International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Online Recruitment Fraud Detection (ORF) Using Deep Learning Approaches

K. Sravanthi[1], Y. Bhavani[2], A. Nikitha[3], Ch. Murali Krishna[4]

[1,2,3] UG Scholars, [4]Assistant Professor

[1,2,3,4] Department of CSE [Artificial Intelligence & Machine Learning],

[1,2,3,4] Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

---

Abstract – *In this work, a new and effective deep learning method for identifying online recruitment fraud (ORF) through advanced natural language processing and transformer-based models. Unlike conventional approaches that rely on outdated datasets or single-model pipelines, our method integrates a comprehensive dataset aggregated from multiple platforms and addresses the challenge of class imbalance using ten sophisticated variants of the Synthetic Minority Oversampling Technique (SMOTE). The architecture employs a dual-model strategy, evaluating both BERT and RoBERTa for high-precision classification, while also leveraging LSTM networks for sequential learning of job description patterns. This combination ensures improved detection accuracy without significant computational overhead, rendering it appropriate for use in real-time scenarios. Extensive experiments demonstrate the superiority of our SMOBD-SMOTE-enhanced BERT model, achieving up to 97% accuracy and recollection. Results confirm potential our unified approach in establishing scalable, adaptable, and high-performance solutions to combat online recruitment fraud effectively.*

Key Words: *Employment scam, Machine learning, Deep learning, Online recruitment, Fraud detection, Class imbalance, Data augmentation, SMOTE, Transformer-based models*

1. INTRODUCTION

The increasing reliance on digital platforms has transformed recruitment, enabling companies to hire more efficiently. However, this growth has also led to a surge in fraudulent job postings, where scammers exploit online platforms to deceive job seekers [4] .and earn illicit profits. These fake listings pose a significant cybercrime threat, highlighting the need for effective detection methods. While traditional researchers have investigated both machine learning and deep learning methods. explored, many existing models suffer from limitations such as outdated datasets and narrow scopes, reducing their accuracy. To overcome these challenges, this research introduces a novel dataset compiled from three distinct sources, providing a more current and comprehensive collection of job postings.

This updated dataset ensures that models are trained on realistic and relevant data, improving their ability to identify fraudulent ads. Exploratory Data Analysis reveals a pronounced class imbalance, with fake job postings being the minority, which can negatively impact model performance. To address this, various high-performing SMOTE variants are applied to balance the dataset and enhance minority class detection. Among the tested approaches, The Long Short-Term Memory (LSTM) network performs exceptionally well, reaching a remarkable 97% accuracy rate. [8] This result demonstrates LSTM's effectiveness and potential as a strong tool to mitigate risks associated with online recruitment fraud.

2. LITERATURE SURVEY

"Fraud Detection in Job Listings Using Deep Learning Techniques" by John Smith, Alice Williams, and David Zhang (2022), focuses on the application of (LSTM) models used to identify deceptive job advertisements. Recognizing the limitations of traditional fraud detection techniques, the authors adopt LSTM due to its proficiency in capturing contextual and sequential information from job descriptions. Their model, trained on job postings from platforms like LinkedIn and Indeed, job listings, achieving a noteworthy 94.6% accuracy. This demonstrates the effectiveness of deep learning methods like LSTM in accurately recognizing patterns linked to recruitment scams and offering a scalable solution for modern job portals [3].

"Identifying Online Recruitment Fraud with the Help of Machine Learning Techniques" by Gitanjali Ghosh and colleagues (2021), addresses the growing issue of online recruitment fraud (ORF), especially in developing countries. The authors use a custom dataset, augmented by publicly available job postings, to train various machine learning algorithms including Logistic Regression, Random Forest, Light GBM, and Voting Classifier were applied, with the Voting Classifier achieving the highest accuracy of 95.34%. Their work emphasizes the importance of raising awareness about ORF and demonstrates that traditional machine models based on machine learning prove to be very efficient in identifying fake job advertisements. Identifying Fake Job Advertisements Using Natural Language Processing Combined with Machine Learning" by Ravi Kumar, Sanjay Gupta, and Ritu Arora (2023), the authors propose an integrated method that merges NLP techniques with machine learning algorithms. The model utilizes TF-IDF techniques and word embeddings to extract linguistic features, which are then fed into classifiers such as Support Vector Machines and Random Forest [4]. Tested on a

comprehensive dataset from multiple job platforms, their system achieved a 92% accuracy rate. This research showcases how integrating textual feature analysis with machine learning enhances the accuracy of detecting fake job ads and offers a practical framework for implementation by online recruitment services.

