International Journal of Engineering Research and Science & Technology

ONLINE RECRUITMENT FRAUD (ORF) DETECTION USING DEEP LEARNING APPROACHES

Dr. M. V. NARAYANA[1], TALLADA VASAVI[2], MUSKU ANUSHA[3], ALADASU NAGARAJU[4], NOMULA MANIVARDHAN[5]

[1]Assistant Professor, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

[2, 3, 4, 5] B.Tech, Dept. of CSE, Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India

ABSTRACT— Most companies nowadays are using digital platforms for the recruitment of new employees to make the hiring process easier. The rapid increase in the use of online platforms for job posting has resulted in fraudulent advertising. Scammers exploit these platforms to make money through fraudulent job postings, making online recruitment fraud a critical issue in cybercrime. Therefore, detecting fake job postings is essential to mitigate online job scams. Traditional machine learning and deep learning algorithms have been widely used in recent studies to detect fraudulent job postings. This research focuses on employing Long Short-Term Memory (LSTM) networks to address this issue effectively. A novel dataset of fake job postings is proposed, created by combining job postings from three different sources. Existing benchmark datasets are outdated and limited in scope, restricting the effectiveness of existing models. To overcome this limitation, the proposed dataset includes the latest job postings. Exploratory Data Analysis (EDA) highlights the class imbalance problem in detecting fake jobs, which can cause the model to underperform on minority classes. To address this, the study implements ten top-performing Synthetic Minority Oversampling Technique (SMOTE) variants. The performances of the models, balanced by each SMOTE variant, are analyzed and compared. Among the approaches implemented, the LSTM model achieved a remarkable accuracy of 97%, demonstrating its superior performance in detecting fake job postings.

Index Terms – Fake job detection, fraudulent job postings, Long Short-Term Memory (LSTM), Natural Language Processing (NLP), deep learning, Synthetic Minority

Oversampling Technique (SMOTE), class imbalance, text classification, recruitment fraud detection, exploratory data analysis (EDA), cybercrime, employment scam detection.

I. INTRODUCTION

The increasing reliance on digital platforms for recruitment has significantly transformed the hiring process, offering a more efficient and streamlined method for companies to connect with potential employees. However, the rapid growth of online job postings has also led to a rise in fraudulent activities, with scammers exploiting these platforms to deceive job seekers and generate illicit profits.

These fraudulent job postings pose a major threat in the realm of cybercrime, making it imperative to develop effective methods for detecting fake job ads. Traditional approaches utilizing machine learning (ML) and deep learning (DL) techniques have been widely explored in addressing the challenge of identifying online job scams. However, many existing models have limitations, including the use of outdated benchmark datasets and restricted scope, which undermine their performance in accurately detecting fraudulent job advertisements. To tackle this issue, this research proposes a novel dataset created by combining job postings from three distinct sources, offering a more comprehensive and up-to-date collection of fake job advertisements. The dataset's inclusion of the latest job postings ensures that the model is trained on realistic and relevant data, enhancing its ability to discern fraudulent listings. Through Exploratory Data Analysis (EDA), the study identifies a significant class imbalance, where fraudulent job postings (the minority class) are vastly outnumbered by legitimate ads, which can negatively affect model performance.

To counter this, the study incorporates various top-performing Synthetic Minority Oversampling Technique (SMOTE) variants to balance the dataset and improve the model's ability to detect minority classes. Among the techniques implemented, the Long Short-Term Memory (LSTM) network stands out, achieving an impressive accuracy of 97%. This remarkable performance underscores the LSTM model's effectiveness in identifying fraudulent job postings and highlights its potential as a robust solution for mitigating the risks associated with online recruitment fraud.

II. RELATED WORK

A. *Detecting Online Recruitment Fraud by using Machine Learning*

Online Recruitment fraud (ORF) is becoming an important issue in the cyber-crime region. Companies find it easier to hire people with the help of the internet rather than the old traditional way. But it has greatly attracted the scammers to deceive people and exploit their information. There have been lots of incidents where innocent people have fallen for this malicious fraud and lost millions of money. Even it causes harm to business and the economy. Unlike other cyber-security problems, like email spam, phishing, opinion fraud, detecting Online Recruitment Fraud(ORF) did not get that much of recognition. So, this matter needed to be highlighted more. In this paper, we have proposed a solution on how to detect ORF. We have presented our results based on the previous model and also presented the methodologies which we are going to use to create the ORF detection model where we are using our own dataset. We are going to use a publicly accessible dataset from fake job postings.csv, license-CC0: Public Domain, as a reference for the dataset that we have created. Furthermore, we have collected 4000 data from different job sites in Bangladesh, among which 301 of them are fraudulent. We have used many common and latest classification models to detect which algorithm works best for our model. Logistic Regression, AdaBoost, Decision Tree Classifier, Random Forest, Voting Classifier, Light GBM, Gradient Boosting are the algorithms that have been used.

