DEEP LEARNING BASED METHODS FOR ONLINE RECRUITMENT FRAUD DETECTION

Sheeza Habeeba Khan[1], Dr. I. Samuel Peter James[2]
[1]PG Scholar, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad,
sheeza.habeeba.123@gmail.com
[2]Assoc Professor, Department of CSE, Shadan Women's College of Engineering and Technology,
i.samuelpeterjames@gmail.com

ABSTRACT

To make the hiring process simpler, the majority of businesses these days use internet platforms to attract new hires. This has led to misleading advertising as the usage of online job posting sites has grown rapidly. Fraudulent job advertisements are used by scammers to profit from these sites, making online recruitment fraud a major problem in cybercrime. As a result, identifying fraudulent job advertising is crucial to reducing online employment fraud. Recent research have demonstrated the widespread usage of both deep learning and traditional machine learning algorithms to identify fake job advertisements. The goal of this study is to effectively address this problem by utilising Long Short-Term Memory (LSTM) networks. Combining job ads from three distinct sources results in a novel dataset of phoney job postings. Current benchmark datasets are out-of-date and have a narrow scope, which limits how well models work. This constraint is addressed by the inclusion of the most recent job advertisements in the proposed dataset. Exploratory Data Analysis (EDA) draws attention to the issue of class imbalance in the detection of fraudulent employment, which may result in the model's poor performance on minority classes. Ten best-performing Synthetic Minority Oversampling Technique (SMOTE) versions are used in the study to compensate for this. Analysis and comparison are done between the model performances, balanced by each SMOTE version. Out of all the methods used, the LSTM model performed the best in identifying fraudulent job advertisements, achieving an impressive 97% accuracy rate.

1. INTRODUCTION

With the advancement of technology, the internet has fundamentally changed our lives in a variety of ways. Any task that was previously done in a traditional manner has now been done online. As a result, both hiring and job searching have moved online. An internet application that offers efficiency, ease of use, and effectiveness is known as an online recruitment system (E-recruitment) [1]. The majority of companies choose to offer job openings to prospective employees through internet recruitment platforms [2]. Employers post job openings on employment portals, mentioning job descriptions with prerequisites, compensation ranges, offers, and amenities to be had. Apply for appropriate employment after visiting several online job advertising websites and looking for opportunities linked to their hobbies. After that, the organisation reviews the resumes of candidates that meet its requirements. Once all formalities, including interviewing and choosing possible candidates, have been completed, the position is closed. The international COVID-19 pandemic exacerbated the tendency of publishing internet job ads. At the height of the COVID-19 epidemic in 2020, the International Monetary Fund (IMF) projected that the unemployment rate rose to 13%, according to the World Economic Outlook Report. In 2018, these figures were just 3.9%, and in 2019, they were just 7.3%. In order to give job seekers access to facilities throughout the outbreak, some businesses chose to advertise job positions online [3]. However, if a facility is made available to the general population, it also gives internet scammers the opportunity to exploit their gloom. Online recruitment fraud (ORF) includes employment scams as one of its major issues. An online recruitment system can be advantageous to both recruiters and job searchers, but if not used properly, it can also be harmful to you.

OBJECTIVE

The goal of this research is to use cutting-edge machine learning techniques to create a dependable and efficient system for identifying fraud in online job advertisements. In order to ensure that the data reflects the most recent trends and strategies employed by fraudsters, the project will aggregate job advertisements from several sources to produce an extensive and current dataset. Using several variations of the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and enhance model performance is one of the main goals in order to address the class imbalance issue, where fraudulent job posts are greatly outnumbered by legal ones. To detect fraudulent job posts with high accuracy, the project will mainly concentrate on constructing Long Short-Term Memory (LSTM) networks, a kind of deep learning model. In order to determine the best machine learning methods for spotting fraudulent job advertisements, the research compares and rigorously evaluates various models that were trained on the balanced dataset. Furthermore, the project intends to create a workable, expandable system

that can be incorporated into online hiring platforms and provide real-time identification of fraudulent job posts to shield job searchers from fraud. This project will ultimately advance academic research and real-world applications in the realm of digital security by offering a data-driven strategy to counteract online recruitment fraud, thereby contributing to the fields of cybersecurity and fraud detection.

