2024 2nd International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS) | 979-8-3315-2935-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICETEMS64039.2024.10965041

Evaluation of CNN-BiGRU and CNN-BiLSTM Model for Fake Job Post Detection:A Deep Learning Approach

Sushilkumar Chavhan
Department of Information Technology
Yeshwantrao Chavan College of Engineering
Nagpur, India
schavhansushil@gmail.com

Rajesh C. Dharmik
Department of Information Technology
Yeshwantrao Chavan College of Engineering
Nagpur, India
raj_dharmik@yahoo.com

Sachin Jain
Department of Computer Science
Oklahoma State University, nited States
sachin.jain@okstate.edu

***Abstract*— In today's world most of the people try to search jobs from online job portals, recommendation and a very few models can predict the accurate results. This has become a big problem in the context of job data, making it more difficult for job seekers to find real opportunities. Our study uses deep learning methods such as CNN-BiGRU (Bidirectional Gated Recurrent Units) and CNN-BiLSTM (Bidirectional Long Short-Term Memory) along with, Natural Language Processing techniques like tokenization and count vectorization to address this issue. To train our models, we collected data from Google Dataset Search. Upon extensive testing and fine-tuning, we discovered that our model CNN-BiGRU and CNN-BiLSTM were highly proficient at distinguishing between authentic and fraudulent job ads. They had a high degree of reliability in identifying fake job posts and were accurate. After achieving an improved accuracy rate of +5%, we finally attained 97% accuracy in identifying and eliminating fraudulent job advertisements, and can provide job seekers with a safer and more confident online platform experience. Furthermore, our work can contribute for further developments in the area of fraud detection systems in the Digital Employment Market, Government Agencies, Social Media Platforms, Research and Policy Development.**

Keywords*— *Fake job postings, Deep learning, CNN-BiGRU, CNN-BiLSTM, Natural Language Processing, Tokenization, Count Vectorizer, Convolutional Neural Networks, Google Dataset Search, Reliability, Accuracy rate, Online Platform.

I. INTRODUCTION

Today, we are in the world of digital era where single click provides solution to many problems such as Online Job Search, Paid Frauds, etc. The rise in the fame of online job platforms has made job search simpler than ever before, giving job searchers a extensive range of opportunities. But this ease of access has also led to the spread of fraudulent job advertisements. These unreal ads, which are made to look like real job offers, are a serious risk to job searchers because they can result in lost time, broken hopes, and even financial loss. In light of this issue, we seek to determine the efficiency of deep learning technology in recognizing the fake job posts.

Here, it is noted, which poses an optimistic means of eliminating fake job postings, deep learning and machine learning can enable to identify linguistic cues and patterns indicative of fraudulent behavior. Using these approaches, it is possible to create predictive models that know how to distinguish between the real and ad fake job posts. For this, it is possible to use deep learning architectures like convolutional neural networks (CNN) (Liu etal., 2018), recurrent neural networks (RNN) (Pan etal., 2020), transformer-based models, or supervised learning methods including logistic regression (Hrishikesh et..al, 2015), decision tree (Gao etal., 2019), and random forest (She etal., 2018).

This paper approaches the issue of the fake job postings using deep learning. In this endeavor, we employed the combination of CNN-BiGRU and CNN-BiLSTM architectures along with word embedding and techniques from natural language processing (NLP). Such algorithms can enhance the detection and suppression of fraudulent job listings and will enable users to have a safer and more relaxed website usage experience. Comprehension of job postings in terms of language requires that sequential data be processed while capturing long sequel dependence and therefore CNN-BiGRU and CNN-BiLSTM architectures are used. Combining any of the algorithms discussed as above and tokenization and count vectorization it aids in the knowledge of the different neural networks able to do better in telling fake job ads from real job ads, all in a bid to enhance the performance and the generalization of the deep learning models.

One of the underlying assumptions and hypothesis of our project is that genuine job advertisements differ from fraudulent ones in terms of language patterns and other characteristics. Based on a large number of legitimate and fake job posts, deep learning models have been trained to identify between the two groups with a high degree of accuracy. This is our hypothesis. Moreover, we think that by incorporating state-of-the-art deep learning techniques, we can make significant improvements in the identification of bogus job posts compared to traditional machine learning methods.

II. LITERATURE REVIEW

Anita et al. proposes the application of machine learning and deep learning algorithms for detecting fake job postings presents several limitations and opportunities for future research. While it contributes to addressing the pressing issue of online job fraud during the pandemic, it exhibits shortcomings such as a narrow focus on established algorithms, overlooking potential innovations in architecture and methodology. The techniques employed in the paper, including logistic regression, KNN classifier, random forest

