IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OFFLINE FAKE JOB PREDICTION USING BI-DIRECTIONAL LSTM METHOD

[1] Kotha Mohan Krishna, [2]Mohammed Khaja Fayazuddin, [3]Kurivella Sai Sri Vidya, [4]Gunturu Gopi Krishna, [5]Istharla Chaithanya Sujith Sinha

[1] Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student [1]Department of Computer Science and Engineering,

[2] Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

[3] Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

[4] Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

[5] Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

Abstract: The Scammers are taking advantage of the growing number of job hunters by creating fake job advertisements, which trick job seekers into applying for the positions. Through this approach, scammers hope to obtain personal identification information (PII) or benefit by requesting payment. It is essential to address this problem statement. The most recent developments in machine learning algorithms are highly helpful in resolving this issue. After doing some research on machine learning techniques. We identified that the Multi-Layer Perceptron and the Bi-directional LSTM are the best algorithms that can be utilized to address this problem quite effectively. This study offers a machine learning technique that detects fraudulent job postings using textual data. From the job post text, we extract different data like the presence of specific keywords. We extract features like the job title and the job description to predict if a job is fraudulent or real. A few metrics are used to evaluate multi-layer perceptron and bi-directional LSTM, including accuracy, F1 score, ROC_AUC score, precision, and more. This research will assist safeguard job seekers from fraud and scams in the employment market by using automated algorithms to detect fake job postings.

Keywords: Machine Learning, MLP, LSTM, Bi-Directional LSTM.

I. INTRODUCTION

Online job portals are now essential for millions of people worldwide due to the internet's rapid growth, which has completely changed the way companies and job seekers interact. But this ease of use has also resulted in a serious problem: an increase of fraudulent job advertisements. In addition to wasting job seekers' time, these dishonest advertisements present serious risks like financial fraud and identity theft. Effective techniques for identifying and lessening the effects of fake job postings are crucial in the fight against this.

Machine learning (AI) and natural language processing (NLP) methods have shown promise in detecting fraudulent content across a range of fields, including sentiment analysis, spam email detection, and false news detection. In this field, Multi-layer perceptron, recurrent neural networks (RNNs), and its variations, such as bi-directional long short-term memory (LSTM) networks, have become effective instruments for textual sequential pattern recognition and sequential data analysis. Bi-directional LSTM (Bi-LSTM) networks, which use contextual information from both previous and upcoming time steps, have demonstrated impressive performance in a variety of NLP applications.

Our paper proposes a new method for detecting fake job postings using Bi-directional LSTM networks and compare the robust nature of Bi-directional LSTM with Multi-Layer Perceptron. We believe that Bi-directional LSTM's ability to effectively understand the intricate structure of textual data can help distinguish between genuine and fraudulent job advertisements. We outline a detailed methodology encompassing text preprocessing, word embedding, and model training, and we assess our proposed model's performance by training dataset which is a part of whole data but not used in training of the model. Through a series of experiments, we showcase the effectiveness of our approach and compare it with other cutting-edge techniques, highlighting the potential of Bi-directional LSTM networks in tackling the escalating problem of fake job postings.

II. PROBLEM IDENTIFICATION

The problem of fake job ads is getting worse because more and more people are looking for jobs online. Even though online job sites and social media make it easy to find jobs, they also make it easier for scammers to trick people. Job seekers often waste their time and money chasing fake job offers that do not exist. These scams can also lead to serious problems like losing money, having their identity stolen, or even being in danger. When fake job ads become common, it makes people doubt if online job sites can be trusted at all. That is why it is important to find better ways to spot and stop fake job ads. This will protect job seekers and help them feel safe when looking for jobs online.

III. LITERATURE SURVEY

Misleading content identification is strongly related to the detection of fake job postings. We have gone through several critical studies and techniques employed in detecting misleading content like fake news, spam mail detection and the previous work done to identify fake job postings.