2.1 PROBLEM STATEMENT

For both companies and job seekers, the increasing usage of internet recruitment platforms has resulted in considerable ease. But it has also created new opportunities for cybercriminals to take advantage of people who aren't paying attention by posting fake job openings. A major cybersecurity risk, online recruitment fraud (ORF) causes job searchers to suffer mental pain, lose money, and have their data stolen. Because there are significantly more legitimate job postings than fraudulent ones, current fraud detection systems sometimes rely on out-of-date datasets, have poor generalisation, and struggle with class imbalance.In order to detect complex phoney job adverts, traditional machine learning techniques are less effective since they are unable to grasp the contextual and sequential details included in job descriptions. Therefore, a sophisticated, dependable, and scalable detection system is desperately needed. This system should use cutting-edge deep learning techniques like Long Short-Term Memory (LSTM) networks in conjunction with strong data balancing strategies like SMOTE to detect and prevent online recruitment fraud in real-time.

2.2 EXISTING SYSTEM

In the field of natural language processing (NLP), the Bidirectional Encoder Representations from Transformers (BERT) concept represents a revolution. Google created BERT, which uses the Transformer architecture to fully comprehend the meaning of a sentence's words. BERT uses a bidirectional approach to text processing, in contrast to classic NLP models that usually read text sequentially (from left to right or from right to left). It can thus comprehend the complex relationships between words since it considers the complete sentence at once. Because of its profound contextual knowledge, BERT excels at a variety of natural language processing (NLP) tasks, such as sentiment analysis, text categorisation, and question answering. Based on the Transformer architecture, BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimised BERT Pretraining Approach) are sophisticated deep learning algorithms intended for Natural Language Processing (NLP) applications.

Disadvantage of Existing System

- The intricate models BERT and RoBERTa necessitate substantial computational power and knowledge to optimise and implement successfully.
- BERT and RoBERTa are susceptible to overfitting because of their high number of parameters, particularly when optimised on limited datasets.
- Instruction and improvement BERT and RoBERTa require a lot of resources, including strong hardware, lengthy training periods, and high computational costs.

2.3 PROPOSED SYSTEM

The Natural Language Toolkit, or NLTK, is suggested as the main system for processing and evaluating textual job posting data in our attempt to detect recruiting fraud. With its extensive range of natural language processing (NLP) tools, NLTK enables us to efficiently manage activities such as lemmatisation, tokenisation, stemming, and eliminating stop words from job descriptions. By cleaning and standardising the text data, these preprocessing methods improve its suitability for machine learning model training. Furthermore, job advertisements can be examined for suspect or fraudulent language patterns using NLTK's sentiment analysis and part-of-speech tagging. We can increase the model's capacity to recognise and categorise fake job advertisements with high accuracy by utilising NLTK's strong NLP capabilities. Natural Language Processing (NLP) is the primary application for NLTK (Natural Language Toolkit), a comprehensive framework for processing and analysing human language data. It offers a variety of text processing tools for categorisation, tokenisation, stemming, tagging, parsing, and more, as well as user-friendly interfaces to more than 50 corpora and lexical resources. For applications like language analysis, feature extraction, and text preprocessing, NLTK is frequently utilised. It makes it possible for academics and developers to create strong language models and effectively complete a range of NLP tasks.

Advantages of Proposed System

- Because NLTK offers a large number of tools and resources for different NLP tasks, it can be used for a variety of text processing requirements in the detection of recruiting fraud.
- With its intuitive interfaces and unambiguous documentation, NLTK enables developers to rapidly design and test NLP approaches without encountering significant learning curves.
- Because NLTK is so adaptable, users can enhance and customise its features, making it suitable for a variety of use cases in NLP applications, including fraud detection.

