A COMPARATIVE STUDY OF FAKE JOB POSTS USING DIFFERENT DATA MINING TECHNIQUES

R.GEYAMRUTHA VELLALA KIRAN KUMAR

Department of Master of Computer Science Department of Master of Computer Science

Miracle Educational Society Group of Institutions Miracle Educational Society Group of Institutions

Vizianagram– 535216 (AP) India Vizianagram– 535216 (AP) India

amrutharevalla8919@gmail.com Kiran.cool60@gmail.com

# ABSTRACT

In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world.So, fake job posting prediction task is going to be a great concern for all. Like many otherclassification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like LogisticRegression, support vector machine, naive bayes classifier, random forest classifier, to predict ajob post if it is real or fraudulent. Our application was built which takes Job Id, Description andjob Requirements to predict whether the given job post is real or fake. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. The trained classifier shows approximately 98% classification accuracy to predicts fraudulent job post

**Introduction**

In modern time, the development in the field of industry and technology has opened a hugeopportunity 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 andsocial 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 posts has increased the percentage of fraudjob postings which causes harassment to the job seekers. So, people lacks in showing interest tonew 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 arearound 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 seekerto 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.

# LITERATURE SURVEY

Many researches occurred to predict if a job post is real orfake. A good number of research worksare to check online fraud job advertiser. Vidros [1] et al. identified job scammers as fake onlinejob advertiser. They found statistics about many real and renowned companies and enterprises whoproduced fake job advertisements orvacancy posts with ill-motive. They experimented onEMSCAD dataset using several classification algorithms like naive bayes classifier, random forestclassifier, Zero R, One R etc. Random Forest Classifier showed the best performance on the datasetwith 89.5% classification accuracy. They found logistic regression performing very poor on thedataset. One R classifier performed well when they balanced the dataset and experimented on that.They tried in their work to find out the problems in ORF model (Online Recruitment Fraud) andtosolvethoseproblemsusingvariousdominantclassifiers.

Alghamdi [2] et al. proposed a model to detect fraud exposure in an onlinerecruitment system. They experimented on EMSCAD dataset using machine learning algorithm.They worked on this dataset in three steps- data pre-processing, feature selection and frauddetection using classifier. In the preprocessing step, they removed noise and html tags from the data so that the general text pattern remained preserved. They applied feature selection technique to reduce the number of attributes effectively and efficiently. Support Vector Machine was used for feature selection and ensemble classifier using random forest was used to detect fake job posts from the test data. Random forest classifier seemed a tree structured classifier which worked asensemble classifier with the help of majority voting technique. This classifier showed 97.4% classification accuracy to detect fake job posts.

Huynh[3] etal. proposed to use different deep neural network models like Text CNN, Bi-GRU-LSTM CNN and BiGRU CNN which are pre-trained with text dataset. Theyworked on classifying IT job dataset. They trained IT job dataset on Text CNN model consisting of convolution layer, pooling layer and fully connected layer. This model trained data throughconvolution and pooling layers. Then the trained weights were flattened and passed to the fullyconnected layer. This model used soft max function for classification technique. They also used ensemble classifier(Bi-GRUCNN, Bi-GRULSTMCNN) using majority voting technique to increase classification accuracy. They found 66% classification accuracy using Text CNN and 70% accuracy for Bi-GRU- LSTM CNN individually. This classification task performed best with ensemble classifier having an accuracy of72.4%.

Zhang[4]etal.proposed an automatic fake detect or model to distinguish between true and fake news (including articles, creators, subjects) using text processing. They had used acustom dataset of news or articles posted by Politick website twitter account. This dataset was used to train the proposed GDU diffusive unit model. Receiving input from multiple sources simultaneously, this trained model performed well a sanautomatic fake detector model.

Researchers experimented a good number of classifiers and feature selection technique to achieve good performance in the field of fake job post classification. Text processing using deep learning model, feature selection using support vector machine, data pre-processing etc. were mentioned approach to apply [8], [9], [10], [11], [12]. We have proposed to use deep neural network to predict job scams. We have applied the training method only ont he categorical attribute of the EMSCAD dataset instead of using text data. This approach reduces the number oftrainable attribute effectively with less processing time. We have made a comparative study on thesame features of EMSCAD dataset using K Nearest Neighbor, Naive Bayes classifier, fuzzy KNN,decision tree,support vector machine,random forest classifier and neural network

