[5 rows x 18 columns]
```

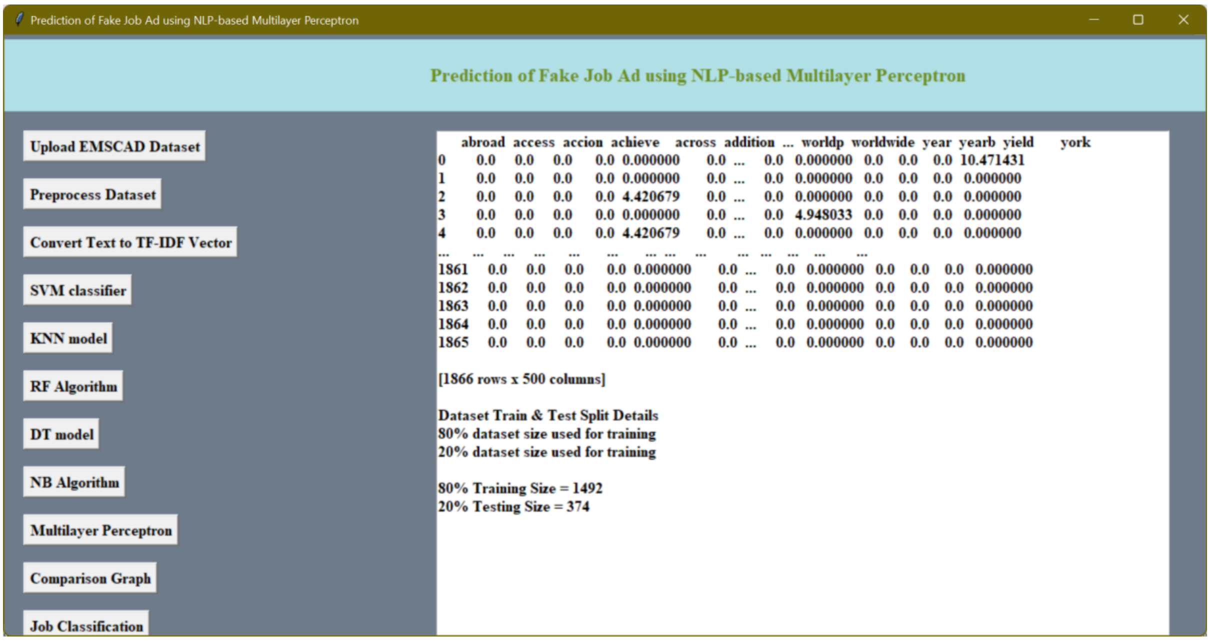


In above screen we can see dataset loaded and in graph x-axis represents 'f' and 't' and y-axis represent counts of fake and true jobs and now close above graph and then click on 'Preprocess Dataset' button to remove missing and stop words from dataset


Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification

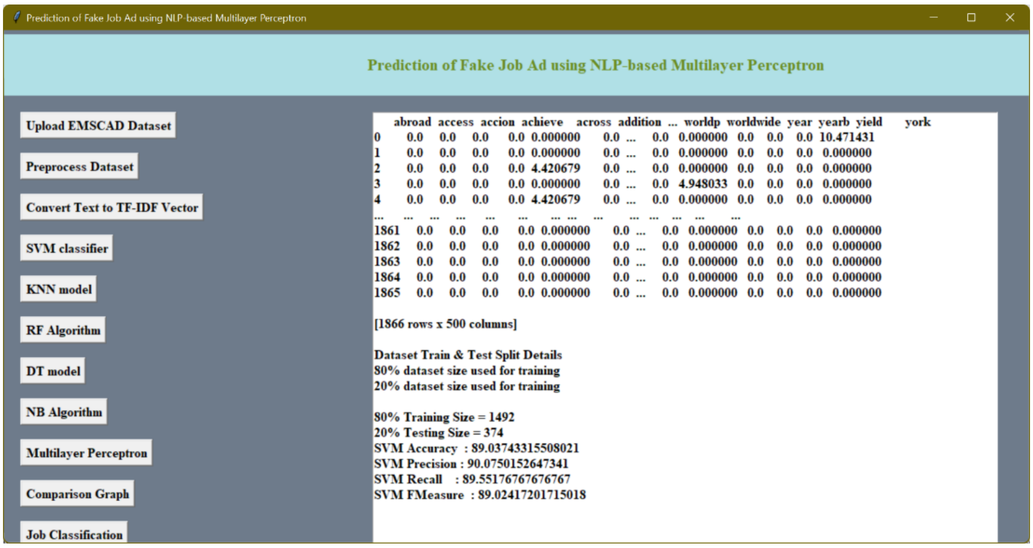
```
weve created groundbreaking awardwinning cooking site support connect celebrate home cook give everything ne
ed one pwe top editorial business engineering team focused using technology find new better way connect people a
round specific food interest offer superb highly curated information food cooking attract talented home cook contri
butor country also publish wellknown professional like mario batali gwyneth paltrow danny meyer partnership whol
e food market random housep named best food website james beard foundation iacp featured new york time npr pa
ndo daily techcrunch today showp pwere located chelsea new york cityp