Fake news identification is the best parallel problem for fake job prediction. We have gone through some base papers in which great work is done in predicting the fake news. Fake news in social media play a key role in identifying fake user accounts and it mainly relies on three perspectives they are – How the fake news is being written, How the fake news is being viral in social media and last how the user is being affected by the fake news. Features related to news content and social context are extracted and a machine learning models are imposed to recognize fake news It helped in understanding models used in our models. Zhou and Zafarani (2018)[1] suggested a hybrid strategy that integrated user behavior analysis with linguistic elements. Drucker et al. (1999)[2] have used Support Vector Machines (SVM) for spam email classification. It gave better

results in mail classification. Sultana Umme[3] proposed various models like naïve bayes classifier, MLP, k-nearest neighbors to identify the fake jobs. That research helped a lot in understanding much deeper about the data and models. S.Vidros [4] have done a great work in correlating the other projects like spam mails and fake news detection with the fake job prediction system and analyzed the EMSCAD dataset thoroughly. William Yang Wang[5] have done a great work in fake news detection using 5 models mails bidirectional LSTM. Rashid Amen[6] did mail filtering project explaining supervised and unsupervised machine learning algorithms and provided valuable information about neural networks and data mining techniques.

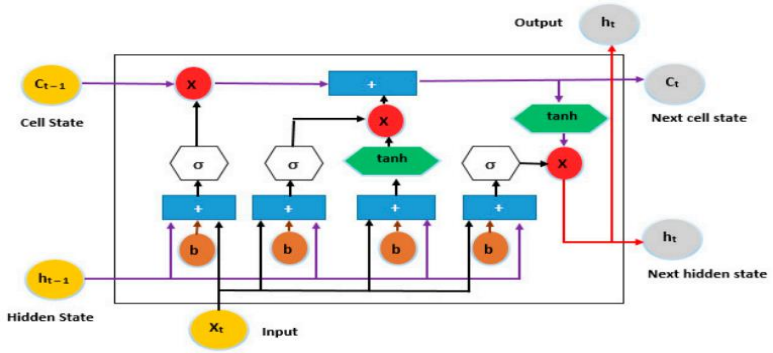
The studies reported in these publications highlight how important it is to use advanced computational techniques to predict the fake job advertisements. Through the utilization of machine learning and natural language processing techniques, these investigations exhibit the capacity to create resilient systems that can accurately discern between authentic employment prospects and deceptive scams.

IV. METHODOLOGY

We have undertaken the training of a range of supervised machine learning models using the EMSCAD dataset. This dataset, after undergoing meticulous cleaning and preprocessing, encompasses approximately 17,880 job postings. This extensive dataset provides a robust foundation for our exploration and evaluation of various machine learning algorithms in the context of identifying fraudulent job postings. The dataset is taken from Kaggle datasets.

A. LSTM- Long-Short Term Memory

Recurrent neural networks with long short-term memory (LSTM) have become an efficient and scalable solution for a variety of learning problems involving sequential data. They are effective and generic, making them valuable for capturing long-term temporal dependencies [7]. The LSTM architecture have gates like input gate, output gate and forget gate. Those three gates play a key role in governing the flow of data among the layers of the LSTM algorithm. The input and forget gate structures can modify information traveling along the cell state, with the ultimate output being a filtered version of the cell state based on context from the inputs [8]. The LSTM design has been criticized for having many components whose purpose is not immediately clear. As a result, it is unclear whether the LSTM is the best design, and it is likely that better ones exist [9]. Below figure shows the structure of a LSTM algorithm [10].



Inputs:		Outputs:		Nonlinearities:		Vector Operations:	
X_t	Current input	C_t	New Update Memory	σ	Sigmoid layer	×	Scaling of information
C_{t-1}	Memory from last LSTM unit	h_t	Current output	tanh	Tanh Layer	+	Adding information
h_{t-1}	Output of last LSTM unit				Bias		

Fig 1: structural representation of Long-Short Term Memory

The following formulas can be used to develop the LSTM's forward training process:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$i_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_i\right)$$

$$C_t = f_t * C_{t-1} + i_t * tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

$$O_t = \sigma(W_0 \cdot [h_{t-1}, x_t] + b_0)$$

$$h_t = o_t * tabh(C_t)$$

where i_t , o_t , and f_t denote the activation of the input gate, output gate, and forget gate, respectively; C_t and h_t denote the activation vector for each cell and memory block, respectively; and W and b denote the weight matrix and bias vector, respectively. In addition, σ(◦) denotes the sigmoid function