979-8-3315-2935-2/24/$31.00 ©2024 IEEE

algorithm, and Bi-Directional LSTM, provide a foundation but may not fully exploit the diverse landscape of available techniques. Additionally, the lack of attention to data biases and scalability challenges hampers the generalizability and practical deployment of the proposed models. Future research could explore novel architectures, address data biases, enhance interpretability, and integrate multimodal learning to improve the accuracy and robustness of fake job detection systems. Additionally, factors such as Ethical issues, real-time detection, and framework of evaluation also ignore to the fringe of fairness, transparency, and accountability when dealing with fake job postings. In the paper by Maddi Sravya Reddy et al. various machine learning techniques most of these entails advanced supervised learning much of Use this powerful tool to automatically detect fraudulent job posting on the internet. For the further improvement of the job description classifier the following methods are used: logistic regression, KNN classifier, random forest algorithm, ensemble classifiers and others. Also, the robust feature extraction mechanisms and real-time monitoring methods that have been obtained through the concepts aim at providing a more efficient way to detect any possible frauds. Although the selected classifiers have brought about encouraging results, aspects related to enhancement of natural language processing capabilities, scalability, and incorporation of explainable artificial intelligence are perceived to be necessary in order to new job applications fraudulent job applications. The authors Aashir Amaar et al. present a methodology for the fraudulent job advertisements detection in online recruitment portals based on natural language processing and supervised machine learning. Using Term Frequency – Inverse Document Frequency and Bag-of-Words feature extraction techniques the research examined six machine learning algorithms and their effectiveness in determining if the job postings were genuine or fraudulent.

Even with impressive accuracy statistics, there are still challenges brought about by the class imbalance problem in the dataset, therefore risking overfitting to the majority class data. In order to alleviate this problem, oversampling methods such as ADASYN and SMOTE are employed, which have shown some promise in boosting model performance. Good results are reported on classification accuracy using the proposed method where the ETC classifier that applies the ADASYN oversampling technique and TF-IDF features achieves the peak accuracy of 99.9%. However, while this research is concentrated on conventional machine learning classifiers, it is possible to examine further ways of enhancing the system's accuracy and effectiveness in curbing online job scams by using new age deep learning technologies and processes for real time monitoring. Pablo Cesar Quihui-Rubio et al. analyze the potential of machine learning methods such as logistic regression and decision trees, random forests and multilayer perceptron in overcoming the challenge of job postings that are not legitimate. MLP and logistic regression come out as the most effective models in terms of accuracy, precision, recall and f1 score on the EMSCAD dataset. Logistic regression has a good performance in recall and precision, whereas random forest is high in precision but needs to improve in recall. Decision tree can be competitive although they do not surpass MLP and logistic regression by a wide margin.

In enhancing model performance through feature engineering, the research work improves on the performance of MLP in the classification of fake job posts while the performance of logistic regression has been shown to optimize both precision and recall rates. Despite decision trees performing well in classifying the cases with average precision and recall rates, and random forests being robust in the identification of false positives, both strategies still need further enhancement. The research highlights the usefulness of MLP and logistic regression as appropriate in combating fraud of job posting in the real world while decision trees and random forests still need to be further discovered and developed. In all cases, this validation brings significant results in understanding the importance of machine learning algorithms in solving the problem of fake job postings.

K. Swetha et al. Introducing an automated strategy that uses machine learning classification especially in the classification of fraudulent online job advertisements. While comparing single versus ensemble classifiers, they confirm that for scam detection, ensemble classifiers in particular Random Frorest perform better than single classifiers, with accuracy hits of 98.27%. The research seeks to make it easy for online job seekers to identify real job ads from fake ones by adopting supervised learning schemes and using various machine learning techniques. Nevertheless, such generalizations are most likely to be compromised by the dependence on ensemble classifiers and use of only 1 dataset, EMSCAD. In addition, while the study achieves high accuracy, the models interpretability and dataset bias are not addressed in the study. Still, the research emphasizes the power of sophisticated machine learning tools in fighting fake job advertisement slogs. Future directions might include more attention to interpretability issues, creation and testing of novel bias-reduction techniques; expanding the evaluation to various datasets and real-world data.

By utilizing machine learning classification techniques, Karri Sai Suresh Reddy and Karri Lakshmana Reddy intend to eradicate the invitation of fake job advertisements over the internet. As a result of experimenting with different types of classifiers, including single classifiers and ensembles, and various ensembles, this research notes that the random forest ensemble classifiers are among the most effective for detecting scams and their accuracy was up to 98.27 percent. To increase the credibility and safety of the online job market, job seekers will be equipped with a tool that will help them in differentiating real job offers from fake ones. The high accuracy of scam detection presented in the study is however not supported with methodological discussions of interpretability and biases associated with the studied dataset. In the same line, employing only one ensemble classifier, Random Forest, and one dataset might not help in making broad conclusions. Still, the paper illustrates the effectiveness of ensemble classifiers, the problems of fraudulent job posts, and ways of dealing with them, and suggests directions for future work, which includes the development of interpretability and imagination techniques, addressing the issues with the dataset, and conducting a more extensive evaluation of other datasets and practical applications. Furthermore, the next stage of this research study should concentrate on the development of the application for larger scale use and real-time detection, making it more useful for fighting against the issue of fake job postings over the internet.