second world cloud video production second world cloud video production service enabling brand agency get high q
uality online video content shot produced anywhere world second make video production fast affordable managed s
eamlessly cloud purchase publish relnofollow pp second remove hassle cost risk speed issue working regular video
production company managing every aspect video project beautiful online experience growing global network rated
 video professional country managed dedicated production success team country second provides success guarante
ep second produced almost video country global brand including world largest including paypal sony barclays office
auckland london sydney tokyo singaporep pa relnofollow relnofollow relnofollow

pvalor service provides workforce solution meet need company across private sector special focus oil amp gas ind
ustry pvalor service involved throughout every step hiring process remain contact way final step signing employme
nt contract new employer pvalor service founded vision employing unique skill experience quality finest veteran pr
ovide private sector company precise concerted valueadded service finest veteran optimized career opportunityp p
we eager get word veteran ample opportunity employment private sector ideal candidate fill position pbvalor servic
e bp piyour success mission ip pp

pour passion improving quality life geography heart everything geographic information system gi technology inspir
es enables government university business worldwide save money life environment deeper understanding changing
 world around themp pcarefully managed growth zero debt give esri stability uncommon today volatile business wor
ld privately held offer exceptional benefit competitive salary profitsharing program opportunity personal professio
nal growth much morep
```

In above screen from dataset TEXT all special symbols and missing values and stop words like 'and the this etc.' are removed out and now click on 'Convert Text to TF-IDF Vector' button to convert text to numeric vector like below screen

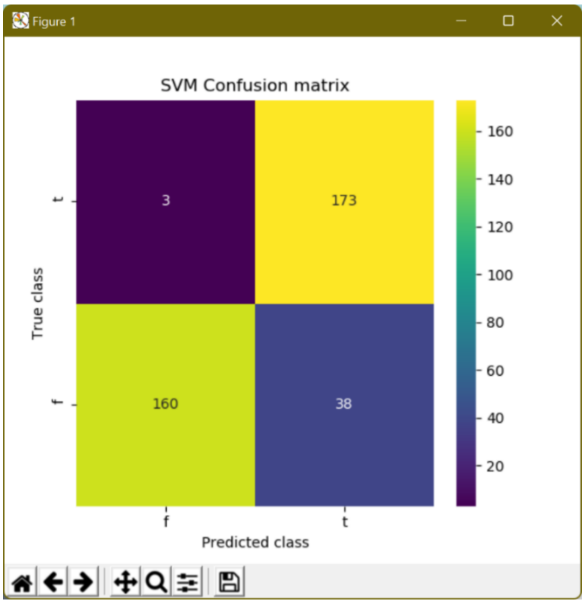

Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification

```
   abroad  access  accion  achieve   across  addition  ...  worldp  worldwide  year  yearb  yield      york
0     0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0  10.471431
1     0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
2     0.0     0.0     0.0      0.0  4.420679      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
3     0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   4.948033   0.0    0.0    0.0   0.000000
4     0.0     0.0     0.0      0.0  4.420679      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
...   ...     ...     ...      ...       ...      ...  ...     ...        ...   ...    ...    ...        ...
1861  0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
1862  0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
1863  0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
1864  0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000
1865  0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0   0.000000   0.0    0.0    0.0   0.000000

[1866 rows x 500 columns]

Dataset Train & Test Split Details
80% dataset size used for training
20% dataset size used for training

80% Training Size = 1492
20% Testing Size = 374
```

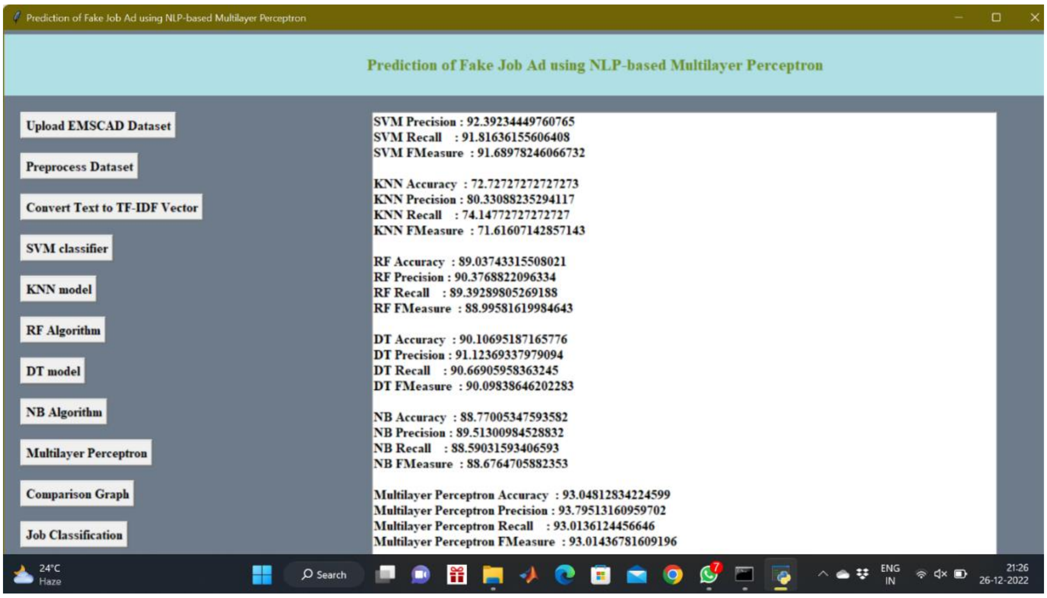
In above screen we can see the values are converted to vector where first row contains WORDS and remaining rows contains average frequency of those words. In above screen in last lines we can see dataset train and test split details.


Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification

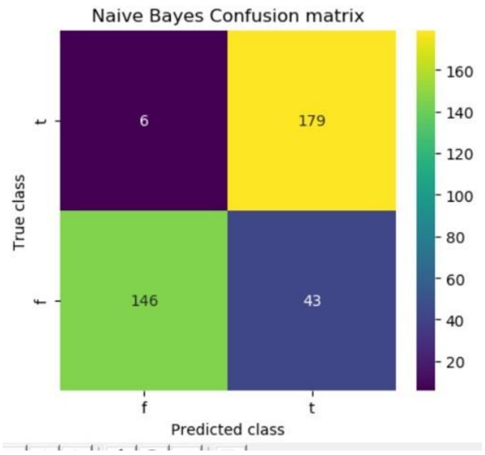
```
     abroad  access  accion  achieve   across  addition  ...  worldp  worldwide  year  yearb  yield      york
0       0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0  10.471431
1       0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
2       0.0     0.0     0.0      0.0  4.420679      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
3       0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  4.948033  0.0   0.0   0.0   0.000000
4       0.0     0.0     0.0      0.0  4.420679      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
...     ...     ...     ...      ...       ...      ... ...     ...       ...  ...   ...   ...        ...
1861    0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
1862    0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
1863    0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
1864    0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000
1865    0.0     0.0     0.0      0.0  0.000000      0.0  ...     0.0  0.000000  0.0   0.0   0.0   0.000000