## PROPOSED SYSTEM

The system has used EMSCAD to detect fake job post. This dataset contains 18000 samples andeach row of the data has 18 attributes including the class label. The attributes are job\_id, title ,location, department,salary range,companyprofile, description ,requirements, benefits, telecommunication, has\_company\_logo, has questions, employment type, required experience,required education, industry, function, fraudulent (class label). Among these 18 attribute, we have used only7 attributes which are converted into categorical attribute.The telecommuting, has\_ company\_logo, has questions, employment type, required experience, required education andfraudulent are changed into categorical value from text value. For example, “employment type”values are replaced like this- 0 for “none”, 1 for ‘full-time”, 2 for “part-time” and 3 for “others”,4 for “contract’ and 5 for “temporary”. The main goal to convert these attributes into categoricalform is to classify fraudulent job advertisements without doing any text processing and natural language processing. In this work, we have used only those categorical attributes

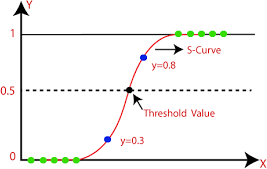
### ADVANTAGES OF PROPOSED SYSTEM

The proposed has been implemented EMSCAD technique which is very accurate and fast. Thesystem is very effective due to accurate detection of Fake job posts which creates inconsistency for the jobseeker to find their preferable jobs causing huge waste of their time.

### LOGISTIC REGRESSION

Logistic regression analysis studies the association between a categorical dependent variable anda set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial Logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced,or Widowed.Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar.

This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likely hood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It canperform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values andprovides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying grows that are not used during the analysis.



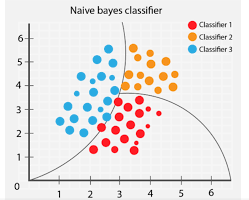
**FIG:LOGISTICREGRESSION**

### NAÏVE BAYES CLASSIFIER

The naive bayes approach is a supervised learning method which is based on asimplistichypothesis: it assumes that the presence (or absence)of a particular feature of a class is unrelated to the presence(or absence)of any other feature.

Yet, despite this, it appears robust and efficient. Its performance is comparable toother supervised learning techniques. Various reasons have been advanced in the literature. In thistutorial, we highlight an explanation based on the representation bias. The naive bayes classifieris a linear classifier, as well as linear discriminant analysis, logistic regression or linear SVM(support vector machine). The difference lies on the method of estimating the parameters of theclassifier(thelearning bias).

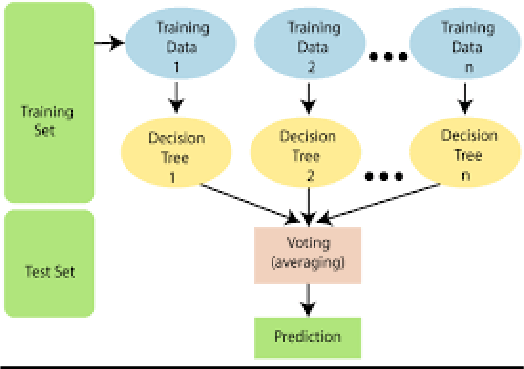
Thus, we introduce in a new presentation of the results of the learning process. Theclassifier is easier to understand, and its deployment is also made easier. In the first part of thistutorial, we present some theoretical aspects of the naive bayes classifier. Then, we implement theapproach on a dataset with Tanagra. We compare the obtained results (the parameters of the model) to those obtained with other linear approaches such as the logistic regression, the linear discriminate analysis and the linear SVM. We note that the results are highly consistent.



**FIG : NAÏVEBAYES**

### RANDOM FOREST CLASSIFIER:

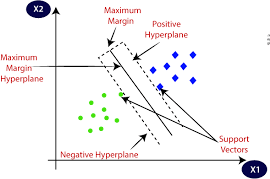
Random forests or random decision forests are an ensemble learning method for classification,regression and other tasks that operates by constructing a multitude of decision trees at trainingtime. For classification tasks, the output of the random forest is the class selected by most trees.For regression tasks, the mean or average prediction of the individual trees is returned. Randomdecision forests correct for decision trees' habit of over fitting to their training set. Random forestsgenerally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristic scan affect their performance.The first algorithm for random decision forests was created in 1995 by Tin KamHo[1] using the random sub space method, which ,in Ho's formulation ,is a way to implement the" stochastic discrimination" approach to classification proposed by Eugene Kleinberg.An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.).Theextension combines Breiman's "bagging" idea and random selection of features, introduced firstby Ho [1] and later independently by Amit and Geman [13] in order to construct a collection ofdecisiontreeswithcontrolled variance.