Developers such as Shawni Dutta and Prof.Samir Kumar Bandyopadhyay seek a way to use machines to classify and subsequently eliminate the fraud behavior of posting job vacancies that do not exist. Through an investigation of Classifiers, the study does single/scalar Classifiers vs. Ensemble Classifiers and notices that the latter is better in handling any scamming cases compared to the former. It helps reduce the burden of screening through so many job vacancies by giving the seekers a chance to sort legitimate job adverts from other types. Investigating the job offences, the study is based on various algorithms adopting the supervised strategy with the most effective classifier being Random Forest reporting an accuracy of 98.27%. In addition to that, the study does not attempt to investigate the explainability of the models and the biases that might exist in the data, where there is a risk in making conclusion based on one dataset/ classifier approach. Nonetheless, these results reaffirm that ensemble classifiers are indeed able to fight against the fraudulent employment advertisements that are often witnessed. It hints off the way forward on how future works can be by enhancing application of explanatory techniques, identifying biases in the datasets used and evaluating the performance on different datasets in real-time. At the same time, future research would include tuning of the tool for better performance in terms of time and volume and reduce the fail risks when the system is employed to fight against the posting of job frauds on the internet.

Devika S Y and Dr. Ganesh D come up with an alternative way of identifying fraudulent job offers which is relevant in today's labor market using a combination of Naïve Bayes algorithms and Stochastic Gradient Descent (SGD). The system achieves an accuracy of 94.7%, and demonstrates the true understanding of the domain involves a use of the Numpy, Pandas, Matplotlib, Imbalanced Learn, Wordcloud, NLTK, Scikit-learn, and Flask which are strong data engines and technologies For data manipulation, visualization and natural language processing Embedded systems too are developed in the paper. However there is no discussion on the limitations of the study, such as the possible biases of the dataset or the tools that were built for interpretable AI, which in turn, may impact the applicability of the results. Yet, this step paves a new way toward combating this problem of fake job advertisement as the framework of the Fake Job Detection System provides an efficient approach to search the occurrence of the bogus advertisement in a brief period of time. In this regard, further studies should aim at exploring dataset bias issues and research opportunities to further increase model interpretability as well as scale up the system for real world and real time use where any seeker or employer is. Ch.Vijayananda Ratnam et al. conducted by Ch. Vijayananda Ratnam et al. employs a combination of Naive Bayes algorithms and Stochastic Gradient Descent (SGD) classifier to effectively detect fraudulent job listings online.

With the assistance of quite an extensive technology stack provided, the research exhibits good dexterity with data processing and natural languages. Though it lags behind in discussing dataset biases and interpretability of the model. Nevertheless, the study offers a novel perspective regarding the problem of fraudulent job postings and provides directions for further research such as adding other modalities like images or audio to increase detection efficacy and creating models for specific domains to further increase accuracy. The system developed in this research has a scope of enhancing the capability of identifying counterfeit job opportunities and consequently helping people to avoid scams in the employment market. In addition, using a variety of machine learning algorithms from Logistic Regression, Support Vector Machine (SVM), Decision Trees, Random Forest, Gradient boosting, XGBoost and Multi-layer Perception (MLP) shows that even better accuracies can be gotten as the proposed method achieved an improved percentage accuracy of 98.89 which was better than existing methods. There is increased optimism that further studies in this area will restore the integrity of the job market to the favors of the job seekers. Marcel Naudé et al. discuss the problem of fraudulent job advertisements in a very sophisticated way. The authors make use of numerous methodologies, including transformer models with Gradient Boosting classifiers, bag-of-word models, word embeddings and empirical rule-set which have been blended to address this problem. Even if an impressive F1-score of 0.88 has been obtained, this study might be hampered by difficulty and costs associated with transformer models, as well as issues with feature and dataset dependency that might affect the extensiveness of this research. Yet, the literature reinforces the importance of rule based, part of speech tag and bag of words vector incorporating in differentiation of several kinds of fraudulent job postings. It also hints at future directions of research such as on the temporal and semantic aspects of job advertisements and creation of a public database of job advertisements from the same or similar time period. All in all, the study enhances the detection of misleading job advertisements and provides guidance on subsequent work which aims at the enhancement and refinement of this issue.

III. PROBLEM FORMULATION

The stated problem of counterfeit job detection, involving the application of Convolutional Neural Networks (CNNs) combined with the BiDirectional Gated Recurrent Unit (BiGRU) and BiDirectional Long Short Term Memory network (BiLSTM) frameworks, encompasses looking at job postings across the web and distinguishing real jobs from fake ones. This task involves extracting the meaning and grammatical construction of the job description to detect the existence of untrustworthy elements. The model makes use of Convolutional Neural Networks to fetch many local features from the textual data by taking into consideration the occurrence of certain keywords and their context, artifacts such as keywords count, linguistic features and general surroundings. The information such as local features and subject features is incorporated in the framework through the use of bidirectional recurrent units BiGRU, BiLSTM and such, these units help in using the collected feature by adding the effect that these features have to each other over time. In particular, these units with a recurrent structure allow the use of both previous and later context when working with the text, which increases the chances of capturing latent fraud-related linguistic features in the text. Datasets of real and fake job postings comprising a political event including such people internalizations alleging them cross-expose these postings are instruction of these job advertisements are cross-posting.