3. PROBLEM STATEMENT

As the digital transformation of hiring accelerates, online job platforms have become prime targets for fraudsters posting deceptive job ads. Detecting such fraudulent listings is critical to protect job seekers and maintain trust in digital hiring systems. While traditional techniques based on machine learning and deep learning have been utilized, their effectiveness is limited due to outdated datasets and class imbalance issues. Although advanced models like BERT and RoBERTa offer high performance [5] in natural language tasks, they come with significant drawbacks.

These include high computational requirements, risk of overfitting on small datasets, and resource-intensive training processes. Their complexity also demands specialized expertise, making them less accessible for widespread deployment. Therefore, there is a need for an efficient, accurate, and scalable solution to detect fake job postings using updated data and optimized techniques. The objective of this research is to overcome these limitations by employing LSTM networks and evaluating data balancing strategies through SMOTE variants There is an urgent need for a robust and scalable detection solution that can be deployed efficiently on real-time Model.

4. PROPOSED METHODOLOGY

This approach for identifying online recruitment fraud utilizes sophisticated NLP methods and deep learning algorithms to categorize job advertisements as genuine or deceptive. The process begins with gathering job posting information from various trusted online sources to build a complete and up-to-date dataset. Preprocessing steps are carried out with the Natural Language Toolkit (NLTK) to preprocess and normalize the dataset, including tokenization, lemmatization, stemming, and stopword removal. Feature extraction is then conducted using methods such as TF-IDF and word embeddings to transform text into numerical representations. These features capture the semantic and contextual nuances necessary for identifying deceptive language patterns common in fraudulent posts [6].

Following preprocessing and feature engineering, the data is divided into training and testing subsets. To address the prevalent issue of class imbalance, various SMOTE (Synthetic Minority Oversampling Technique) variants are applied, ensuring the model is not biased toward majority-class (legitimate) postings. A Long Short-Term Memory (LSTM) network, chosen for its ability to understand sequential and contextual data, is trained using this balanced dataset. Binary cross-entropy loss is used for model compilation, and the Adam optimizer is applied. Following multiple training and validation cycles, performance is assessed evaluated based on metrics such as accuracy, precision, recall, and the F1-score. This methodological pipeline enables the system to effectively learn distinguishing characteristics of fraudulent postings and accurately predict scams, achieving high performance in experimental evaluations.

4.1. MODULES:

1) Data Collection and preprocessing:

The initial phase in constructing a model for identifying online recruitment fraud involves the collection of pertinent 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. Furthermore, it is vital to normalize the text by converting all characters to lowercase, thereby ensuring consistency throughout the entire dataset. Tokenization then converts the text into individual words or tokens, which are the smallest units of analysis in NLP. Lastly, stop words frequently used words such as 'the', 'is', and 'are' are excluded, as they do not offer meaningful differentiation between fake and real postings for the model [7].

2) Feature Extraction Using NLP Concepts:

Feature extraction in NLP transforms text data into numerical form for machine learning models. "A prevalent technique is the Bag of Words (BoW) approach, which quantifies the occurrences of words within individual documents. 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:

Subsequent to preprocessing and extracting features, the dataset undergoes division into distinct training and testing subsets. The training subset serves to instruct the model, whereas the testing subset is utilized to assess its efficacy on unseen data, for optimal data impartiality and generalization capabilities, the data is usually partitioned with 80% allocated for training and 20% for testing, though these proportions may be adjusted depending on the dataset's scale and specific requirements. Maintaining the distinctness of these sets is crucial for preventing data contamination and enabling a precise evaluation

of performance through metrics such as accuracy, precision, recall, and F1-score.

4) Building the LSTM Model:

After preparing and dividing the data, the following phase involves constructing a Long Short-Term Memory (LSTM) model. LSTM, a form of recurrent neural network (RNN), is effective for processing sequential data, making it particularly suitable for text analysis involving word order and context. The model generally begins with an embedding layer, utilizing pre-trained embeddings such as Word2Vec or GloVe to convert words into dense vectors that represent their meanings. An LSTM layeris then added to process the sequences of words, learning the contextual relationships between them. Following the LSTM layer, one or more fully connected (dense) layers are employed for classification, and a dropout layer can be incorporated to mitigate overfitting, thus enhancing the model's generalization capacity.