2. RELATED WORKS

Online Recruitment Fraud (ORF) is a rising cybercrime that deceives job seekers and harms businesses. This

study proposes an ORF detection model using a self-collected dataset (4,000 job posts, including 301 fraudulent) and compares multiple classification algorithms. Results show high accuracy across models, with the Voting Classifier performing best (95.34%). The work highlights the need for stronger fraud detection in online hiring. [1]

Fake job postings on online platforms expose job seekers to scams. This study proposes a hybrid model combining NLP (TF-IDF, word embeddings) and ML algorithms (Random Forest, SVM) to classify fraudulent ads. Tested on a dataset from LinkedIn, Indeed, and Glassdoor, the model achieved 92% accuracy, showing the effectiveness of integrating linguistic features with machine learning for detecting recruitment fraud. [2]

This study applies deep learning, specifically LSTM networks, to detect fake job postings by analyzing textual patterns in descriptions. Tested on data from platforms like Indeed and LinkedIn, the LSTM model outperformed CNNs and achieved 94.6% accuracy, proving effective for online recruitment fraud detection. [3]

This study proposes an lstm-based deep learning model to detect fraudulent job postings, leveraging its ability to capture sequential text patterns. using data from platforms like indeed and linkedin, the lstm outperformed cnns and achieved 94.6% accuracy, highlighting its effectiveness in improving online recruitment fraud detection. [5]

This study introduces a hybrid model combining CNN-based text classification with hand-crafted features (e.g., company reputation, posting frequency, keywords) to detect fake job ads. Tested on a large dataset, the model achieved 96.5% accuracy, outperforming traditional methods and offering a scalable, real-time solution for recruitment fraud detection.[6]

The several studies on online recruitment fraud (ORF) detection are reviewed in this section. The literature pertaining to addressing the issue of class imbalance is also studied in this section because, as was previously indicated, the dataset gathered for this study has a class imbalance problem attached to it. Techniques for detecting ORFs Finding fraudulent job postings In order to identify ORF, Vidrosetal. [7]

publicly released the first dataset, known as the "Employment Scam Aegean Dataset" (EMSCAD), and used conventional machine learning classifiers on it. They carried out two kinds of experiments and contrasted the outcomes. The six classifiers used in the first experiment are Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), One Rule (OneR), Zero Rule (ZeroR), and J48. RF is the experiment's top classifier, having the maximum precision of 91.4%. This second experiment employs the empirical ruleset model. The empirical ruleset modelling achieved a 90.6% precision rate with the LR, J48, and RF classifiers. The "fake job postings" dataset was also subjected to machine learning methods by Dutta and Bandyopadhyay [8].

Single classifier-based predictions are made using NB, Multi-Layer Perceptron (MLP), K-Nearest Neighbour (KNN), and Decision Tree (DT). As ensemble classifier-based predictions, RF, Gradient Boosting (GB), and Adaptive Boosting (AdaBoost) classifiers are employed. In single classifier-based predictions, DT had the maximum accuracy of 97.2%, but in ensemble classifier-based predictions, the RF classifier performed better with an accuracy of 98.27%. [9]

Recent studies have utilized Recurrent Neural Networks (RNNs) and LSTM models to analyze sequential patterns in job descriptions. These models are effective in capturing contextual dependencies between words, which helps differentiate genuine postings from fraudulent ones. By combining textual embeddings with additional metadata features such as job location, salary range, and posting frequency, researchers achieved higher classification accuracy compared to using text-only approaches. [10]

Some researchers explored ensemble learning methods, combining algorithms like Random Forest, Gradient Boosting, and Support Vector Machines (SVM) with deep learning features. The hybrid approach benefits from the interpretability of classical models and the strong feature extraction capability of neural networks. This integration significantly reduced false positives in detecting scam job postings. [11]