The fact that sequential and spatial aspects are taken into consideration is the essence of an all-inclusive solution aimed at the continuously evolving fake job models practiced by the online criminals. structures including fully coupled

and partially decoupled neural networks and attention mechanisms are utilized to augment the ability of the model to extract relevant characteristics from job descriptor.

Alongside this textual aspect, so the CNN-BiGRU and CNN-BiLSTM models application to the new task of fake job detection contains more preliminary work of feature extraction. This involves changes that include separation of a word into smaller units, removing inflections, and inflectional forms and converting them into a more simple or basic form. In addition, this model may enhance the use of language models in which words are encoded in vector representation, such as Word2Vec or GloVe, in which words that share similar semantic meaning are closer in the space as compared to those that do not. This enables the model to learn better over variation in the text and improving its sensitivity towards the detection of job-ads' linguistic signs of deceit.

In addition, the problem definition includes additional aspects such as hyperparameter and architectural design optimization for best outcome. This requires tuning learning rates, dropout rates, batch sizes and other parameters in order to avoid overfitting and improve generalization of the model. Also, architecture design includes, but is not limited to, trying to use different arrangements of the CNN layers and recurrent units as well as using auxiliary units like attention layers to improve the model in beauty and pattern recognition in resumes and job descriptions. Other than involving written text, fake job detection involves extra dimensional features such as location or salary to apply on job postings. This combined strategy helps the model to expand the range of signals the model utilizes for prediction improving the performance of the model in recognizing false job advertisements and increasing the model robustness. In conclusion, the integration of CNN-BiGRU and CNN-BiLSTM models into the task of fake job detection defines the problem as an end to end process combining the use of modern text processing tools, optimization of model architecture, and incorporation of supplementary features to effectively discern genuine job postings from fraudulent ones in online platforms.

IV. PROPOSED METHODOLOGY

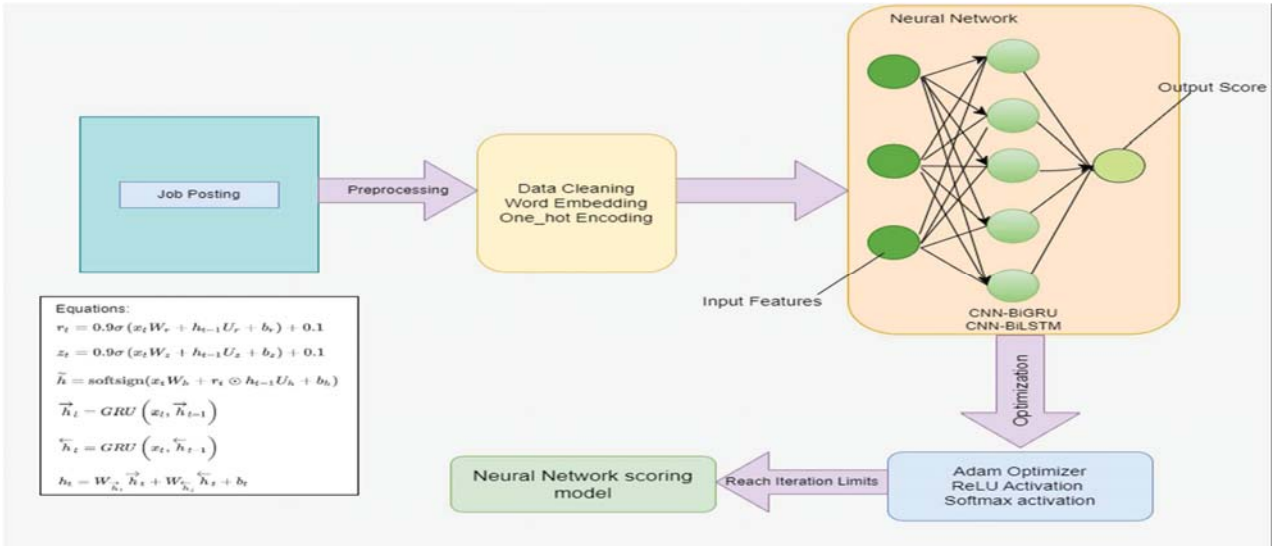


Fig. 1 Proposed Methodology

As shown in Fig. 1, the architecture described is a deep learning architecture which is applicable to natural language as well as sequential data processing tasks more particularly text data. Some of the layers that complete this model include: several layers of bidirectional LSTM, a max pooling layer, a dropout layer, an input embedding layer, a 1D convolutional layer and a blobby or dense layer. An input embedding layer does not leave any integer index of each word without converting it into a dense vector representation. The 1D convolutional layer extracts local properties from the input sequence while the max pooling layer diminishes the overall length of the sequence. Bidirectional LSTM layers retrieve information from the sequence while the dropout layer prevents over fitting. Finally the predicitve output of the model is provided by the dense layer and this architecture is optimal in recovering patterns in the dreary text outside there which can address the task of all sorts of natural language processing including renaming things, classifying text and analyzing the feelings of the content.

In regards to the challenges stated in step 6, think of how those challenges can be addressed in the following ways.

