Fake Job Post Prediction Using Machine Learning Algorithms

Mr. Gulshan P.[1], Mr. Mukund T.[2], Mr. Ajay A.[3], Mr. Pankaj Kumar[4], Mrs. Aruna M G[5], Dr. Malatesh S H[6]

[1,2,3]*Student, Department of CSE, M. S. Engineering College, Bangalore, India*
[4]*Assistant Professor, Department of CSE, M. S. Engineering College, Bangalore, India*
[5]*Associate Professor, Department of CSE, M. S. Engineering College, Bangalore, India*
[6]*Professor and HOD of CSE, M. S. Engineering College, Bangalore, India*

Abstract **- During the pandemic, there is strong rise in the number of online job posted on various job portals. So, fake job posting prediction task is going to be big problems for all. Thus, these fake jobs can be precisely detected and classified from a pool of job posts of both fake and real jobs by using advanced deep learning as well as machine learning classification algorithms. . This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naive 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 EMSCAD which containing 18000 employee samples. 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.**

Index Terms **- Random Forest, KNN, Naive Bayes, Real and Fake, support vector machine, deep learning, and classification.**

I.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. 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.

II.EXISTING SYSTEM

Vidros et al. identified job scammers as fake online job advertiser. They found statistics about many real and renowned companies and enterprises who produced fake job advertisements or vacancy posts with ill-motive. They experimented on EMSCAD dataset using several classification algorithms like naive bayes classifier, random forest classifier, Zero R, One R etc. Random Forest Classifier showed the best performance on the dataset with 89.5% classification accuracy.

Alghamdi et al. proposed a model to detect fraud exposure in an online recruitment system. They

experimented on EMSCAD dataset using machine learning algorithm.

Huynh et al. proposed to use different deep neural network models like Text CNN, BiGRU-LSTM CNN and BiGRU CNN which are pre-trained with text dataset. They worked on classifying IT job dataset

TABLE I: LITERATURE SUMMARY

S.NO	TITLE	METHODOLGY	ADVANTAGES
1.	An Intelligent Model for Online Recruitment Fraud Detaction-2019	Random Forest	It detect the online recruitment fraud automatically
2.	Job Prediction: From Deep Neural Network Models to Applications-2020	TextCNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN	It determine the job is suitable for a student or a person.
3.	FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network-2020	DNN	It is used to identify the fake news timely.

III.PROPOSED SYSTEM

In our proposed model uses to use different machine learning techniques and classification algorithm like KNN, decision tree, support vector machine, naive 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 EMSCAD which containing 18000 employee samples. Deep neural network as a classifier, performs great for this classification task.

IV. TRAINED DATA AND PRE- PROCESSING

We are passing raw data set to the input we are using sum, average, standard deviation mathematical approaches to fill the missing values and removing the repeated data

A. Pre- Processing

Prior to training and data evaluation using machine learning, data processing is a normal first step. Algorithms for machine learning are always as useful as information you fed them. It is important to format correct data and to include relevant items so that they are consistent enough to produce best outcomes possible. Stopword removal, tokenization, lower case and punctuation removal are all examples of data refinement. This allows us to reduce the size of the real data by removing irrelevant information. We created a simple processing function for each document to remove punctuation and non-letter characters, followed by the letter case in the document was lowered. Make different steps to clean text (remove all non- alphanumeric characters, delete stop words, delete missing rows, etc.).

B. Feature Extraction

Feature selection is the method of reduction that reduces an original batch of actual data to even more controllable computing categories. Ngram are a type of grammatical unit. Every job post is mined for unigrams and bigrams. Tfidf Vectorizer is used to score the relative importance words in a document. Count Vectorizer is used for creating vectors that have a dimensionality equal to the size of our vocabulary, and if the text data features that vocab word, we will put such in that dimension. The result of this will be very large vectors, if we use them on real text data, however, we will get very accurate counts of the word content of our text data.

Trained data:

In case of conventional machine learning algorithms like KNN, Random Forest, SVM etc. we have used hold out cross validation. 80% of the total data was used for training and 20% was used for testing and checking the model performance. In KNN model, we have applied K value from 1 to 40 and minimum error is found when k= 13. Mean error rate was less than 0.05 during the training process. RBF kernel is used in SVM and gamma value = 0.001 is also used.

Prediction:

User passes the input parameters like Telecommuting, has_company_logo,has_questions, employment_type, required experience, required_education system will predict the given post is fake or real using deep neural network.

Algorithms:

For the prediction, multiple supervised learning algorithms are trained using the training set, after which using the testing set performance evaluation happens. These algorithms are:

A.Random Forest

STEP 1: START

STEP 2: SPLIT dataset into 67 percent training set, 33 percent testing set .

STEP 3: FOR train dataset
CALL RFClassifier
TRAINRFClssifier
STEP 4: FOR test dataset
CALL RFClassifier
PREDICT the label
COMPUTE
AccuracyScore
SAVE AccuracyScore
DISPLAY
ConfusionMatrix
STEP 5: STOP

B. KNN Classifier
Pseudo code
Procedure Train()
// Input: train set, test set
// Output: Trained model
Step 1: Read Train set and test set
Step 2: Build KNN classifier
Step 3: Train the model using fit()
Step 4: Performance Graph Returned Trained Model

C. Naïve Bayes
Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

STEP 1: START
STEP 2: SPLIT dataset into 67 percent training set, 33 percent testing set
STEP 3: FOR train dataset
CALL MultinomialNB
TRAIN MultinomialNB
STEP 4: FOR test dataset
CALL MultinomialNB
PREDICT the label
COMPUTE AccuracyScore
SAVE AccuracyScore
DISPLAY ConfusionMatrix
STEP 5: STOP