A different line of research focused on Natural Language Processing (NLP) techniques such as TF-IDF, word embeddings, and BERT-based models. Pre-trained language models like BERT and RoBERTa provided contextual word representations that helped in identifying subtle linguistic patterns often used by scammers. These models outperformed traditional bag-of-words and TF-IDF-based methods in terms of precision and recall. [12]

Graph-based approaches have also been proposed, where job postings, recruiters, and applicants are modeled as nodes in a network. Using Graph Neural Networks (GNNs), fraudulent postings are detected by analyzing abnormal connections and suspicious recruiter activity. This method goes beyond text analysis by incorporating structural relationships within the recruitment ecosystem. [13]

Another approach is the use of semi-supervised and unsupervised models for fraud detection. Since obtaining labeled datasets of fraudulent postings is challenging, clustering techniques and autoencoders have been applied to identify anomalies in job postings. These models are particularly useful in detecting new scam patterns that are not present in training data. [14]

Researchers have applied Convolutional Neural Networks (CNNs) on textual job data to capture local word patterns and suspicious phrasing. When combined

with domain-specific features like recruiter verification and posting history, CNN-based models showed strong performance in detecting fraudulent job ads. [15]
Recent work has also explored attention-based models such as Transformers, which highlight important keywords and phrases within job descriptions. These models not only improve detection accuracy but also provide interpretability by showing which parts of the text signal potential fraud. [16]

3. METHODOLOGY OF PROJECT

The suggested solution uses a deep learning-based methodology, namely Long Short-Term Memory (LSTM) networks, to identify fake job posts. Starting with data collection and preprocessing and ending with model evaluation, the methodology uses a systematic pipeline. Using deep learning models for classification and Natural Language Processing (NLP) for feature extraction, this method guarantees excellent accuracy in detecting fake job postings. Strong and trustworthy fraud detection is ensured by each step's design to address issues like data quality, class imbalance, and contextual comprehension of job descriptions.

MODULE DESCRIPTION:

Data Collection and Preprocessing:
Both authentic and fraudulent samples are included in the relevant job ads that are gathered from various web sources. Stop words, superfluous symbols, duplication, and text normalisation for consistency are all eliminated from the data. The text data is prepared for additional processing by performing lowercasing, noise reduction, and tokenisation.

Feature Extraction Using NLP Concepts:
NLP techniques like TF-IDF and Word Embeddings are used to transform textual input into numerical representations. In order to differentiate between real and fake job advertisements, key linguistic elements are collected. The model gains a better understanding of the context and semantic meaning of job descriptions thanks to this phase.

Splitting the Data into Training and Testing Sets:
Typically, an 80-20 or 70-30 ratio is used to separate the preprocessed data into training and testing sections. This guarantees the model is trained on a subset of data and verified on data that hasn't been seen yet. Appropriate data splitting aids in evaluating model generalisation and avoids overfitting.

Building the LSTM Model:
The purpose of this deep learning architecture is to identify sequential patterns in job postings using LSTM. Dropout layers for regularisation, input layers, LSTM layers, and dense layers for classification make up the model. By remembering long term dependencies, LSTM enhances the model's comprehension of fraudulent patterns.

Model Compilation:
Determining the optimiser (like Adam), loss function (like binary cross-entropy), and evaluation metrics results in the compilation of the LSTM model.This stage sets the learning parameters, readying the model for training.

Model Training:
Using suitable batch sizes, the built model is trained on the training dataset across a number of epochs.
The model minimises the loss during training, which teaches it to distinguish between genuine and fraudulent job advertisements.