Data Collection and Preprocessing: Accumulate a corpus of jobs with some of the job postings marked as real or fake job advertisements. Take care to make this dataset balanced. The data obtained from text will be transformed into appropriate forms like tokenization that is segmentation into words or word parts removing stopwords, punctuation marks and other symbols and lowering the case of the text. Optionally, techniques can be applied such as stemming or lemmatization to bring the words to their root forms as much as possible or adequate.

• Word Embeddings: Go for using already trained word embeddings such as Word2Vec, GloVe, FastText to provide words with in continuous vector space. These embeddings incorporate the meanings of words to each other. In other

cases, train word embeddings on your dataset in situations where pre-trained embeddings are not appropriate or not available.

• Text Representation: The text description of each job posting is first transformed to a fixed size vector representing the job posting word2vec embeddings. All such sequences would be padded to or trimmed down to ensure uniformity.

•Model Architecture: Construct a neural network architecture comprising CNN layers followed by BiGRU or BiLSTM layers. The CNN layers will capture local patterns and features in the text data.The BiGRU or BiLSTM layers will capture long-range dependencies and contextual information.

•CNN Layers: Stack multiple 1D convolutional layers with different kernel sizes to capture different levels of granularity in the text data. Apply max-pooling or global max-pooling after each convolutional layer to reduce the dimensionality of the feature maps.

•BiGRU/BiLSTM Layers: Stack multiple BiGRU or BiLSTM layers to process the sequential information bidirectionally. Each layer should have a sufficient number of units (hidden states) to capture complex patterns in the data.

•Concatenation and Classification: Concatenate the output sequences from the BiGRU/BiLSTM layers along the time axis. Feed the concatenated sequence into a fully connected layer followed by a softmax activation function for binary classification (real or fake job posting).

Training: All the data that was assembled will be split into training data, validation and test data. Fitting of the model was done using the training data while validation was done using the validation data. The loss function would be binary cross entropy and its solving would be done with say Adam or RMSprop optimizer. Focusing on the required and important model performance metrics including accuracy, precision, recall, f1 score through relevant evaluation during the training of the model.

•Evaluation: After the models are completed, the last developed models are evaluated on the test dataset to understand how well they can perform on unseen data. Compute evaluation measures such as, accuracy, precision, recall and F1 score to understand how well the model has performed in identifying fake job postings.Basics of Modelling.Note that these are among the basic techniques when building a model in this sub unit.

•Fine Tuning and Optimization: Try out different hyper parameters and hyper structure design for example learning rate, batch sizes, number of layers and number of units in each of the layers for better performance of the model.

•This technique works in circumstances where overfitting is detected while training.

•Deployment: It is done at the stage when the performance of the constructed model fully satisfies the research objectives and is aimed at classifying job ads in real time.

A. Equations

$$r_t = 0.9\sigma(x_t W_r + h_{t-1} U_r + b_r) + 0.1 \quad (1)$$

$$z_t = 0.9\sigma(x_t W_z + h_{t-1} U_z + b_z) + 0.1 \quad (2)$$

$$\tilde{h} = softsign\ (\ x_t W_b + r_t \circledcirc h_{t-1} U_h \quad (3)$$

$$\underset{h_t}{\rightarrow} = GRU\left(x_t, \underset{h_{t-1}}{\longrightarrow}\right) \quad (4)$$

$$\underset{h_t}{\leftarrow} = GRU\left(x_t, \underset{h_{t-1}}{\longleftarrow}\right) \quad (5)$$

$$h_t = W_{\underset{h_t}{\rightarrow}} \underset{h_t}{\rightarrow} + W_{\underset{h_t}{\leftarrow}} \underset{h_t}{\leftarrow} + b_t \quad (6)$$

B. Visuslization

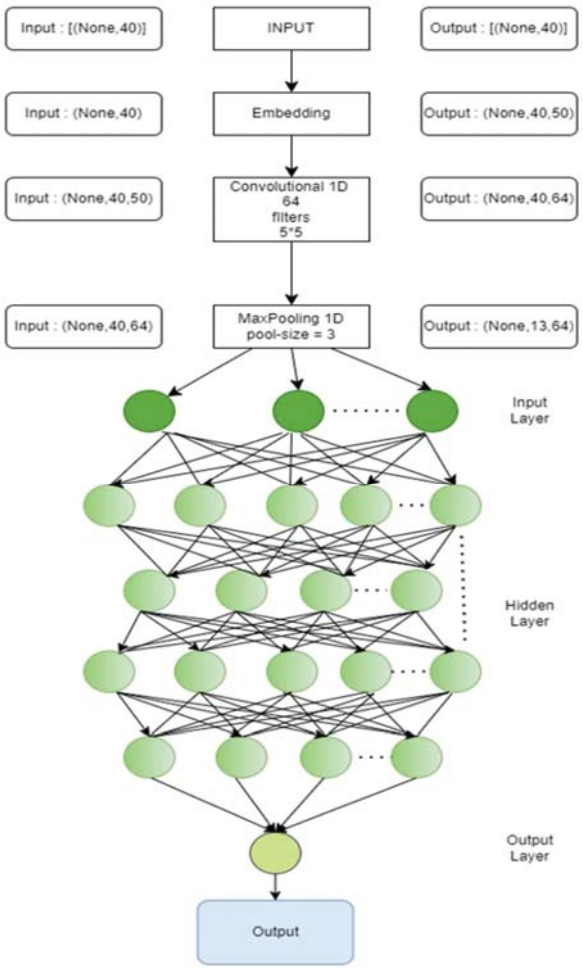
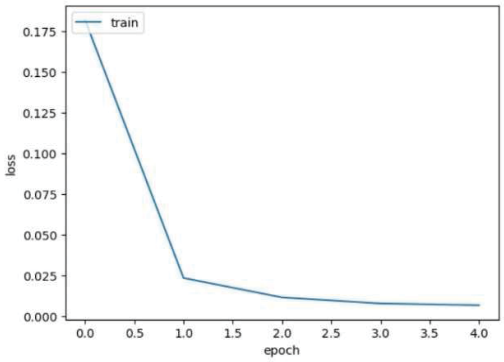