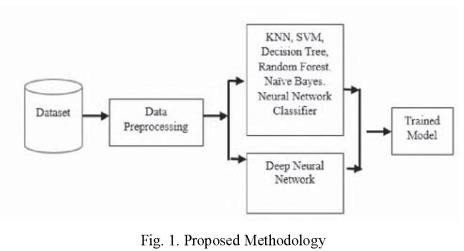
**FIG:RANDOMFOREST**

### SUPPORT VECTOR MACHINE:

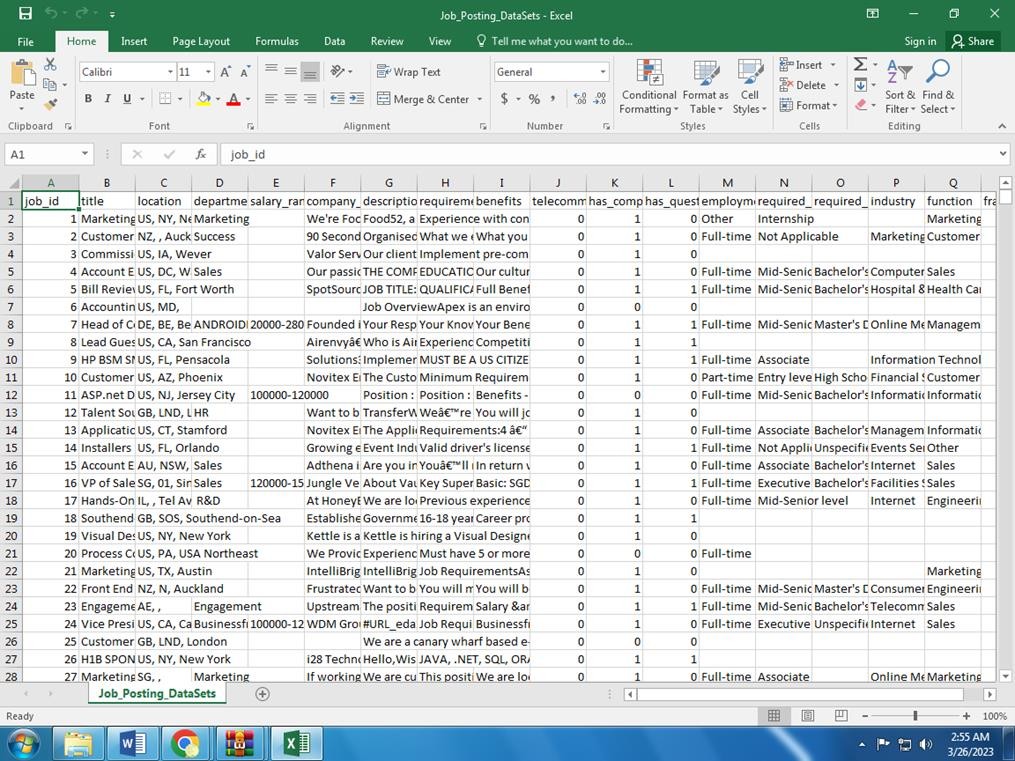
In classification tasks a discriminant machine learning technique aims at finding, based on an independent and identically distributed training dataset ,a discriminate function that can correctly predict labels for newly acquired instances. Unlike generative machine learning approaches, whichrequire computations of conditional probability distributions, a discriminant classification function takes a data point*s* and assigns it to one of the different classes that area part of the classification task. Less powerful than generative approaches, which are mostly used when prediction involves outlier detection, discriminant approaches require fewer computational resources and less trainingdata, especially for a multidimensional feature space and when only posterior probabilities areneeded. From a geometric perspective, learning a classifier is equivalent to finding the equation for a multidimensional surface that best separates the different classes in the feature space.The SVM kernelis a function that takes low-dimensional input space and transforms it into higher-dimensional space, i.e. it converts no separable problems to separable problems. It is mostly useful in non-linear separation problems. Simply put the kernel, does some extremely complex data transformations and then finds out the process to separate the database don the labels or outputs defined.



