A Comparative Analysis on the Identification of Fake Job Posts Using Various Data Mining Techniques

Shaik Salman Hussain[1] Mrs. Mohammed Asma[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

[2]Associate Professor, Dept. of Computer Science and Engineering, Lords Institute of Engineering & Technology, Hyderabad, Telangana

***Abstract* - As per with the development of social media and modern technologies, advertising new job openings has recently become a very prevalent problem in the current world. Therefore, everyone will have a lot of reason to be concerned about bogus job postings. Fake job posing prediction presents a variety of difficulties, just like many other categorization tasks. In order to determine whether a job posting is legitimate or fake, this study advocated using a variety of data mining techniques and classification algorithms, including KNN, decision trees, support vector machines, naive bayes classifiers, random forest classifiers, and deep neural networks. The For this deep neural network classifier, three thick layers were used. A bogus job advertisement can be predicted with a classification accuracy of about 98% by the trained classifier using DNN. The article suggests an automated application that uses machine learning-based classification approaches to prevent fraudulent job postings online. The outputs of various classifiers are evaluated in order to determine the best employment scam detection model. These classifiers are used to verify fraudulent posts on the web. It assists in identifying phoney job postings among a large number of postings. For the purpose of identifying fake job postings, two main categories of classifiers—single classifiers and ensemble classifiers—are taken into consideration. Nevertheless, experimental findings show that ensemble classifiers are the most effective classification to identify fraud over the single classifiers.**

***Index Terms* — Support vector machines, Deep learning, Machine learning algorithms, Neural networks, Data mining, Task analysis, Random forests.**

I. INTRODUCTION

The advancement of business and technology in the current day has given job seekers a great deal of chance to find new and varied positions. The adverts for these job openings enable job searchers determine their options based on their availability, qualifications, experience, suitability, etc. The influence of social media and the internet on the hiring process has increased. The effectiveness of a recruitment process depends on how well it is advertised, therefore social media has a significant influence here. Job information can now be shared in ever-new ways thanks to social media and electronic media marketing.Instead of this, the opportunity to disseminate job advertisements quickly has increased the number of fraudulent job postings, which irritate job seekers. People don't respond to fresh job postings because they want to keep their personal, academic, and professional information secure and consistent.

The genuine goal of legitimate job advertisements via social and electronic media thus has a very difficult struggle to win over

people's trust and trustworthiness. Technologies are all around us to improve and ease our lives, not to create unsafe working conditions. Recruiting new personnel will improve greatly if job postings can be appropriately filtered to identify fake job postings. False job postings make it difficult for job seekers to find the positions they desire, which is a significant waste of their time. A new door is opened to deal with challenges in the area of human resource management by an automated system that predicts fake job postings.

II. SYSTEM ANALYSIS

Problem statement:

The task of predicting fake job postings will cause everyone a lot of stress. Fake job posing prediction presents a number of difficulties, just like many other categorization tasks. For identifying bogus posts, a machine learning approach is used, which makes use of numerous classification algorithms. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of recognising scammers on job postings. A classifier uses training data to map input variables to target classes. There is a brief description of the classifiers used in the paper to distinguish phoney job postings from the others. Generally speaking, these classifier-based predictions can be divided into two types: single classifier-based predictions and ensemble classifier-based predictions. For identifying bogus posts, a machine learning approach is used, which makes use of numerous classification algorithms. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of recognising scammers on job postings. A classifier uses training data to map input variables to target classes. addressed are classifiers.

Objective:

One of the important challenges recently addressed in the area of online recruitment frauds (ORF) is employment scam. Nowadays, a lot of businesses choose to list their open positions online so that job hunters can quickly and easily access them. However, since the con artists offer employment to job seekers in exchange for money, this could be one of their scams. A reputable organization may be targeted by fraudulent job postings in order to damage their reputation. These fraudulent job post detections attract a lot of interest in developing an automated method for recognizing bogus jobs and alerting people to them so they won't apply for them. In order to accomplish this, a machine learning approach is used, which makes use of a number of classification algorithms. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of recognising scammers on job postings. A classifier uses training data to map input variables to target classes. There is a brief description of the classifiers used in the paper to distinguish phoney job postings from the others. Generally speaking, these classifier-based predictions can be divided into two types: single classifier-based predictions and ensemble classifier-based predictions.