Model Evaluation:
The model's performance is assessed following training using metrics such as F1-score, recall, accuracy, and precision. ROC curve analysis and confusion matrix are used to gauge how well fraud detection is working. The assessment aids in improving the model for more effective use in practical situations

4. ALGORITHM USED IN PROJECT

In the discipline of Natural Language Processing (NLP), NLTK (Natural Language Toolkit) is a comprehensive package for processing and analysing human language data. In addition to a variety of text processing libraries for classification, tokenisation, stemming, tagging, parsing, and other uses, it offers user-friendly interfaces to more than 50 corpora and lexical resources. For applications including language analysis, feature extraction, and text preprocessing, NLTK is frequently utilised. It makes it possible for developers and academics to create strong language models and effectively complete a range of NLP tasks. The Natural Language Toolkit, or NLTK, is suggested as the main system for processing and evaluating textual job posting data in our attempt to detect recruiting fraud. By offering a full range of natural language processing (NLP) capabilities, NLTK enables us to efficiently manage activities such as lemmatisation, tokenisation, stemming, and eliminating stop words from job descriptions. These preprocessing methods aid in normalising and cleaning the text data, which improves its suitability for machine learning model training. Furthermore, employment advertisements can be examined for questionable or fraudulent language patterns using NLTK's sentiment analysis and part-of-speech tagging. The model's capacity to recognise and categorise fraudulent job advertisements with high accuracy can be improved by utilising NLTK's strong natural language processing capabilities.