[1866 rows x 500 columns]

Dataset Train & Test Split Details
80% dataset size used for training
20% dataset size used for training

80% Training Size = 1492
20% Testing Size = 374
SVM Accuracy  : 89.03743315508021
SVM Precision : 90.0750152647341
SVM Recall    : 89.55176767676767
SVM FMeasure  : 89.02417201715018
```

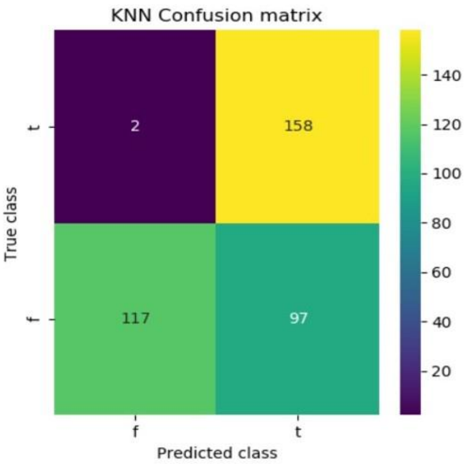


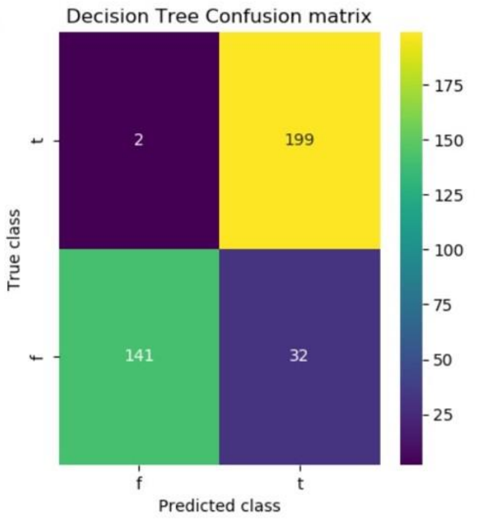
In above screen with SVM we got 89% accuracy, and we can see its confusion matrix graph. Similarly, we can run all the algorithms to get below output.

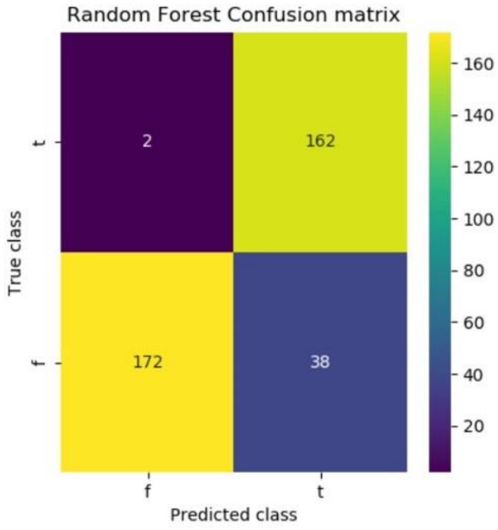


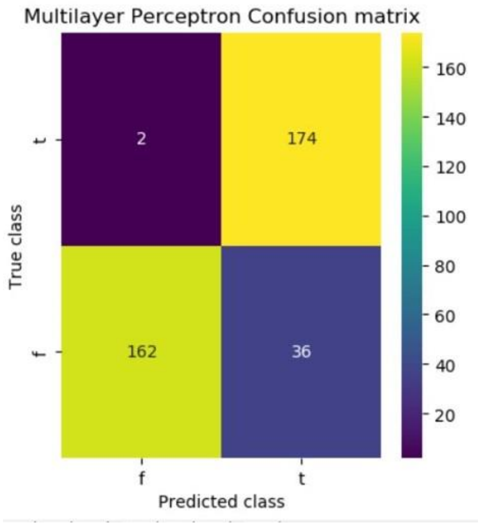
```
SVM Precision : 92.39234449760765
SVM Recall    : 91.81636155606408
SVM FMeasure  : 91.68978246066732