Proposed System:

For identifying bogus posts, a machine learning approach is used, which makes use of numerous classification algorithms. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of recognising scammers on job postings. A classifier uses training data to map input variables to target classes. There is a brief description of the classifiers used in the paper to

distinguish phoney job postings from the others. Generally speaking, these classifier-based predictions can be divided into two types: single classifier-based predictions and ensemble classifier-based predictions.

Advantages:

• For identifying bogus posts, a machine learning approach is used, which makes use of numerous categorization algorithms. In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user.

• In this instance, a classification tool separates the fraudulent job postings from a broader collection of job postings and notifies the user. supervised learning algorithms are initially taken into consideration as classification techniques to address the issue of recognising scammers on job postings. A classifier uses training data to map input variables to target classes. addressed are classifiers.

III. PROPOSED MODULAR IMPLEMENTATION

Algorithm/ Technique Used:

After Feature Transformation and Data Pre-processing, the dataset is fitted to a model. The algorithm receives the training set in order to learn how to forecast values. After creating a target variable to predict, testing data is provided as input. The models are created using a technical implementation for the Drug Review dataset for the Drug Recommendation System.

Data preparation techniques include feature extraction, feature extraction with TF-IDF and N-gram analysis, splitting the data set into test and training data, bidirectional LSTM data modelling, data evaluation and prediction, and construction of the fake job prediction system.

Data Preparation for common classification algorithms:

Prior to feeding the data into the widely used classification algorithms, the data must be prepared as follows.

1. Review the fake_job_postings.csv dataset.
2. Eliminate undesirable columns from the raw data.
3. Combine every categorical variable into a single data frame.
4. Codify categorical data
5. Eliminate HTML Tags
6. Take a dataset sample
7. Make the dependent variables encoded
8. Remove stop words, change the text to lower case, and lemmatize the text data.
9. The following methods for preparing training data are used to provide it to machine learning algorithms

a. When using SelectKBest and count vectorizer with unigrams.

b. Using bigrams and the count vectorizer and SelectKBest

Using TfidfVectorizer and SelectKBest c. Using count vectorizer and SelectKBest with trigrams d.

10. The outcomes for the 3 ways mentioned above are as follows:

	Accuracy			
Models	**Unigram**	**Bigram**	**Trigram**	**Tf-Idf**
Gaussian Naïve Bayes	0.9250	0.9481	0.9654	0.9423
Support Vector Classifier	0.8760	0.9164	0.8530	0.9221
XGBoost	0.8876	0.8760	0.7838	0.8847
Random Forest	0.8962	0.9077	0.8991	0.8991
LightGBM	0.9020	0.8760	0.7636	0.9106

Data Preparation for Bidirectional LSTM model:

Before feeding the Bidirectional LSTM model with the data, the following preparations must be made.

1. Review the fake_job_postings.csv dataset.
2. Consolidate all textual features into a single feature to enable the use of NLP techniques.
3. Remove superfluous features
4. Purify the dataset by carrying out the subsequent actions.

a. Lowercase the entire text.

b. Eliminate stop words c. Carry out lemmatization. Natural Language Processing (NLP) uses a text normalisation technique that changes the mode of any form of word to that of its basic root.

d. Use a tokenizer to create tokens for the words.
e. Divide the dataset into sets for training and testing.
5. Supply the Bidirectional LSTM with the training data.

Below is the proposed modular implementation

of the project.