B. Bi-Directional LSTM

Bi-directional LSTM (Bi-LSTM) represents a prevalent neural network architecture frequently utilized for sequential data analysis, such as natural language processing tasks. Unlike traditional LSTM networks, Bi-LSTM incorporates information from both past and future time steps, enhancing its ability to capture temporal dependencies in the data. Each neuron in a Bi-LSTM layer receives input from both the previous and subsequent time steps, facilitating a comprehensive understanding of the input sequence. By leveraging bidirectional information flow, Bi-LSTM models can effectively capture context and dependencies in sequential data, making them particularly adept at tasks such as sentiment analysis, machine translation, and speech recognition. Training a Bi-LSTM network typically involves backpropagation through time (BPTT), where gradients of the loss function with respect to the network's parameters are computed and updated iteratively. Regularization techniques such as dropout and recurrent dropout can be employed to mitigate overfitting and improve generalization performance. While Bi-LSTM models exhibit strong predictive capabilities and excel at capturing long-range dependencies in sequential data, they require substantial computational resources and training data to achieve optimal performance.

Formulas for Bi-Directional LSTM:

$$a_h^t = \sum_{l=1}^{L} x_l^t \, w_{lh} + \sum_{h',\, t>0}^{H} b_{h'}^{t-1} \, w_{h'h}$$

$$a_h^t = \theta_h(a_h^t)$$

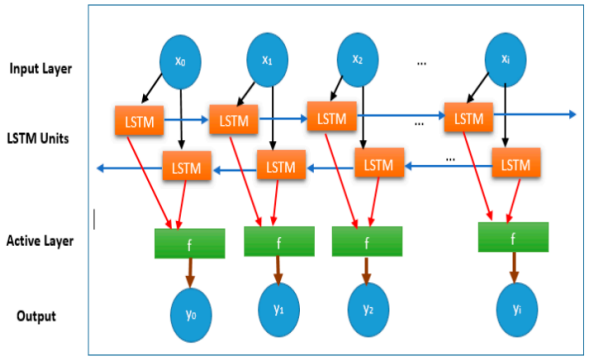


Fig 2: Bi-directional LSTM Architecture

C. Multi-Layer Perceptron

The Multi-Layer Perceptron (MLP) stands as a widely employed neural network architecture for tackling classification tasks. Comprised of numerous layers of interconnected nodes, or neurons, each node gathers input from the preceding layer, amalgamates these inputs utilizing weighted formulas, and subsequently applies an activation function to generate an output. Typically, the activation function adopts a non-linear form like ReLU, infusing the model with non-linearity and facilitating the learning of intricate relationships between inputs and outputs. Training the MLP involves backpropagation, wherein gradients of the loss function concerning the network's weights and biases are computed and then updated using gradient descent or similar optimization techniques. The model can be regularized to mitigate overfitting and enhance generalization performance through strategies such as dropout, early stopping, and weight decay. While the MLP is formidable, capable of discerning complex patterns in data and excelling at classification tasks, it necessitates substantial amounts of data and computational power for training and refinement.

v. PROPOSED METHODOLOGY

The proposed methodology uses Bi-directional LSTM as the base model to train the data. To send the data as input to the model the data must be cleaned. So, firstly the data cleaning is done using various python libraries. As we are working with sequential data the important task is to remove the punctuation marks, html tags, and stop words from the data. Now the it is crucial to have uniformly maintained data. In order to have uniformity among all the instances in the dataset the data is first converted into numerical form using one hot representation and then to main same length of that representation we are using embedding technique. Finally, the data is now consistent and can be used as an input for machine learning models. As per requirements the whole data should be divided into training and testing data. We split the data in 80:20 ratio. The 80% of the data is being used to train the model to predict the class label and rest 20% of the data is used to Test the trained model to find the efficiency of our proposed model. And based on metrics the we mentioned the build models are compared with each other namely Multi-Layer perceptron and Bi-directional LSTM.