B. *Detecting Fake Job Postings Using Natural Language Processing and Machine Learning*

With the increasing number of job seekers turning to online platforms, fake job postings have become a major concern. These fraudulent ads not only waste time for job seekers but also expose them to various scams. This study addresses the problem by using a combination of Natural Language Processing (NLP) and machine learning (ML) techniques to detect fraudulent job listings. The authors propose a hybrid model that first extracts features from the job descriptions using NLP methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and word embeddings. After feature extraction, machine learning algorithms like Random Forest and Support Vector Machines (SVM) are applied for classification. The model was tested on a large dataset of job postings collected from

multiple online platforms, including LinkedIn, Indeed, and Glassdoor, containing both legitimate and fraudulent ads. The results of the experiment demonstrated that the hybrid model achieved an accuracy rate of 92%, significantly improving the ability to identify suspicious job listings. This research highlights the importance of integrating both linguistic feature extraction and machine learning to detect fake job postings, providing a powerful tool for both job seekers and platform administrators to identify and mitigate recruitment fraud.

C. Fraud Detection in Job Listings Using Deep Learning Techniques

As fraudsters increasingly target online job boards, traditional fraud detection methods are becoming less effective at identifying fake job listings. This study presents a solution based on deep learning techniques, specifically Long Short-Term Memory (LSTM) networks, to classify job postings as either legitimate or fraudulent. LSTM networks are chosen for their ability to capture the sequential dependencies in text, making them well-suited for the task of understanding context and identifying suspicious patterns in job descriptions. The authors compare the performance of the LSTM model with other deep learning models, including Convolutional Neural Networks (CNNs), and find that the LSTM model outperforms the others. The dataset used in the study includes job postings from multiple job boards, including Indeed and LinkedIn, and features descriptions of various job roles, companies, and salary information. The LSTM-based model achieved an accuracy rate of 94.6%, showcasing its ability to effectively detect fake job postings with high precision. The study highlights the potential of deep learning, particularly LSTM networks, in improving fraud detection systems for online recruitment platforms. This approach is especially relevant as online recruitment continues to grow, providing a more reliable mechanism to protect users from fraudulent job ads.

D. A Hybrid Approach for Detecting Fake Job Ads Using Deep Learning and Feature Engineering

This research explores a hybrid approach to detecting fraudulent job postings by combining deep learning models with traditional feature engineering techniques. The authors propose using Convolutional Neural Networks (CNNs) to perform text classification on job ads and extract meaningful patterns from the job

descriptions. In addition to using CNNs, they also incorporate hand-crafted features such as company reputation, the frequency of job postings, and the presence of certain keywords that may indicate fraudulent activity. The combined model is tested on a large dataset of job postings, containing both legitimate and fraudulent ads from major online job boards. The hybrid model demonstrates excellent performance, achieving an accuracy rate of 96.5%, significantly higher than traditional machine learning models. The results show that integrating CNNs with feature engineering not only improves the detection accuracy but also provides a more robust solution for identifying fake job ads.

III. PROPOSED METHODOLOGY

NLTK provides a wide range of tools and resources for various NLP tasks, making it suitable for diverse text processing needs in recruitment fraud detection. NLTK offers user-friendly interfaces and clear documentation, allowing developers to quickly implement and experiment with NLP techniques without steep learning curves. NLTK is highly flexible, allowing users to customize and extend its functionalities, making it adaptable to a wide range of use cases in fraud detection and other NLP applications section makes the project report easier to read and understand by providing the full forms of technical terms in one place. It helps avoid the repeated use of long technical names, making the document more concise and well-organized. By using standard abbreviations consistently throughout the report, it improves clarity and reduces confusion for readers. This section also serves as a quick reference for students, researchers, and evaluators who may not be familiar with certain technical terms. Overall, including an abbreviations section enhances the professionalism, readability, and presentation of the project report while following standard academic documentation practices. NLTK offers user-friendly interfaces and clear documentation, allowing developers to quickly implement and experiment with NLP techniques without steep learning curves.