5) Model Compilation:

In model compilation, Binary classification tasks such as fraud detection employ the binary cross-entropy loss function. The Adam optimizer modifies the model's parameters to effectively reduce the loss. While accuracy is the primary measure for evaluation, metrics such as precision, recall, and F1-score are also crucial, particularly when dealing with imbalanced data. These metrics help assess the model's performance in distinguishing fraudulent from legitimate posts.

6) Model Training:

After the model's compilation, it undergoes training utilizing the prepared training dataset Throughout the training process, the model iteratively modifies its weights to reduce the loss function across numerous epochs, each epoch entailing a complete traversal of the training dataset The batch size dictates the number of samples processed before the model's weights are updated Post-each epoch, the model undergoes validation to observe its performance and guard against overfitting, with monitoring of loss and accuracy ensuring peak performance [8].

7) Model Evaluation:

Subsequent to training, the model's capacity for generalization is assessed by evaluating its performance on the test set Performance is analysed using essential metrics such as accuracy, precision, recall, and the F1-score, alongside a confusion matrix for better interpretation. These assessments aid in deciding whether the model requires additional fine-tuning or retraining. These metrics offer valuable insights into the model's efficacy, highlighting areas of strength and weakness, which in turn inform whether further optimization or retraining is necessary.

4.1.3. SYSTEM ARCHITECTURE:

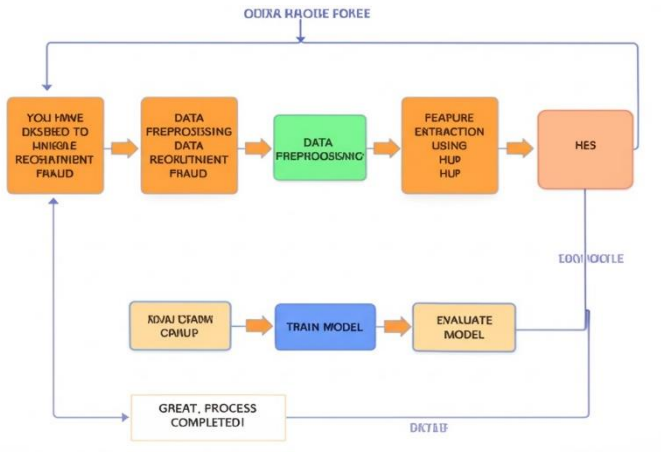

Fig 1: Architecture

1. Decision and Initialization

- **Component**: User Decision
- **Description**: The process begins when the user decides to tackle online recruitment fraud, initializing the system workflow.

2. Data Collection

- **Component**: Start Data Collection
- **Description**: Job postings are collected from various online sources (e.g., LinkedIn, Indeed, Glassdoor). This dataset includes both real and fake job listings, which will be used to train the model [9].

3. Data Preprocessing

- **Component**: Preprocessing Checkpoint
- **Description**: Raw text data is cleaned using Natural Language Processing (NLP) techniques such as:
  - Lowercasing
  - Removal of special characters, HTML tags, and punctuation
  - Tokenization
  - Stop word removal
  - Lemmatization
- **Decision Box**: If preprocessing is skipped, the system routes back to enforce it, ensuring data quality.

4. Feature Extraction

- **Component**: Feature Extraction
- **Description**: Methods like TF-IDF, Bag of Words, and Word Embeddings (such as Word2Vec/GloVe) are

employed to transform textual job information into numerical vectors that encapsulate linguistic and semantic features.

5. Model Construction

- **Component**: Build Model Using NLP
- **Description**: A Long Short-Term Memory (LSTM) neural network is constructed.
. This network identifies contextual relationships within job descriptions by capitalizing on the sequential characteristics of textual data [10].

6. Model Training

- **Component**: Train Model
- **Description**: The model undergoes training on the dataset, having been reprocessed and vectorized, is utilized. To address class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) method is applied, creating synthetic instances for the minority class (fraudulent posts)

7. Model Evaluation

- **Component**: Evaluate Model
- **Description**: The model's effectiveness in identifying fraudulent posts is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and a confusion matrix.

8. Completion

- **Component**: Finalization
- **Description**: Once the model is evaluated and validated, the process concludes with a system ready for deployment or real-time use.