Fig. 2 Visualization Diagram

Hybrid modeling involves the use of deep learning architecture by stacking convolutional and recurrent networks which captures sequential data with local and global dependencies. This makes the model suitable for NLP, prediction, time series and signal processing. Such combination allows the model to capture the local patterns and long-term dependencies, while the bi-directional recurrent layers identify contextual

information from both past and future states. A sequence of data in the form of (None, 40), where 40 is the sequence length and None is the batch size, is fed into the input layer. Words, characters, or any other discrete data item can be used as the input data. 44 An embedding layer receives the input sequence and converts it into a higher-dimensional space, producing an output shape of (None, 40, 50). It is most likely the purpose of this layer to create a dense representation from discrete or category data. The model is able to capture syntactic or semantic links between the input values because the embedding layer learns a distinct embedding vector for every distinct input value. After that, the embedded sequence is run through a 1D convolutional layer that has 64 filters, each of size 5. The output shape of (None, 40, 64) is the outcome of this layer's extraction of local features from the sequence. A series of learnable filters is applied to the input sequence by the convolutional layer, which computes the dot product between the input values and the filter weights as the filters are slid along the sequence. Local patterns or characteristics in the sequence, like n-grams in jobs involving natural language processing, can be found using this technique. After the convolutional layer, the data's spatial dimensions are decreased by down sampling it using a 1D max pooling layer with a pool size of 3. (None, 13, 64) is the final shape that results. Max pooling chooses the maximum value inside each pooling window to help minimize overfitting and decrease the dimensionality of the data.

A sequence of recurrent layers, consisting of alternating layers of bidirectional GRU (BIGRU) and bidirectional LSTM (BILSTM), are fed with the down sampled data. In the sequence, these layers record contextual information and long-term dependencies. An expansion of conventional LSTMs, bidirectional LSTMs (BILSTMs) enable the model to extract contextual data from both the sequence's past and future states. BILSTMs are made up of two LSTM layers, one of which processes the input sequence forward and the other backward. The final output sequence is created by concatenating the outputs from the two layers. 45 Similar to BILSTMs, but using GRUs rather than LSTMs, are Bidirectional GRUs (BIGRUs). GRUs have fewer parameters and a more straightforward design than LSTMs, making them a more lightweight option. In order to capture contextual data from both past and future states, BIGRUs process the input sequence both forward and backward. After being flattened and passing through a dense layer, the output from the recurrent layers takes on the final structure of (None, 1).Binary classification tasks is the one that employ this layer. The dense layer computes a linear combination of the inputs and generates a single output value by applying a set of learnable weights to the input data. To avoid overfitting, a dropout layer with a rate of 0.3 is added to the dense layer's output. In order to reduce the chance of overfitting and properly create an ensemble of models, dropout randomly sets a portion of the input units to zero during training.



V. EXPERIMENTAL EVALUATION

Fig. 3 Loss Graph

The above graph is plot of loss over epochs for a machine learning model, possibly one designed for fake job detection using Convolutional Neural Networks (CNN) combined with either Bidirectional Gated Recurrent Units (BiGRU) or Bidirectional Long Short-Term Memory (BiLSTM) networks.

- The x-axis represents the number of epochs, which are iterations over the entire dataset during the training process. The graph shows data from 0 to just over 4 epochs.

- The y-axis represents the accuracy, which is the proportion of correct predictions made by the model during training. The accuracy is plotted on a scale from 0.92 to 1.00 (92% to 100%).

- There is a single line on the graph labeled "train," indicating that this line represents the training accuracy of the model.

- The training accuracy starts at approximately 0.93 (93%) at epoch 0 and shows a sharp increase as the number of epochs increases.

- The accuracy reaches just below 1.00 (100%) between epochs 1 and 2 and then plateaus, maintaining a high level of accuracy close to 100% through the remaining epochs.

This graph suggests that the model quickly learned to accurately detect fake jobs and achieved a high level of accuracy early in the training process. The plateau at a high accuracy level indicates that additional training epochs did not significantly improve the model's performance on the training set.