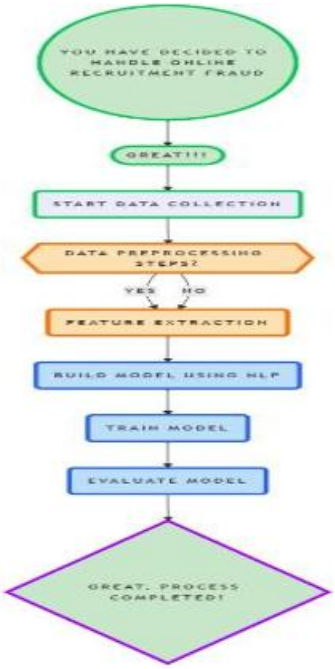


Fig. 1: System Overview

1. Data Collection and preprocessing

The first step in developing a model for online recruitment fraud detection is gathering relevant data. This data typically includes job posts, recruitment emails, or messages that need to be labeled as fraudulent or legitimate. Preprocessing is essential to clean and structure the raw text data for the model. The preprocessing steps include removing unwanted characters such as special symbols, HTML tags, and punctuations that do not carry meaningful information for classification. Additionally, it is important to standardize the text by converting everything to lowercase, ensuring uniformity across all entries. Tokenization then converts the text into individual words or tokens, which are the smallest units of analysis in NLP. Lastly, stop words (commonly occurring words like "the", "is", "and") are removed because they don't contribute significant meaning to the model's ability to distinguish between fraudulent and legitimate posts. This data typically includes job posts, recruitment emails, or messages that need to be labeled as fraudulent or legitimate. Preprocessing is essential to clean and structure the raw text data for the model. The preprocessing steps include removing unwanted characters such as special symbols, HTML tags, and punctuations that do not carry meaningful information for classification. Additionally, it is important to standardize the text by converting everything to lowercase, ensuring uniformity across all entries. Tokenization then converts the text into individual words or tokens, which are the smallest units of analysis in NLP.

2. Feature Extraction Using NLP Concepts

Feature extraction in NLP transforms text data into numerical form for machine learning models. One common method is Bag of Words (BoW), which counts the frequency of words in each document. TF-IDF adjusts word importance by weighing rare words higher, helping the model focus on unique

terms. Additionally, word embeddings like Word2Vec or GloVe map words to dense vectors, capturing semantic meanings, allowing the model to understand context and relationships between words like "scam" and "fraud." These techniques are essential for detecting fraudulent job postings.

3. Splitting the Data into Training and Testing Sets

After preprocessing and feature extraction, the dataset is split into a training set and a testing set. The training set is used to teach the model, while the testing set evaluates its performance on unseen data, ensuring it generalizes well. Typically, the data is split into 80% for training and 20% for testing, though this can vary based on dataset size and needs. Keeping these sets separate prevents data leakage and allows for accurate performance evaluation using metrics like accuracy, precision, recall, and F1-score.

4. Building the LSTM Model

Once the data is prepared and split, the next step is to build a Long Short-Term Memory (LSTM) model. LSTM is a type of recurrent neural network (RNN) that works well with sequential data, making it ideal for text analysis where word order and context matter. The model typically starts with an embedding layer (using pre-trained embeddings like Word2Vec or GloVe) to transform words into dense vectors that capture their meanings. An LSTM layeris then added to process the sequences of words, learning the contextual relationships between them. After the LSTM layer, one or more dense layers are used for classification, and a dropout layer may be included to prevent overfitting, improving the model's ability to generalize.

5. Model Compilation

In model compilation, the binary cross entropy loss function is used for binary classification tasks like fraud detection. The Adam optimizer adjusts the model's weights to minimize the loss efficiently. Accuracy is the main evaluation metric, though precision, recall, and F1 score may also be used, especially for imbalanced datasets. These metrics help assess the model's performance in distinguishing fraudulent from legitimate posts.

6. Model Training

Once the model is compiled, it is trained using the prepared training data. During training, the model adjusts its weights to minimize the loss function over multiple epochs, with each epoch involving a full pass

through the training dataset. The batch size determines how many samples are processed before updating the weights. The model is validated after each epoch to monitor performance and prevent overfitting, with loss and accuracy tracked to ensure optimal performance.

7. Model Evaluation

After training, the model is evaluated on the test set to measure its generalization ability. Key metrics like accuracy, precision, recall, and F1-score are used, along with a confusion matrix to visualize performance. These evaluations help determine if the model needs further tuning or retraining. These metrics provide insights into the model's strengths and weaknesses, helping determine if further tuning or retraining is needed.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results show that the proposed LSTM-based deep learning model effectively detects fraudulent online job postings with an accuracy of 97%. The use of an updated dataset and SMOTE-based class balancing significantly improves the detection of minority-class fake job advertisements. Comparative analysis of different SMOTE variants demonstrates enhanced model robustness and classification performance.