KNN Accuracy  : 72.72727272727273
KNN Precision : 80.33088235294117
KNN Recall    : 74.14772727272727
KNN FMeasure  : 71.61607142857143

RF Accuracy  : 89.03743315508021
RF Precision : 90.3768822096334
RF Recall    : 89.39289805269188
RF FMeasure  : 88.99581619984643

DT Accuracy  : 90.10695187165776
DT Precision : 91.12369337979094
DT Recall    : 90.66905958363245
DT FMeasure  : 90.09838646202283

NB Accuracy  : 88.77005347593582
NB Precision : 89.51300984528832
NB Recall    : 88.59031593406593
NB FMeasure  : 88.6764705882353

Multilayer Perceptron Accuracy  : 93.04812834224599
Multilayer Perceptron Precision : 93.79513160959702
Multilayer Perceptron Recall    : 93.0136124456646
Multilayer Perceptron FMeasure  : 93.01436781609196
```

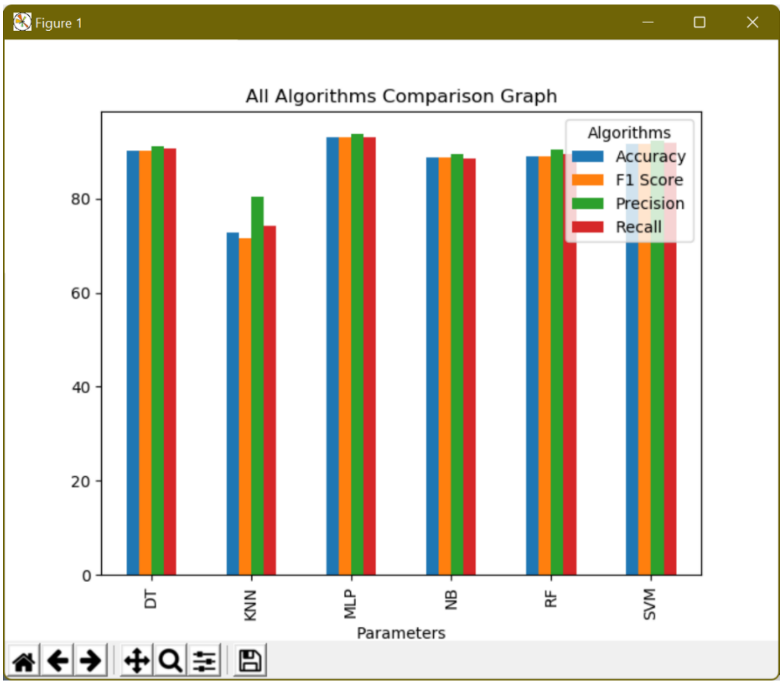

Naive Bayes Confusion matrix
True class
t
f
6
179
146
43
f
t
Predicted class
160
140
120
100
80
60
40
20


KNN Confusion matrix
True class
t
f
2
158
117
97
f
t
Predicted class
140
120
100
80
60
40
20

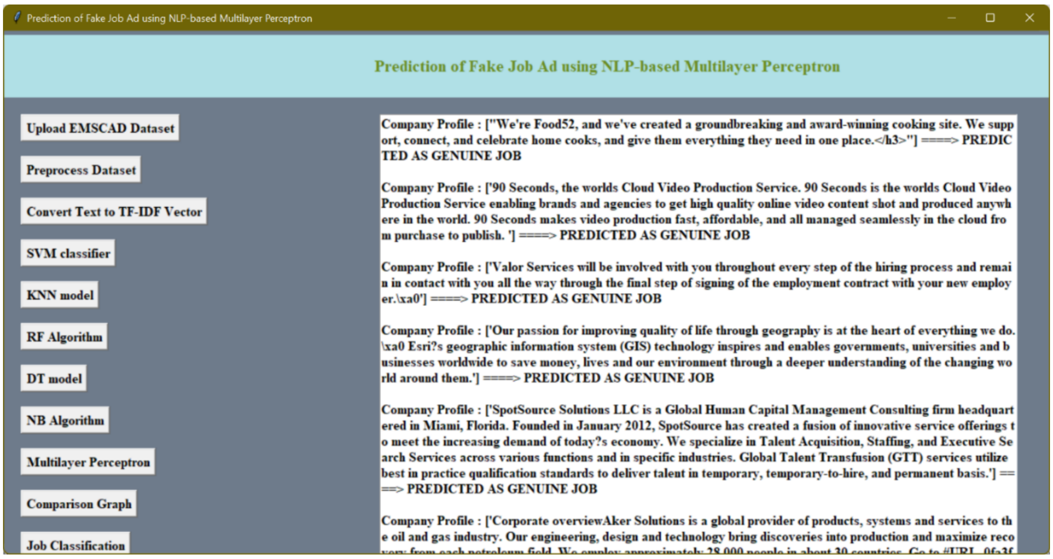

Decision Tree Confusion matrix
True class
t
f
2
199
141
32
f
t
Predicted class
175
150
125
100
75
50
25


Random Forest Confusion matrix
True class
t
f
2
162
172
38
f
t
Predicted class
160
140
120
100
80
60
40
20


Multilayer Perceptron Confusion matrix
True class
t
f
2
174
162
36
f
t
Predicted class
160
140
120
100
80
60
40
20



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.


Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Prediction of Fake Job Ad using NLP-based Multilayer Perceptron
Upload EMSCAD Dataset
Preprocess Dataset
Convert Text to TF-IDF Vector
SVM classifier
KNN model
RF Algorithm
DT model
NB Algorithm
Multilayer Perceptron
Comparison Graph
Job Classification