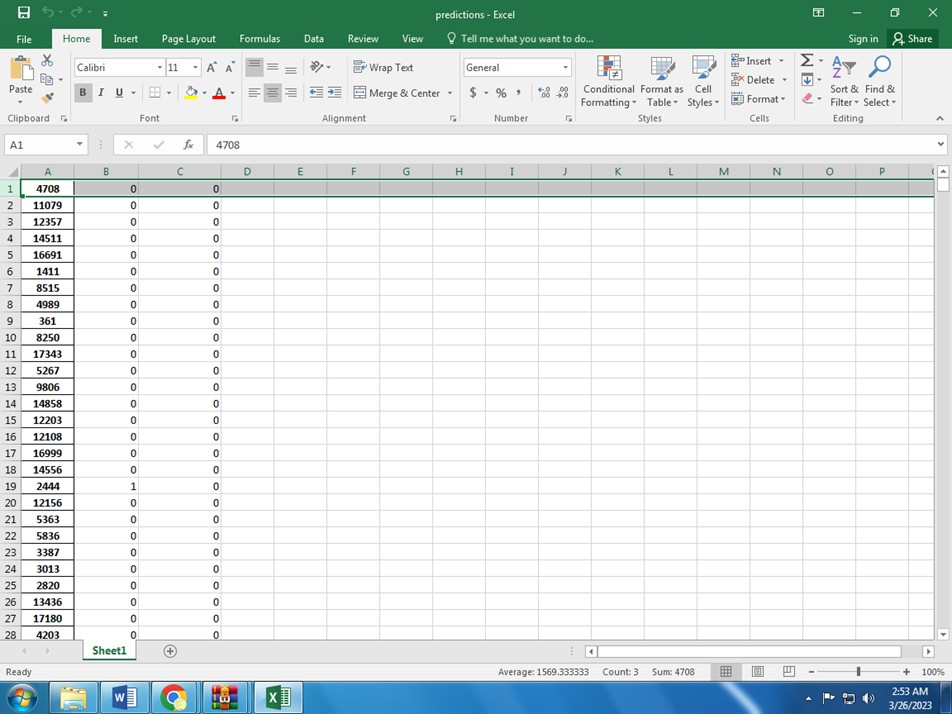
**FIG:SVM**



### SAMPLE DATA SET OF JOB PREDICTION:



**FIG: DATASET**

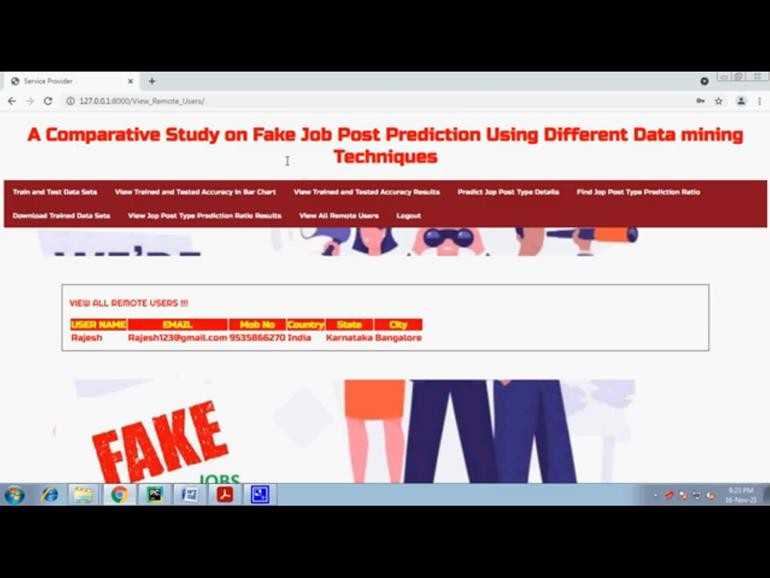


**FIG: DATASET**

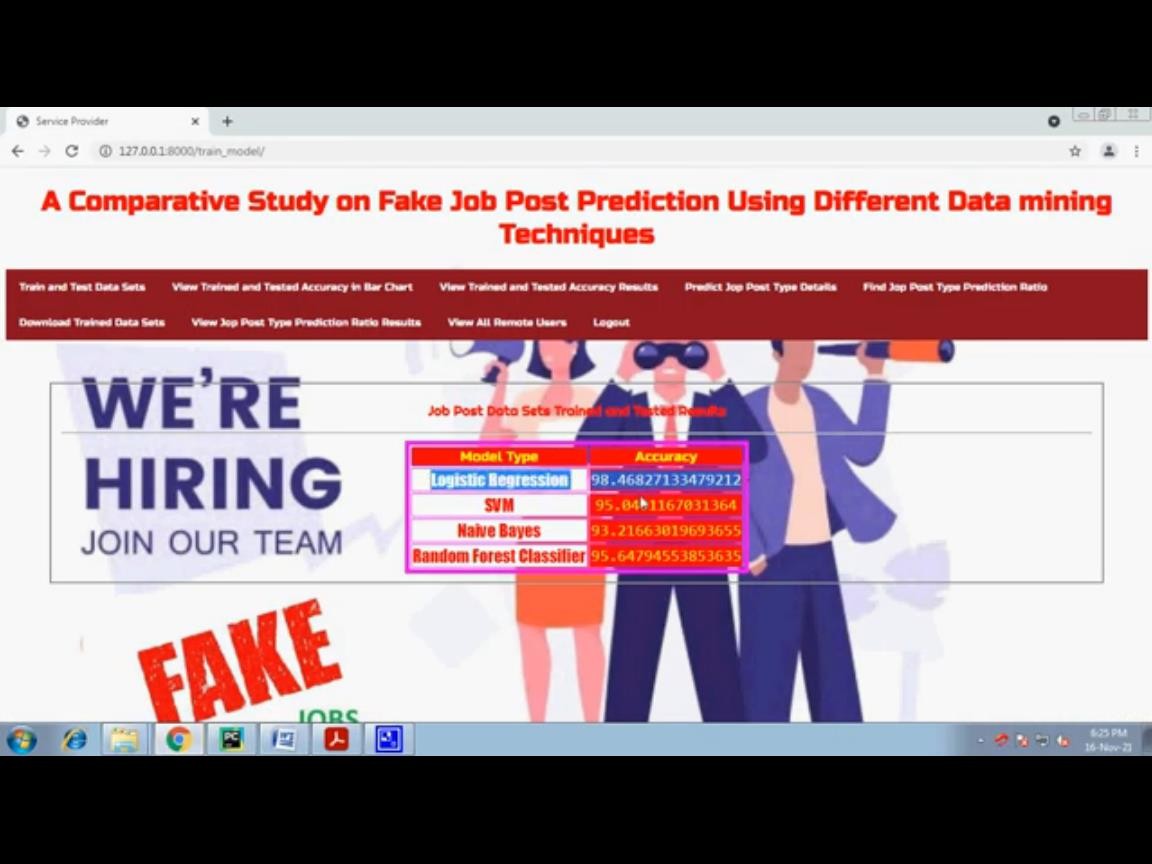
**RESULTS**



**FIG:SIGN UP PAGE**



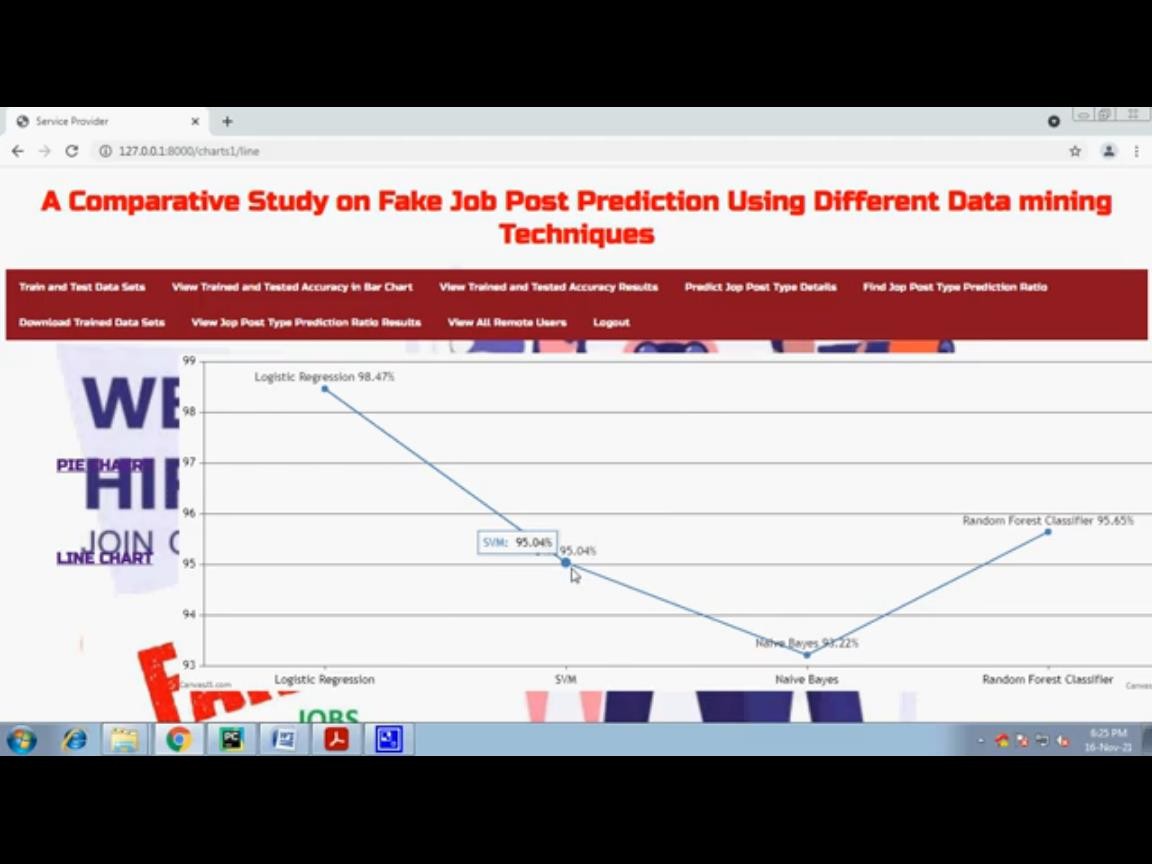
**FIG:VIEW ALL REMOTE USERS**



**FIG:TRAINED AND TEST DATASETS**



**FIG:BAR GRAPHS**



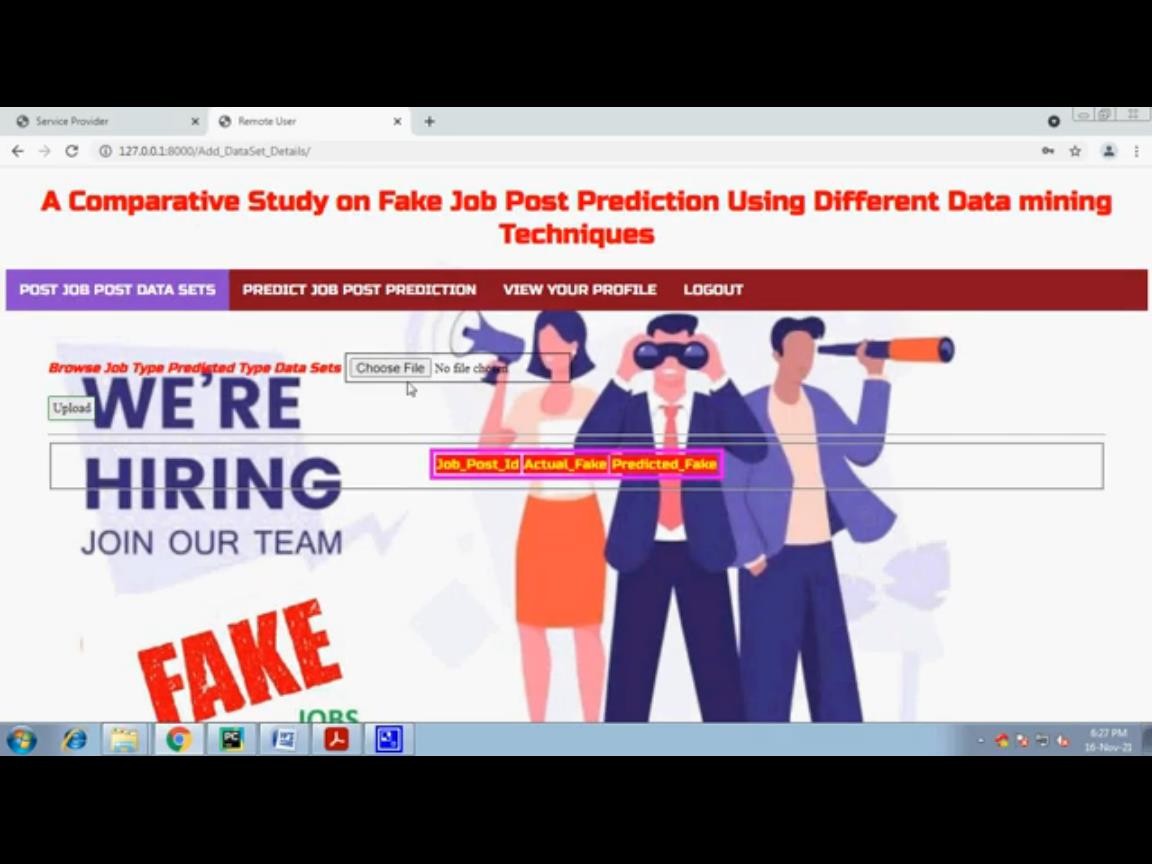
**FIG16:ACCURACY GRAPHS**



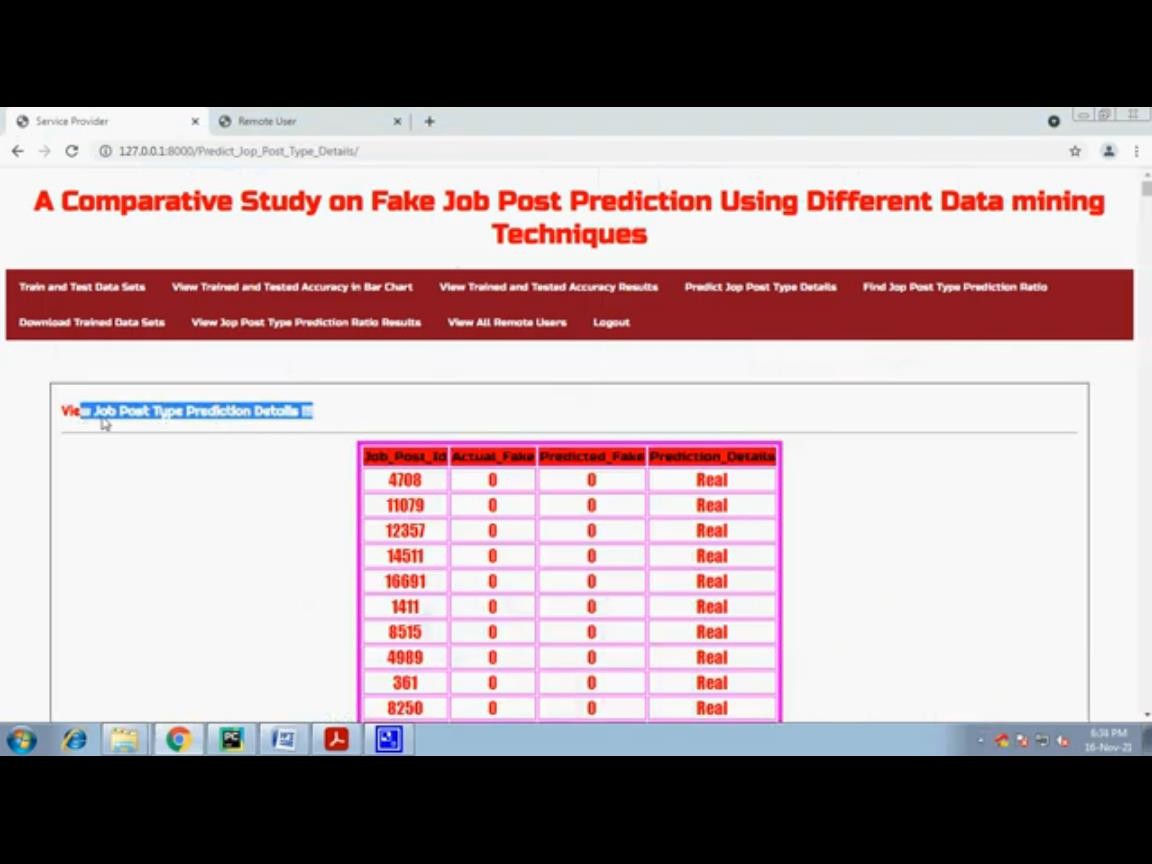
**FIG:PIE CHARTS**



**FIG: REGISTER**



**FIG:POST JOB DATASETS**



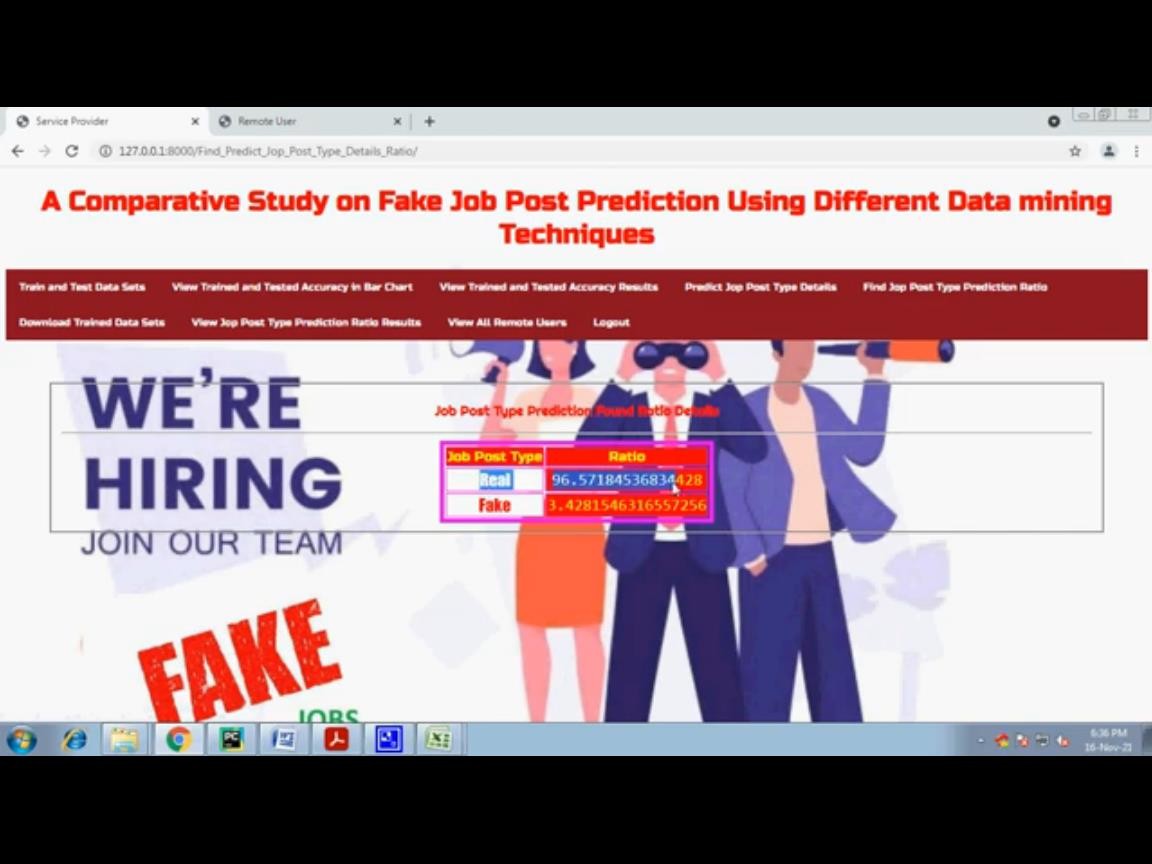
**FIG: FAKE AND REAL JOBS**



**FIG:PREDICT FAKE AND REAL JOBS**



**FIG: PREDICTION STATUS**



**FIG:JOBPREDICTIONRATIO**



**FIG:PIE CHART DISPLAYING RATIO**

## CONCLUSION

Job scam detection has become a great