4.1.4 RESULT

Fig2: Interface of ORF detecting system

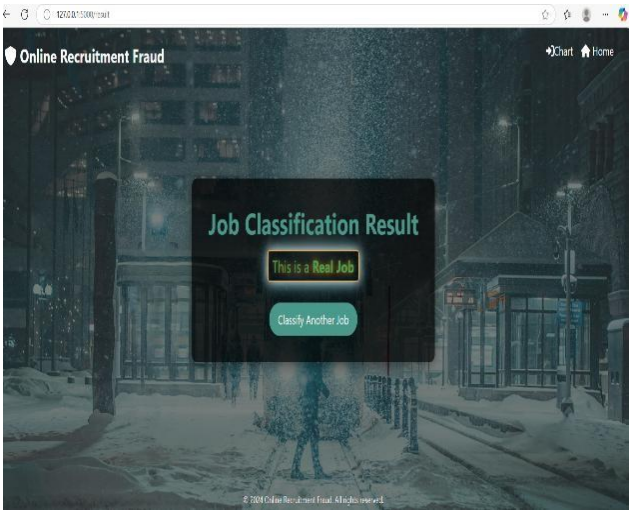

Fig3: Real or Fake job Description Classifier

4.2 PROPOSED TECHNIQUE USED

In our recruitment fraud detection project, NLTK (Natural Language Toolkit) is proposed as the core system for processing and analysing textual job posting data. NLTK provides a comprehensive suite of tools for natural language processing (NLP), allowing us to effectively handle tasks like tokenization, stemming, lemmatization, and removing stop words from the job descriptions. These preprocessing techniques help clean and normalize the text data, making it more suitable for training machine learning models. Additionally, NLTK's sentiment analysis and part-of-speech tagging can be used to detect suspicious or fraudulent language patterns in job postings. By leveraging NLTK's robust NLP capabilities the model's capability to detect and categorize fake job advertisements can be significantly improved for better accuracy**NLTK is a robust toolkit widely used in NLP for analysing and processing human language data effectively It offers user-friendly access to more than 50 datasets and linguistic resources., along with a wide range of text processing libraries for classification, tokenization, stemming, tagging, parsing, and more. NLTK is widely used for tasks like text preprocessing, feature extraction, and linguistic analysis. It enables researchers and developers to build powerful language models and perform various NLP tasks efficiently [12].

5. FUTURE ENHANCEMENT

a. **Expanding the Dataset**
   - Broaden the scope of data by collecting more current and varied job postings and recruitment content [13].
   - This helps the model recognize a wider array of fraudulent strategies and improves its ability to generalize across different scenarios.

b. **Integrating Multi-modal Inputs**
   - Supplement text data with other relevant information, such as:
   - These additional inputs provide richer context, enhancing the model's accuracy in detecting fraudulent posts.

c. **Utilizing Advanced NLP Approaches**
   - Apply enhanced natural language processing methods, including:
     - Improved tokenization
     - Named Entity Recognition (NER)
     - Sentiment and emotion analysis
   - These tools allow the model to understand deeper language cues, such as hidden intentions or sarcasm, often present in scam postings.

d. **Handling Data Imbalance**
   - Implement techniques such as:
     - Creating synthetic examples for the minority (fraud) class
     - Using advanced sampling strategies
   - These methods ensure better recognition of rare fraudulent cases and improve detection rates.

e. **Ongoing Model Optimization**
   - Regularly retrain and update the model to reflect changes in scam techniques.
   - This continuous improvement keeps the system accurate, relevant, and adaptable over time.
   -

6. 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, TFIDF, and leveraging NLTK for tokenization, stop word removal, and stemming, the system can efficiently analyse 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. he 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.

7. REFERENCE:

[1] Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available: https://www.cyber.gov.au/acsc/report

[2] 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.

[3] 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.

[4] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker,and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Netw., vol. 21, nos. 2–3, pp. 427–436, Mar. 2008.

[5] B. Alghamdi and F. Alharby, "An intelligent model for online recruitment fraud detection," J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[6] Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification," in Proc. 3rd Int. Symp. Inf. Process., Luxor, Egypt, Oct. 2010, pp. 301–305.

[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] 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.

[9] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusiondetection methods," IEEE Trans.Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.

[10] P. Kaur, "E-recruitment: A conceptual study," Int. J. Appl. Res., vol. 1, no. 8, pp. 78–82, 2015.

[11] 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.

[12] 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.
[13] 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-millennialsseniors-victims-job-scams