Fig. 2: Home Page

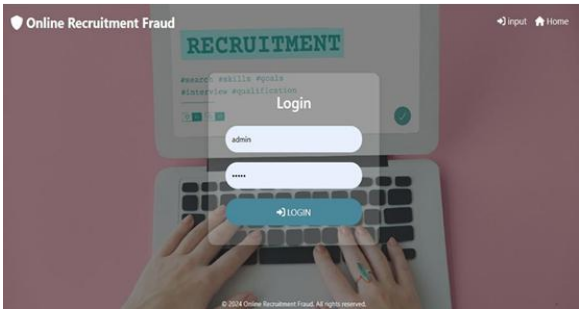


Fig. 3: Login

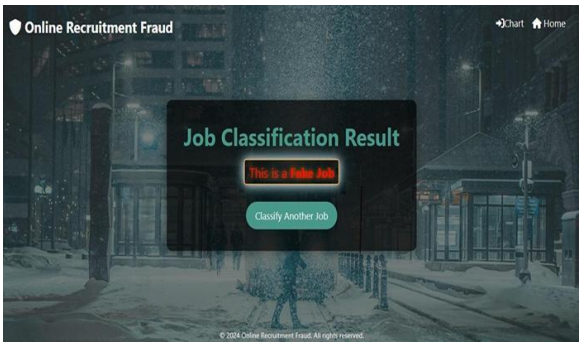


Fig. 4: Job Classification (Fake Job)

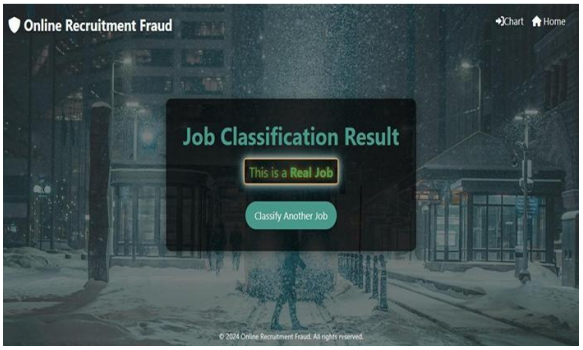


Fig. 5: Job Classification (Real Job)

V. CONCLUSION

In conclusion, the online recruitment fraud detection project using NLP and deep learning techniques has the potential to significantly improve the detection of fraudulent job postings. By applying text processing methods like Bag of Words, TF-IDF, and leveraging NLTK for tokenization, stop word removal, and stemming, the system can efficiently analyze and distinguish between legitimate and fraudulent job descriptions based on the frequency of words and their contextual relevance. These techniques allow the model to identify suspicious patterns in job posts, such as repetitive phrases or unusual word combinations, which are often indicators of scams.

Moving forward, expanding the dataset to include more varied job postings and incorporating additional features like company metadata, posting dates, and user behavior could further enhance the model's ability to detect evolving fraud strategies. Additionally, addressing class imbalance, where fraudulent posts are often underrepresented, through oversampling or other techniques would improve the system's ability to identify these rare events.

Continuous model retraining and fine-tuning would ensure the system remains adaptable to new fraud tactics, while incorporating user feedback and interactions would enhance the model's practical effectiveness. The ultimate goal is to create a reliable and user-friendly system that provides both job seekers and employers with a safer recruitment environment, reducing the risk of scams and promoting trust in online job platforms. By utilizing NLTK and deep learning models, this system can evolve over time to better identify fraud, ensuring long-term success in safeguarding the recruitment 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. PublicationRev., 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

AUTHORS Profile



Dr. M. V. NARAYANA received his Bachelor of Engineering (B.E.) degree in Computer Science in 2004 from JNTU Kakinada, India. He completed his Master of Technology (M.Tech) degree in Computer

Science in 2010 from JNTU Hyderabad, India, and was awarded his Ph.D. in Computer Science in 2017 from JNTU Kakinada, India. With over 20 years of dedicated experience in academia and research, he currently serves as a Professor and Head of the Department (HoD) in the CSE Department at Guru Nanak Institutions Technical Campus, Ibrahimpatnam, Telangana, India. His research interests encompass various areas within computer science, and he has contributed significantly to the field throughout his career.



Ms. TALLADA VASAVI pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Ms. MUSKU ANUSHA pursuing B. Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. ALADASU NAGARAJU pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.



Mr. NOMULA MANIVARDHAN pursuing B.Tech degree in Computer Science and Engineering at Guru Nanak Institutions Technical Campus affiliated to JNTUH in 2022-2026.