concern all over the world at present. In this paper, 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 Kaggle dataset which contains real life fake job posts. In this paper we have experimented both machine learning algorithms (SVM, Logistic Regression, Naïve Bayes, and Random Forest Algorithm).This workshows a comparative study on the evaluation of traditional machine learning. We have foundhighest classification accuracy for Random Forest Classifier among traditional machine learning algorithms and 99*%* accuracy for Logistic Regression.

## REFERENCES:

1. S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, “Automatic Detection of OnlineRecruitment 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”,JournalofInformationSecurity,2019,Vol10,pp.155176,https://doi.org/10.4236/iis.2019.103009.
3. Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen,“JobPrediction:From DeepNeural NetworkModels toApplications”,RIVF InternationalConferenceonComputingandCommunicationTechnologies(RIVF),2020.
4. JiaweiZhang,BowenDong,PhilipS.Yu,“FAKEDETECTOR:EffectiveFakeNewsDetection with Deep Diffusive Neural Network”, IEEE 36th International Conference on DataEngineering(ICDE),2020.
5. Scanlon, J.R. and Gerber, M.S., “Automatic Detection of Cyber Recruitment by ViolentExtremists”,SecurityInformatics,3,5,2014,https://doi.org/10.1186/s13388-014-0005-5
6. Y.Kim,“Convolutionalneuralnetworksforsentenceclassification,”arXivPrepr.arXiv1408.5882,2014.
7. T.VanHuynh,V.D.Nguyen,K.VanNguyen,N.L.-T.Nguyen,andA.G.-T.Nguyen,“HateSpeech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model,”arXivPrepr.arXiv1911.03644,2019.
8. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, “Semantic expansion using wordembedding clustering and convolutional neural network for improving short text classification,”Neurocomputing,vol.174,pp.806814,2016.
9. C. Li, G. Zhan, and Z. Li, “News Text Classification Based on Improved BiLSTM-CNN,” in20189thInternationalConferenceonInformationTechnologyinMedicineandEducation(ITME),2018,pp.890-893.
10. K. R. Remya and J. S. Ramya, “Using weighted majority voting classifier combination forrelation classification in biomedical texts,”International Conference on Control, Instrumentation,CommunicationandComputationalTechnologies(ICCICCT),2014,pp.1205-1209.
11. Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing EmailDetection.InternationalJournalofNetworkSecurity&ItsApplications,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-VuNguyen,KietVanNguyen,andNganLuu-ThuyNguyen."EmotionRecognitionforVietnameseSocialMediaText",arXivPrepr.ArXiv:1911.09339,2019.
13. Thin Van Dang,Vu Duc Nguyen, Kiet Van Nguyen and Ngan Luu- Thuy Nguyen, “Deeplearning for aspect detection on vietnamese reviews” in In Proceeding of the 2018 5th NAFOSTEDConferenceonInformationandComputerScience(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 December2014;pp.899-904.
15. Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online reviewcommunities. InProceedings of the 21st international conference on World Wide Web, Lyon,France,16-20 April2012;ACM:NewYork,NY,USA,2012;pp.201-210.
16. Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of FraudulentEmails by Employing Advanced FeatureAbundance. Egyptian Informatics Journal, Vol.15,pp.169-174.

<https://doi.org/10.1016/j.eij.2014.07.002>