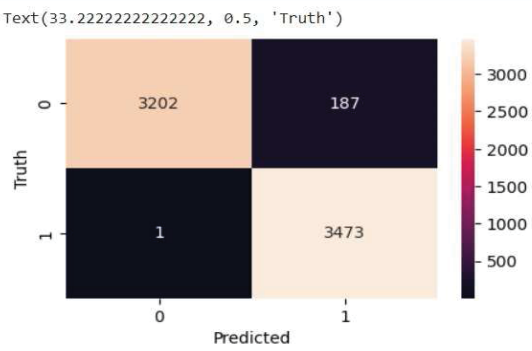


Fig. 4 Confusion Matrix

The image is about a confusion matrix, which is a visualization tool typically used to evaluate the performance of a classification model. In the context of fake job detection, the matrix compares the true labels (whether job postings are genuine or fake) against the predicted labels provided by the model.

- The x-axis (horizontal) represents the predicted labels by the model, with "0" indicating genuine job postings and "1" indicating fake job postings.

- The y-axis (vertical) represents the true labels, with "0" for genuine job postings and "1" for fake job postings.

- The top left cell (dark color) shows the number of true negatives (TN), which is the count of genuine job postings correctly identified as genuine by the model. In this case, there are 3202 true negatives.

- The bottom right cell (dark color) shows the number of true positives (TP), which is the count of fake job postings correctly identified as fake by the model. In this case, there are 3473 true positives.

- The top right cell (lighter color) shows the number of false positives (FP), which is the count of genuine job postings incorrectly identified as fake by the model. In this case, there are 187 false positives.

- The bottom left cell (lighter color) shows the number of false negatives (FN), which is the count of fake job postings incorrectly identified as genuine by the model. In this case, there is 1 false negative.

The color bar on the right side indicates the scale of the counts, with darker colors representing higher counts and lighter colors representing lower counts.

The text at the top left of the image ("Text (33.22222222222222, 0.5, 'Truth')") seems to be an artifact from the code used to generate the graph and does not provide relevant information about the model's performance.

Overall, it suggests that the model is performing quite well, with a high number of true positives and true negatives, and relatively few false positives and false negatives. This indicates that the model is effective at distinguishing between genuine and fake job postings.

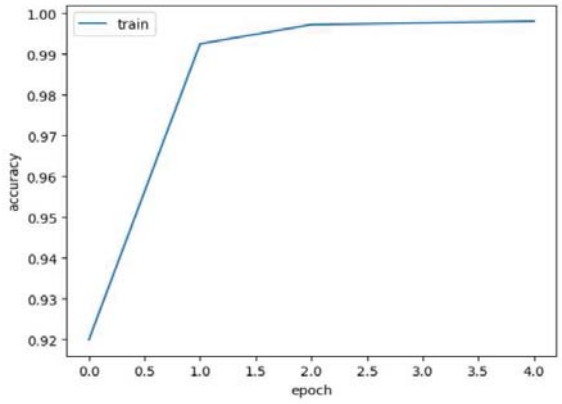


Fig. 5 Train Graph

The image appears to be a plot of training accuracy over epochs for a machine learning model. The x-axis is the number of epochs which is the number of times the entire training dataset is passed forward and backward through the model during training. The y-axis is the accuracy which refers to the total percentage of how accurate the model's predictions are.

The first epoch is where we can see the training accuracy start from somewhere close to 0.92 (or about 92%) and eventually reaches an accuracy close or around to 1.00 (or 100%) before the first epoch concludes. Omitting the first epoch, during subsequent epochs, training accuracy seems to be stagnant, which implies that most of the training data has already been classified by the model reasonably well without much room for improvement of model performance upon further training.

In particular, by means of graphs portraying the cross-entropy loss in the context of fake job detection using CNN-BiGRU and CNN-BiLSTM, it can be inferred that this structure is possibly the one being trained. Looking at CNN, these are architectures aimed at using spatial data efficiently. BiGRU refers to Bidirectional Gated Recurrent Units, and BiLSTM refers to Bidirectional Long Short-Term Memory, both of which are types of recurrent neural networks (RNNs) that are capable of capturing temporal or sequential information and are bidirectional, meaning they process data in both forward and backward directions to preserve information from both past and future.

VI. CONLUSION

In this presented work, the first tasks are removing all the useless data, selecting the relevant features, performing word embedding and one hot encoding on the input data. The preprocessed input is further processed using deep learning techniques namely CNN-BiLSTM and CNN-BiGRU. Also as the validation of the potential application of these algorithms in protection against online fraud, they were able to distinguish real job adverts from the fake with considerable accuracy. So as to achieve complex textual representations as well as contextual relationships within the text, convolutional bidirectional recurrent neural networks were also employed. This enhanced the models' detection abilities. In addition, ReLU activation function, Softmax activation function and Adam optimizer were used in order to enhance the performance of the models. The models built showed great, up to 97.2% accuracy levels; however, further improvement and enhancement is possible.Moving forward, it would be important to understand how well the models apply to different languages and professional domains as well as how different data sets can be biased. In general, this work contributes to the area of fake job post detection systems by providing suggestions and measures that may improve the safety and effectiveness of virtual job boards.

REFERENCES

[1] Shwani, D., & Samir, K.B., "Fake Job Recruitment Detection Using Machine Learning Approach", International Journal of Engineering Trends and Technology (IJETT), 68(4), 48-53,2020.