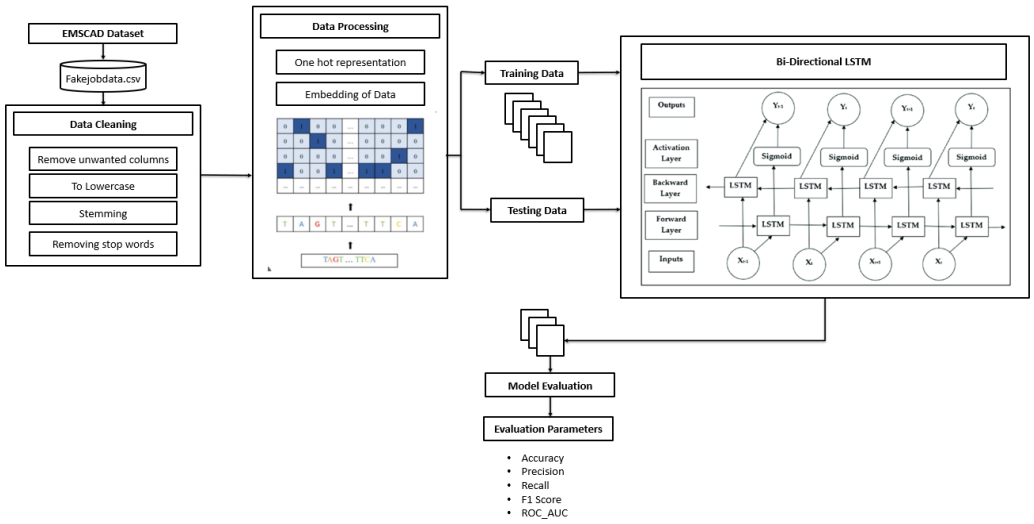


Fig 3: Architecture of proposed methodology

VI. DATASET

The dataset contains 17,880 rows and 18 columns in which 5 features (title, company_profile, description, benefits and requirements) are long texts and the rest 13 features are mainly numeric fields or categorical data. The dataset includes a Fraudulent column, with a value of 1 signifying a fraudulent job and 0 signifying a legitimate one.

Many missing values in the dataset can affect the efficiency of the model. So, it is important to perform data cleaning and data preprocessing before training the models.

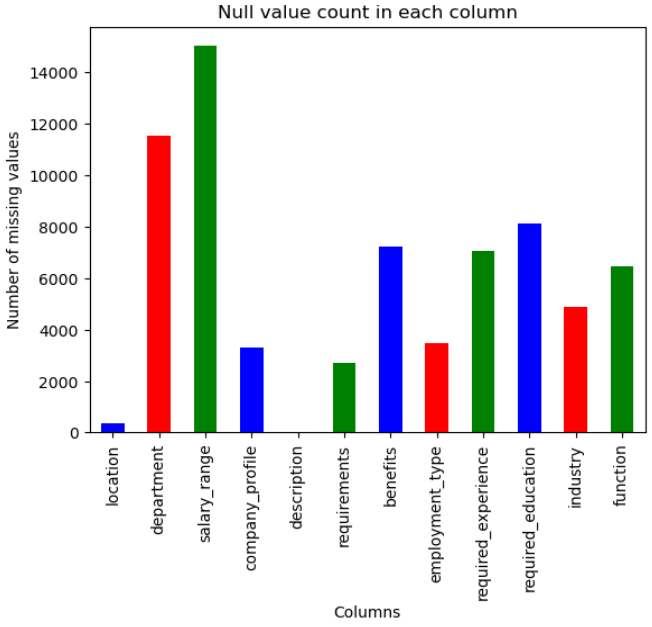


Fig:4 Simple visualization of missing values from the dataset.

Except the description column and location column most of the columns have a lot of null values. But when the location column data is observed more than 60% of the values are US.