Admin Module:

Log in, then upload the Kaggle-downloaded used vehicle market dataset.
3. Observational Data Analysis

4. Data Preprocessing a. Concatenate all textual features for NLP analysis that is simpler to perform.
B. Remove superfluous features C. Purify the dataset I. Remove punctuation II. Change the text's case
Tokenizing the text, getting rid of stop words, lemmatizing, and using the Bag of Words are the next three steps.
a. Using pad sequences to ensure that all reviews are the same length.
Establish dependent and independent variables.
b. Designing Split arrays for Train and Test

5. Applying various classification techniques to the dataset
A Random Forest and B SVM
c. XGBoost; d. Gaussian Naive Baiyes; e. Light GBM
6. Model construction utilising Bidirectional LSTM

IV. PROJECT EXECUTION

Admin Login:

The admin module's login page is located here. In order to carry out actions like uploading the dataset, the administrator must first log in with his credentials. During dataset training, a dataset's exploratory data analysis, using a dataset to train many machine learning algorithms to determine the optimal algorithm for accuracy and Create a model that can be utilised by users and hosted on the Flask application.

Fig: Admin Login

Upload Dataset:

The system administrator can upload datasets that are used to train machine learning models on this page. To upload a file to a server, an administrator must first choose the file by clicking the Choose file button, then click the Upload button. A success message indicating that the file was successfully uploaded would be shown once the upload was finished. We are utilising the dataset fake_job_dataset.csv for this project.

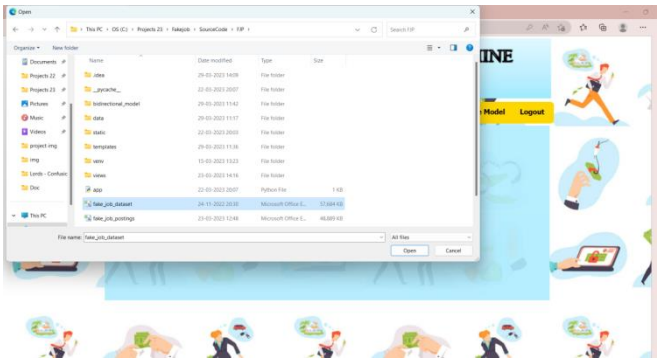

Fig: Upload Dataset & File Uploaded Successfully.

Data Analysis:

Exploratory data analysis is used to examine the dataset for any missing data, spot trends, and establish connections between different output characteristics using graphs, statistics, and other visual aids.

Fake Post Analysis:

The graph below displays the Fake Post Analysis over the dataset's data.

Fig: Fake Post Analysis

Title Analysis:

The graph below displays the Title Analysis over the dataset's data.

Fig: Title Analysis

Location Analysis:

The graph below displays the location analysis over the dataset's data.

Fig: Location Analysis

Company Logo Analysis:

The graph below displays a company logo analysis over the dataset's data.

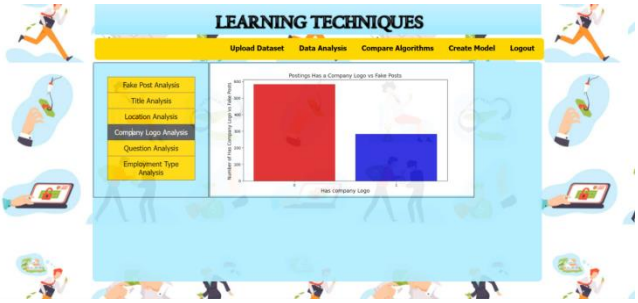

Fig Company Logo Analysis

Question Analysis:

The graph below displays the question analysis over the dataset's data.

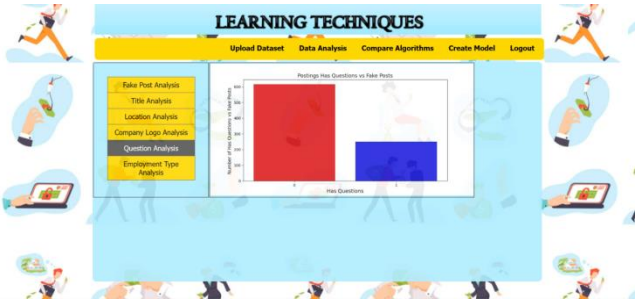

Fig Question Analysis

Employment Type Analysis:

The Employment Type Analysis over the data in the dataset is displayed in the graph below.