```
Company Profile : ["We're Food52, and we've created a groundbreaking and award-winning cooking site. We supp
ort, connect, and celebrate home cooks, and give them everything they need in one place.</h3>"] ====> PREDIC
TED AS GENUINE JOB

Company Profile : ['90 Seconds, the worlds Cloud Video Production Service. 90 Seconds is the worlds Cloud Video
Production Service enabling brands and agencies to get high quality online video content shot and produced anywh
ere in the world. 90 Seconds makes video production fast, affordable, and all managed seamlessly in the cloud fro
m purchase to publish. '] ====> PREDICTED AS GENUINE JOB

Company Profile : ['Valor Services will be involved with you throughout every step of the hiring process and remai
n in contact with you all the way through the final step of signing of the employment contract with your new employ
er.\xa0'] ====> PREDICTED AS GENUINE JOB

Company Profile : ['Our passion for improving quality of life through geography is at the heart of everything we do.
\xa0 Esri?s geographic information system (GIS) technology inspires and enables governments, universities and b
usinesses worldwide to save money, lives and our environment through a deeper understanding of the changing wo
rld around them.'] ====> PREDICTED AS GENUINE JOB

Company Profile : ['SpotSource Solutions LLC is a Global Human Capital Management Consulting firm headquart
ered in Miami, Florida. Founded in January 2012, SpotSource has created a fusion of innovative service offerings t
o meet the increasing demand of today?s economy. We specialize in Talent Acquisition, Staffing, and Executive Se
arch Services across various functions and in specific industries. Global Talent Transfusion (GTT) services utilize
best in practice qualification standards to deliver talent in temporary, temporary-to-hire, and permanent basis.'] ==
==> PREDICTED AS GENUINE JOB

Company Profile : ['Corporate overviewAker Solutions is a global provider of products, systems and services to th
e oil and gas industry. Our engineering, design and technology bring discoveries into production and maximize reco
very from each petroleum field. We employ approximately 28 000 people in about 30 countries. Go to #URL_0fa3f
```

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