A. Understanding the data

The schema of the dataset is as follows. There are 18 columns in the dataset in which there is one class label. The columns have textual, categorical, and numerical data. There is a class label having column name as fraudulent representing that row contains information of real job posting or fake job posting. Only some columns have consistent data. Most of the columns are having a lot of null values which do not provide much help prediction but more over it will affect the real process. So, it is important to address it and it can be done easily with help of functions in python.

	column	description	data_type
0	job_id	A unique ID assigned to each job	int64
1	title	The title of the advertised position or job	object
2	location	Information about where is the job	object
3	department	The department offering the job	object
4	salary_range	The amount that the job pays	object
5	company_profile	Information about the company advertising the job	object
6	description	Job description	object
7	requirements	Requirements enlisted	object
8	benefits	Benefits offered by the company with the job	object
9	telecommuting	1 if work from home allowed, 0 otherwise	int64
10	has_company_logo	1 if the company has a logo, 0 otherwise	int64
11	has_questions	1 if the job has any questions, 0 otherwise	int64
12	employment_type	full-time, part-time, or contract etc.	object
13	required_experience	the experience required for the job	object
14	required_education	educational requirements for the job	object
15	industry	The industry the job is in e.g engineering	object
16	function	what work is required from the applicant	object
17	fraudulent	1 if fake job, 0 if real job	int64

Fig:5 Basic Description of every column from the dataset.

B. Data Cleaning

After getting insights from the data we identified that only some columns are helpful in predicting if a job is fake or real. So, we dropped the columns which are not useful in the prediction process. We applied data cleaning process on the text data by removing HTML tags, URLs, punctuation, digits, underscores, single characters, and multiple spaces using regular expressions and some libraries are used to remove the stop words.

C. Data Preprocessing

One-hot encoding is used to represent the required columns in numerical format. It represents each unique category as a binary vector where each dimension corresponds to a category, and the value is 1 if the category is present and 0 otherwise. As the data of each instance is of different sizes, we used embedding technique to make sure that all the data is uniformly mentioned to be sent as an input to the models.

VII. MODEL TRAINING AND EVALUATION

MLP and Bidirectional LSTM are compared using various metrics like accuracy, F1 score, ROC_AUC score, and precision. We mainly worked on two columns namely title and description. As the mentioned architecture we first understood the data and performed data cleaning and data processing on those to columns and the two columns are concatenated by the features identified in them and are given as input for the models. Before them as input to the models it is split into test data and train data being in 80:20 ratio. The training data is used to train the built models and then the effectiveness of the models are evaluated using test data. At last, the metrics like accuracy, recall, f1 score, precision and ROC_AUC are calculated. Both of the models are compared on the bases of those metrics.

A. Mathematical Background

a. Confusion Matrix:

An in-depth analysis of a machine learning model's performance on a set of test data is provided via a confusion matrix. Based on the model's predictions, it shows the quantity of accurate and inaccurate instances.

True positives (TP): When a positive data point is correctly predicted by the model, this is known as a true positive.
True Negatives (TN): When a negative data point is correctly predicted by the model, this is known as a true negative.
False positives (FP): When a positive data point is incorrectly predicted by the model, this is known as a true positive.
False negatives (FN): When a negative data point is incorrectly predicted by the model, this is known as a true negative.

b. Accuracy: The percentage of correctly identified cases relative to the total number of examples is known as accuracy. The model performs better the higher its precision.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

c. Precision: The percentage of correctly identified true positive predictions among all positive predictions made by the classifier is known as precision. A high precision means that while the algorithm can detect most fake job listings, it can miss some genuine ones.

$$\text{Precision} = TP / (TP + FP)$$

d. F1 score: The harmonic mean of recall and precision is the F1 score. It offers a solitary score that harmonizes recall and precision.

$$\text{F1 Score} = 2*(\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

e. ROC-AUC measures the area under the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rate at various threshold settings Plotting the true positive rate (TPR) versus the false positive rate (FPR) yields this result. It measures the model's ability to discriminate between the positive and negative classes.

$$\text{True Positive Rate (TPR)} = TP / (TP + FN)$$

$$\text{False Positive rate (FPR)} = FP / (FP + TN)$$

VIII. RESULTS AND CONCLUSION

Detecting fake job postings early can prevent job seekers from falling victim to scams and ensure they only apply for legitimate job opportunities. In our study, we have created software for predicting fake job postings by utilizing different supervised machine learning algorithms to classify posts provided by users as either genuine or fraudulent.