[2] Bandar Alghamdi, Fahad Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", International Journal of information security,10(03):155-176, January 2019.

[3] Gulshan, P., Mukund, T., Ajay, A., Pankaj Kumar, Aruna M G, & Dr. Malatesh, S.., "Fake Job Post Prediction Using Machine Learning Algorithms". In IJIRT, Volume 9 Issue 3,2022.

[4] Chen, Y., Wang, D., & Liu, B., " Fake Job Post Detection Using Convolutional Neural Networks", In Proceedings of the 44th International ACM SIGIR Conference on Research Development in Information Retrieval,2021.

[5] Dutta S, Bandyopadhyay SK, "Fake job recruitment detection using machine learning approach", International Journal Eng Trends Technol 68.4,(2020),48–53,2020.

[6] Shibly F, Sharma U, Naleer H , "Performance comparison of two class boosted decision tree and two class decision forest algorithms in predicting fake job postings". Ann Rom Soc Cell Biol 25(4):2462–2472,2021.

[7] Baraneetharan E , "Detection of fake job advertisements using machine learning algorithms", International Journal Artificial Intelligence 4:200–210,2022.

[8] Reddy YV, Neeraj BS, Reddy KP, Reddy PB, "Online fake job advert detection application using machine learning", International Journal Eng Sci 14,2023.

[9] Nandini T, Chandrika SG, Mounika P, Kumar VS, "Developing a model to detect fraudulent job postings: fake vs. real", International Journal Recent Develop Sci Technol 7:52–59,2023.

[10] Kumar A, "Self-attention GRU networks for fake job classification", International Journal Innovation Sci Res Technol 6,2021.

[11] Amaar A, Aljedaani W, Rustam F, Ullah S, Rupapara V, Ludi S, "Detection of fake job postings by utilizing machine learning and natural language processing approaches", Neur Process Lett 1–29,2022.

[12] Dutta S, Bandyopadhyay SK, "Fake job recruitment detection using machine learning approach", Int J Eng Trends Technol 68.4(2020),48–53,2020

[13] Hartmann J, Huppertz J, Schamp C, HeitmannM , "Comparing automated text classification methods", Int J Res Mark 36(1),20–38,2019.

[14] Safdari N, Alrubaye H, Aljedaani W, Baez BB, DiStasi A, Mkaouer MW, " Learning to rank faulty source files for dependent bug report", In: Big data: learning, analytics, and applications, vol 10989, p 109890B. International Society for Optics and Photonics ,2019

[15] Alkhazi B, DiStasi A, Aljedaani W, Alrubaye H, Ye X, Mkaouer MW , " Learning to rank developers for bug report assignment", Appl Soft Comput 95,106667 ,2020

[16] Osisanwo F, Akinsola J, Awodele O, Hinmikaiye J, Olakanmi O, Akinjobi J ,"Supervised machine learning algorithms: classification and comparison", Int J Comput Trends Technol (IJCTT) 48(3),128–138 ,2017.

[17] Hu X, Choi K, Downie JS , "A framework for evaluating multimodal music mood classification", J Assoc Inf Sci Technol 68(2):273–285,2017.

[18] S. Dutta and S. K. Bandyopadhyay, "Fake job recruitment detection using machine learning approach," International Journal of Engineering Trends and Technology, vol. 68, no. 4, pp. 48–53, 2020.

[19] S. Moon, M.-Y. Kim, and D. Iacobucci, "Content analysis of fake consumer reviews by survey-based text categorization," International Journal of Research in Marketing, vol. 38, no. 2, pp. 343–364, 2021.

[20] M. Naude, K. J. Adebayo, and R. Nanda, "A machine learning approach ′ to detecting fraudulent job types," AI & SOCIETY, pp. 1–12, 2022.

[21] S. Bandyopadhyay and S. Dutta, "Fake job recruitment detection using machine learning approach," International Journal of Engineering Trends and Technology, vol. 68, 04 2020.

[22] A. S. Pillai, "Detecting fake job postings using bidirectional lstm," arXiv preprint arXiv:2304.02019, 2023.

[23] S. U. Habiba, M. K. Islam, and F. Tasnim, "A comparative study on fake job post prediction using different data mining techniques," in 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 543–546.

[24] K. Mehta, N. K. Rajagopal, S. B. Gaikwad, and S. Yadav, "Machine learning techniques for true and fake job posting," ECS Transactions, vol. 107, no. 1, p. 2383, apr 2022.

[25] D. R. Cox, "The application of the logistic function to psychophysical data," Journal of the Royal Statistical Society, Series B (Methodological), vol. 20, no. 2, pp. 215–242, 1958.

[26] J. L. McClelland, D. E. Rumelhart, P. R. Group et al., Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models. MIT press, 1987, vol. 2.

[27] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5–32, 2001.

[28] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol. 7, no. 2, pp. 179–188, 1936.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[30] A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, and S. Ludi, "Detection of fake job postings by utilizing machine learning and natural language processing approaches," Neural Processing Letters, pp. 1–29, 2022