D. Support Vector Machine (SVM)
It shows many unique advantages in a small sample, nonlinear, and high-dimensional pattern recognition and can be extended to other functions such as function fitting ML problems. Before the rise of deep learning, SVM was considered the most successful and best-performing machine learning method in recent decades. The SVM method is based on the Vapnik Chervonenkis(VC) dimension theory of statistical learning theory and the principle of structural risk minimization.
STEP 1: START
STEP 2: SPLIT dataset into 67 percent training set, 33 percent testing set
STEP 3: FOR train dataset
CALL SVMClassifier TRAIN SVMClassifier
STEP 4: FOR test dataset
CALL SVMClassifier PREDICT the label
COMPUTEAccuracyScore
DISPLAY ConfusionMatrix
STEP 5: STOP

V.SYSTEM ARCHITECTURE

The system "design" is defined as the process of applying various requirements and permits it physical realization. Various designs are followed to develop the system the design specification describes the features of the system, the opponent or elements of the system and their appearance to the end-users.
The below figure. 1 illustrates the steps in fake job post detection system, Data preprocessing first performed in our fake job post detection structure, including duplicate, outlier, and missing value processing. Then, we are applying various machine learning and deep learning algorithms to train the model to detect the fake job post detection.

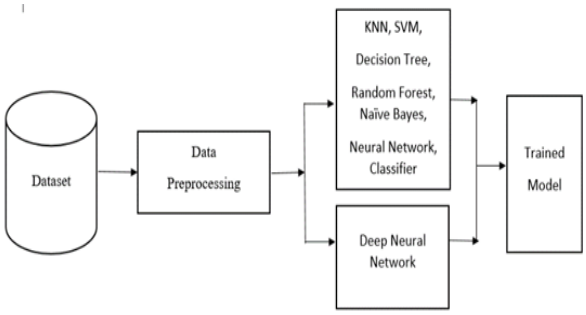

Fig. 1 System Architecture

VI. RESULTS AND DISCUSSION

The system was created using Windows 10 as well as a 64-bit processor with 8 GB of RAM. The model implemented with the help of Python v3.7.8

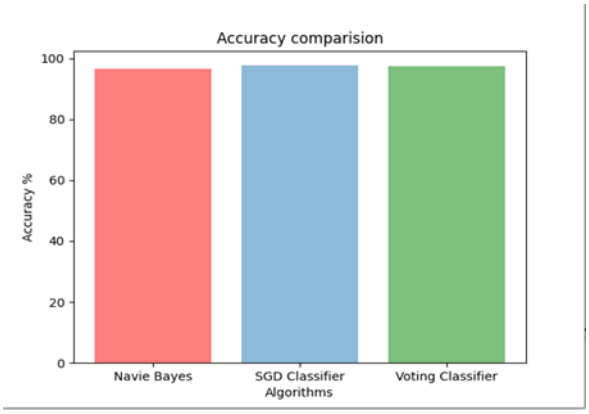

Fig: 2 Accuracy comparison between algorithms

From the values calculated in the confusion matrix, an accuracy graph (Fig. 2) is generated for each algorithm for comparison for best algorithm with highest accuracy.

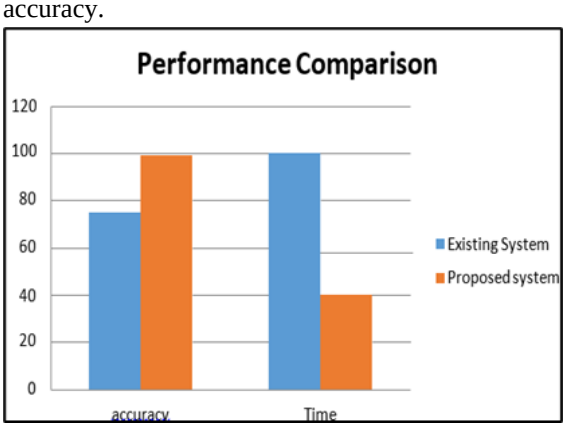

Fig. 3 Performance Comparison

Figure 3 shows the phase involves the evaluation of performance of machine learning algorithms on training and building the model, and then predicting the label of the news article given by the user. The impact is measured in average accuracy and time taken to train and predict.

The accuracy is determined by accuracy score of the ML models, which is measured in percentage. The average time taken is determined by comparing the evaluation time taken for training and prediction by a model.

VII. CONCLUSION

In this work, we have developed a fake job post detection software by applying supervised machine learning algorithms to classify a given job post taken as input from user, as real or fake.

A simple approach for fake job post detection using is performed using KNN classifier. The way they get these probabilities is by using KNN, which describes the probability of a feature which has misclassification and less prediction. In this proposed model, initially for training 80% data is being used, and for testing 20% of data are pre-processed.

The dataset are taken from kaggle.com, which have more-than 18000 data. After pre-processing, data features extraction and tokenization take place. Updated dataset after pre-processing and feature extraction is used to train the Machine Learning Model(s). After training, the model is used to classify the given news articles, into two binary classifications, that of “real” and “fake”.

REFERENCE

[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] 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 DNN Models to Applications”, RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R, “Automatic Detection of Cyber Recruitment by Violent Extremists”, Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s 13388-014-0005-5

[6] Kim, “Convolutional neural networks for sentence classification,” arXiv Prepr. arXiv1408.5882, 2014.

[7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, “Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model,” arXiv Prepr. arXiv1911.03644, 2019.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806- 814, 2016.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.