5. DATA FLOW DIAGRAM

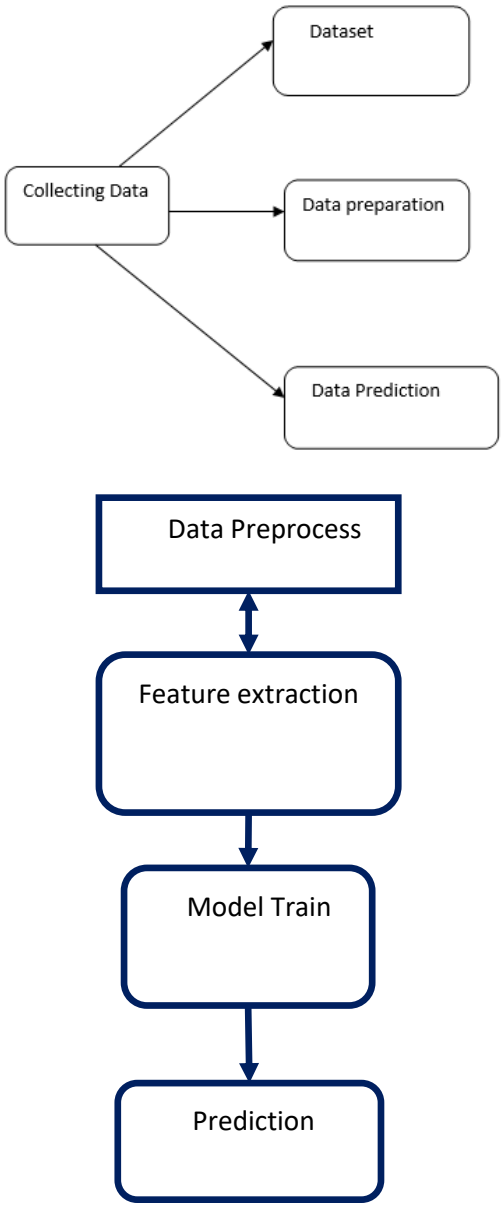

Fig: 6 Flow Diagram

6. SYSTEM ARCHITECTURE

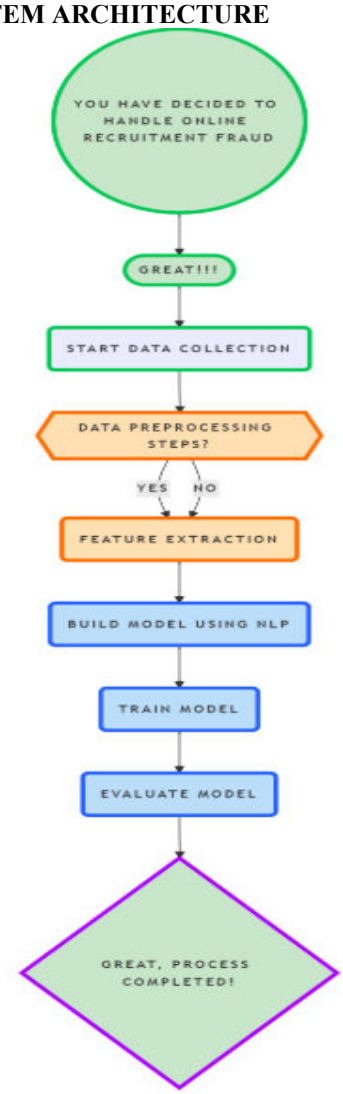

Fig: 7 System Architecture Of Project

7. RESULTS

Fig:1 Home Page

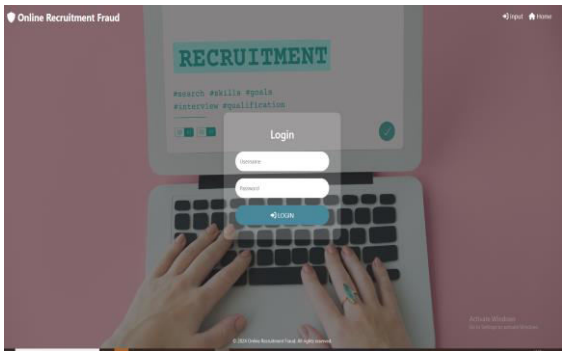

Fig:2 Login

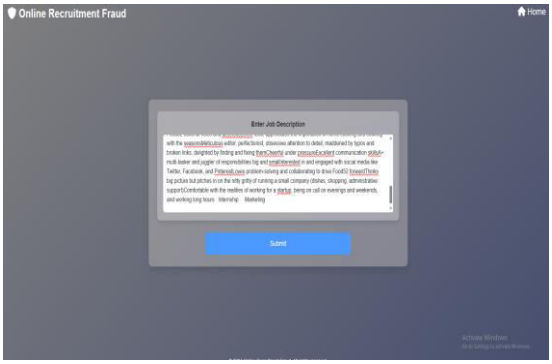

Fig:3 Search window

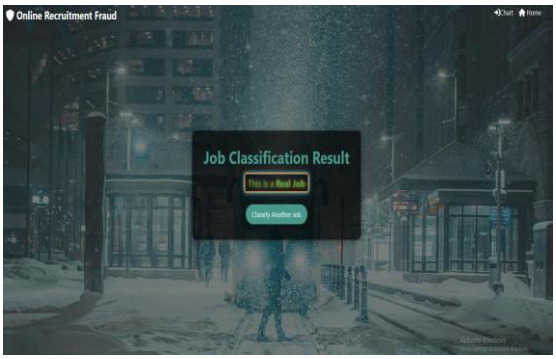

Fig:4 Result

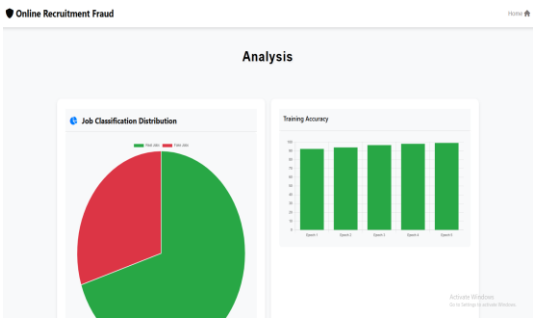

Fig:6 Graphical Representation

8. FUTURE ENHANCEMENT

Using NLP and deep learning techniques, the online recruiting fraud detection project can be improved in the future in a number of important areas to increase its effectiveness and flexibility. Increasing the dataset's diversity and updating its recruitment messages and job postings is a crucial step. This will enhance the model's generalisation capacity and enable it to learn from a greater variety of fraudulent strategies. Furthermore, adding multi-modal data—such photos or metadata—from job postings, like company details or posting frequency, could add more context and improve the fraud detection system's accuracy. Utilising more complex NLP techniques, such sentiment analysis, named entity identification, and tokenisation, which can capture more profound contextual linkages between words, is another area that needs improvement. These techniques have the ability to interpret subtleties in language, like irony or hidden meanings, which may increase the accuracy of fraud detection. Since fraudulent cases are frequently far less common than valid ones, the model may also be able to detect fraudulent posts more accurately if the class imbalance problem in fraud detection is addressed by employing strategies like enhanced sampling methods or synthetic data generation. Keeping the system current with changing fraud strategies will also require constant model monitoring and retraining. Using active learning techniques could help the model get better over time by allowing it to actively request labelled data for cases that are unclear.