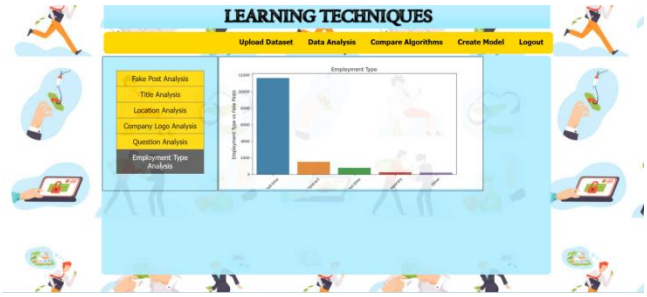

Fig Employment Type Analysis

Unigram Accuracy Comparison:

When the dataset is fed into several Unigram Accuracy Comparison methods

Fig: Unigram Accuracy Comparison

Bigram Accuracy Comparison:

When the dataset is fed into several Unigram Accuracy Comparison methods

Fig: Bigram Accuracy Comparison

Trigram Accuracy Comparison:

When the dataset is fed into several trigram accuracy comparison algorithms

Fig: Trigram Accuracy Comparison

TF-IDF Accuracy Comparison:

When the dataset is fed to different TF algorithmsAccuracy Comparison for IDD

Fig: TF-IDF Accuracy Comparison

Accuracy Summary:

The administrator can use this page to train several algorithms on a dataset and determine each algorithm's test accuracy.

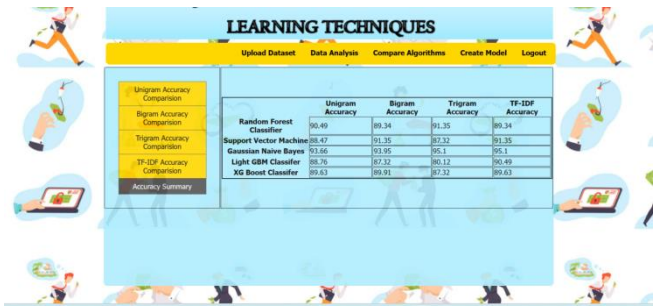

Fig: Accuracy Summary

Create Model:

This screen displays the model's training accuracy as 98.01% and test accuracy as 90.8%.

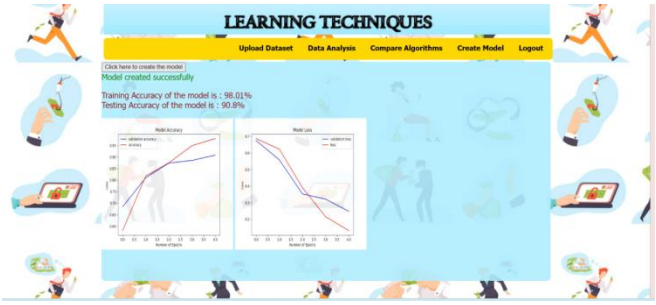

Fig: Create Model

CONCLUSION

The need to identify work scams globally has grown in importance recently. We looked at the effects of employment scams in this work since they might be a very lucrative area of research and make it challenging to spot dishonest job adverts. The primary contribution of this study is to provide support for the claim that deep learning may be utilised in specific situations to detect fake employment. Our results show that, after thorough pre-processing of a short dataset, a bidirectional LSTM model is capable of identifying a number of potentially subtle language patterns that a person may employ (or may not be able to perceive). They are useful for classifying phoney work since many of these linguistic traits are relatively simple to pick up. A few of the fundamental characteristics that our computer has identified to distinguish fake jobs include exaggerations, slang phrases, and generalisations.

Detecting employment scams will direct job searchers to only get genuine offers from businesses. In this study, numerous machine learning algorithms are suggested as counters to job fraud detection. A supervised technique is used to demonstrate how various classifiers may be utilised to identify employment scams. According to experimental findings, Bidirectional LSTM surpasses its rival classification technology. The accuracy of the suggested strategy, 98.01%, was significantly higher than that of the existing approaches.

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 , 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] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882.

[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.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 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), pp. 1205-1209.