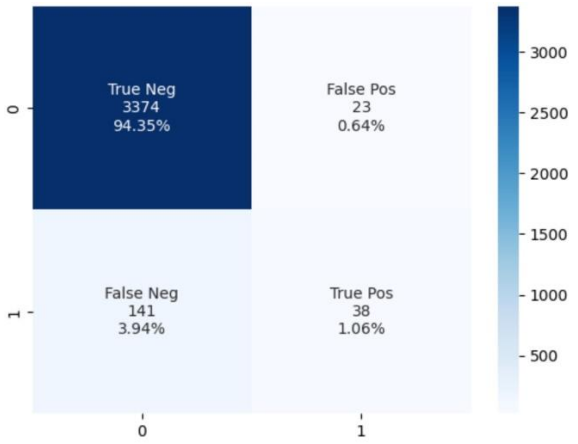


Fig 6: Confusion Matrix of Multi-Layer Perceptron

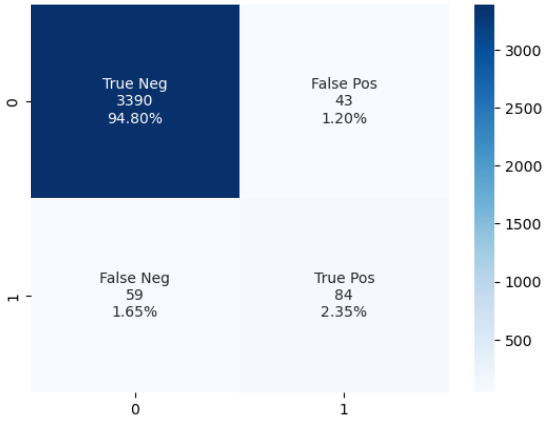


Fig 7: Confusion Matrix of Bidirectional LSTM.

By observing above confusion matrices we can say that Bi-Directional LSTM is working pretty well compared to Multi-Layer Perceptron in predicting the class label properly.

Let us compare the two models with help of visualising accuracy, f1-score, precision and roc_auc:

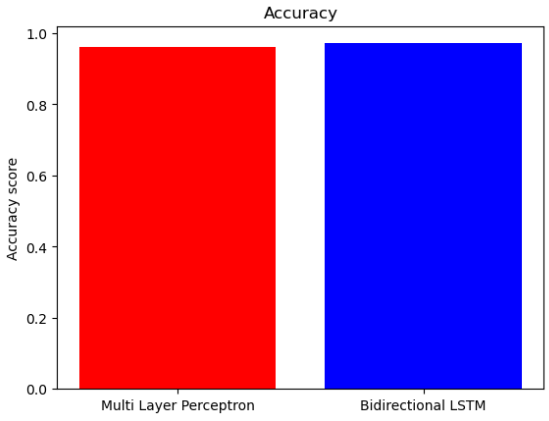


Fig:8 Comparing models based on accuracy

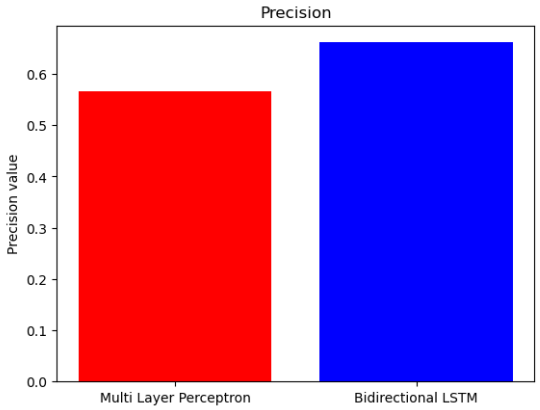


Fig:9 Comparing models based on Precision

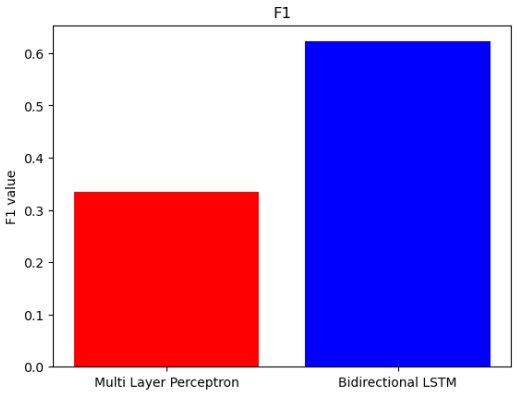


Fig:10 Comparing models based on F1 score

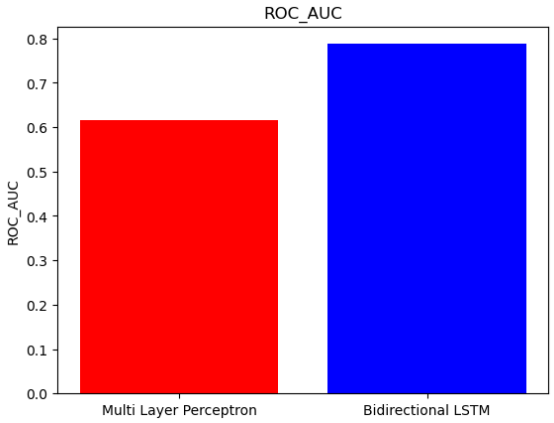


Fig:11 Comparing models based on ROC_AUC

Final comparison of MLP and Bi-Directional LSTM

Model	Accuracy	Precision	F1_score	ROC_AUC
MLP	96	0.57	0.34	0.62
Bidirectional LSTM	97	0.66	0.62	0.79

IX. FUTURE SCOPE

One of the largest challenges in detecting fraudulent job postings is class imbalance, which can lead to models that are biased towards the majority class. While the Multi-layer Perceptron and bidirectional LSTM have demonstrated promising results, future research should explore strategies for managing class imbalance more successfully, such as employing cost-sensitive learning strategies or oversampling or undersampling data. To achieve more prominent results, co-relating the dataset's columns and creating an ensemble learning can be employed, as can voting classifiers to combine those models. Domain-specific models can be prepared to improve accuracy. Using more efficient dataset will help to find more insights and relation of the job postings to predict the fake jobs.

REFERENCES

[1] M. Abbasi, T. Zubair, X. Zhou, and R. Zafarani, "Fake News: A Survey of Research, Detection Methods, and Opportunities," ACM Comput. Surv, vol. 1, 2018.

[2] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," IEEE Trans Neural Netw, vol. 10, no. 5, 1999, doi: 10.1109/72.788645.

[3] Sultana Umme Habiba,Md. Khairul Islam,Farzana 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) IEEE|DOI:10.1109/ICREST51555.2021.9331230

[4] 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

[5] William Yang Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection Department of Computer Science University of California, Santa Barbara Santa Barbara, CA 93106 USA

[6] Rashid Amen, Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges(2021).

[7] Greff, K.; Srivastava, R.K.; Koutnik, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A Search Space Odyssey. IEEE Trans. Neural Netw. Learn. Syst. 2017, 28, 2222–2232. [CrossRef]

[8] Sherratt, F.; Plummer, A.; Iravani, P. Understanding LSTM Network Behaviour of IMU-Based Locomotion Mode Recognition for Applications in Prostheses and Wearables. Sensors 2021, 21, 1264. [CrossRef] [PubMed]

[9] Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of Recurrent Network architectures. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; Volume 3, pp. 2332–2340.

[10] Le, X.-H.; Ho, H.V.; Lee, G.; Jung, S. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. Water 2019, 11, 1387. [CrossRef]

[11] Yang, S. Research on Network Behavior Anomaly Analysis Based on Bidirectional LSTM. In Proceedings of the IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 15–17 March 2019.