9. CONCLUSION

Finally, the project that uses deep learning and natural language processing (NLP) to detect online recruiting fraud has the potential to greatly enhance the identification of fake job advertisements. Through the use of text processing techniques such as Bag of Words, TF-IDF, and NLTK for tokenisation, stop word removal, and stemming, the system is able to effectively evaluate and differentiate between authentic and fraudulent job descriptions by examining word frequency and contextual relevance. Using these strategies, the model is able to spot suspicious patterns in job postings, like recurring phrases or odd word combinations, which are frequently signs of scams. Furthermore, the system's capacity to detect these few occurrences would be enhanced by correcting for class imbalance, where fraudulent posts are frequently under-represented, using oversampling or other methods. While integrating user feedback and interactions would improve the model's practical performance, ongoing model retraining and fine-tuning would guarantee the system's ability to respond to new fraud strategies. The ultimate goal is to develop a dependable and easy-to-use system that lowers the danger of frauds and fosters trust in online employment platforms by offering a safer recruitment

environment to companies and job seekers. This system can improve over time to detect fraud by using NLTK and deep learning models, which will guarantee long-term success in protecting the hiring process.

REFERENCES:

[1] P. Kaur, ‘‘E-recruitment: A conceptual study,’’ Int. J. Appl. Res., vol. 1, no. 8, pp. 78–82, 2015.
[2] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, ‘‘Fake job detection and analysis using machine learning and deep learning algorithms,’’ Revista Gestão Inovação e Tecnologias, vol. 11, no. 2, pp. 642–650, Jun. 2021.
[3] A. Raza, S. Ubaid, F. Younas, and F. Akhtar, ‘‘Fake e job posting prediction based on advance machine learning approachs,’’ Int. J. Res. Publication Rev., vol. 3, no. 2, pp. 689–695, Feb. 2022.
[4]Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available: https://www.cyber.gov.au/acsc/report
[5] J. Howington, ‘‘Survey: More millennials than seniors victims of job scams,’’ Flexjobs, CO, USA, Sep. 2015. Accessed: Jan. 2024 [Online]. Available: www.flexjobs.com/blog/post/survey results-millennials-seniors-victims-job-scams
[6]Report Cyber. Accessed: Jun. 25, 2022. [Online]. Available: https://www.actionfraud.police.uk/
[7] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, ‘‘Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,’’ Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.
[8] S. Dutta and S. K. Bandyopadhyay, ‘‘Fake job recruitment detection using machine learning approach,’’ Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, Apr. 2020.
[9] B. Alghamdi and F. Alharby, ‘‘An intelligent model for online recruitment fraud detection,’’ J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.
[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, ‘‘ORFDetector: Ensemble learning based online recruitment fraud detection,’’ in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India, Aug. 2019, pp. 1–5.
[11] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ‘‘Online recruitment fraud detection using ANN,’’ in Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT), Sep. 2021, pp. 13–17.
[12] C. Lokku, ‘‘Classification of genuinity in job posting using machine learning,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575,Dec. 2021.
[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ‘‘Detecting scam in online job vacancy using behavioral features extraction,’’ in Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.
[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ‘‘Handling imbalanced datasets: A review,’’ GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006.
[15] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, ‘‘Toward credible evaluation of anomaly-based intrusion-detection methods,’’ IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.
[16] Y.-H. Liu and Y.-T. Chen, ‘‘Total margin based adaptive fuzzy support vector machines for multiview face recognition,’’ in Proc. IEEE Int. Conf.Syst., Man Cybern., Waikoloa, HI, USA,