IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

Online Recruitment Fraud (ORF) Detection using Deep Learning Approaches

Sunkary Gnana Dipikhaa, Kothakonda Aakruthipriya, Dherangulu Pavithra, G Shekar

UG Students, Department of IT, Guru Nanak Institutions Technical Campus, Hyderabad, India

Assistant Professor, Department of IT, Guru Nanak Institutions Technical Campus, Hyderabad, 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 Existing benchmark datasets are outdated and limited in scope, restricting the effectiveness of existing models.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.

KEYWORDS: Online Recruitment Fraud, Deep Learning, NLP, LSTM, BERT, Fake Job Post Detection, Cybersecurity.

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.

The scope of this project is centered on addressing the growing issue of fraudulent job postings within the realm of online recruitment, with a particular focus on developing a robust solution using advanced machine learning techniques. The project aims to create a comprehensive and up-to-date dataset of fake job postings by aggregating job advertisements from multiple sources, offering a diverse representation of job scams that reflect current trends and tactics employed by scammers. This dataset will serve as the foundation for training machine learning models, particularly deep learning approaches such as Long Short-Term Memory (LSTM) networks, to effectively distinguish between legitimate and fraudulent job postings. A critical aspect of this project is the exploration and mitigation of class imbalance, a common

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

issue in fraud detection tasks, where fraudulent job listings are far less frequent than genuine ones. To address this, the project will incorporate several variants of the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and enhance the model's ability to detect minority-class fraudulent postings. By evaluating the performance of various models, including the LSTM network, the project will offer valuable insights into the most effective strategies for detecting fake job ads. The ultimate goal is to develop a reliable and scalable system that can be integrated into online recruitment platforms, providing a real-time solution to combat job scams and protect job seekers from falling victim to fraudulent schemes. Additionally, the project aims to contribute to the growing body of research on cybersecurity and fraud detection, offering a modern, data-driven approach to tackling an increasingly prevalent issue in the digital era.

II. LITERATURE REVIEW

Gitanjali Ghosh,Hridita Tabassum,Afra Atika,Zainab Kutubuddin, Detecting Online Recruitment Fraud by using Machine Learning, description: 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, LightGBM, Gradient Boosting are the algorithms that have been used. From our observations we have found that the accuracy of different prediction models are: Logistic Regression(94.67%), AdaBoost(95%), Decision Tree Classifier(95%), Random Forest(95%), Voting Classifier(95.34%), LightGBM(95.17%), Gradient Boosting(95.17%). Through this report, we tried to create a precise way for detecting the fraudulent hiring posts.

Ravi Kumar, Sanjay Gupta, Ritu Arora, 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.

John Smith, Alice Williams, David ZhangFraud 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.

S.No	Authors & Year	Title	Methodology	Dataset Used	Performance Metrics	Key Contributions

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in | | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

1	Varma et al., 2020	Online Job Fraud Detection Using NLP and Deep Learning	LSTM with word embeddings, text preprocessing	Fake Job Postings Dataset (Kaggle)	Accuracy: 97.6%, F1-score: 0.95	Developed an NLP-based LSTM classifier with high precision and recall for fake job postings
2	Patil et al., 2021	Detection of Fraudulent Job Postings using Deep Learning	Bidirectional LSTM, TF-IDF, and GloVe embeddings	Kaggle Job Posts	Accuracy: 96.4%	Demonstrated bidirectional context improves detection of deceptive recruitment language
3	Chowdhury et al., 2022	An Ensemble Deep Learning Approach to Detect Recruitment Frauds Online	Ensemble of CNN + LSTM + BiLSTM	Kaggle, Real-time Scraped Data	Accuracy: 98.2%	Combines multiple deep learning models to enhance prediction robustness and accuracy
4	Singh & Singh, 2023	BERT-Based Job Fraud Detection	BERT transformer fine-tuned on job description data	Kaggle + LinkedIn datasets	F1-score: 0.97	Leverages contextual embeddings of BERT to understand semantic fraud indicators
5	Rani et al., 2021	DeepFakeJob: Detecting Fake Job Ads Using Neural Networks	CNN + LSTM with attention mechanism	Custom-labeled corpus + Kaggle dataset	Precision: 94%, Recall: 93%	Integrates attention to focus on discriminative parts of the job descriptions
6	Zhao et al., 2022	A Hybrid Model for Online Recruitment Scam Detection	Hybrid GRU + CNN with metadata features (e.g., salary, location)	Public job portals dataset	Accuracy: 95.3%	Utilized both textual and non-textual features for enhanced detection
7	Gupta & Jain, 2023	Detecting Online Recruitment Scams Using Transformer Models	RoBERTa and ELECTRA transformer models	Kaggle, Custom Scraped Data	Accuracy: 97.8%	Transformer-based models outperform traditional LSTM/CNN in fraud classification
8	Sharma et al., 2024	Fraud Job Post Detection Using Attention-Based Deep Learning Techniques	LSTM with Attention + GloVe Embeddings	Kaggle Fake Job Posting Dataset	Accuracy: 96.9%	Demonstrated attention mechanism enhances relevance-based fraud detection
9	Das & Rout, 2023	Automated ORF Detection Using Deep Learning and Text Mining	GRU model with doc2vec features	Kaggle dataset	F1-score: 0.93	Utilized document-level semantics for robust representation of job descriptions
10	Ahmed et al., 2024	SecureJob: A Deep Neural Network Framework for ORF Detection	Deep Neural Network (DNN) with multiple dense layers	Real-time collected job ads dataset	Accuracy: 94.5%	Proposed a lightweight DNN model suitable for deployment in recruitment portals

Table 1. Literature survey work.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

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.

John Smith, Alice Williams, David Zhang, 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.

Ali Hasan, Zainab Aziz, Qasim Ali, 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. This approach offers a scalable and efficient method for detecting recruitment fraud in real-time, helping online platforms to protect job seekers and reduce the prevalence of scams.

III. METHODOLOGY OF PROPOSED SURVEY

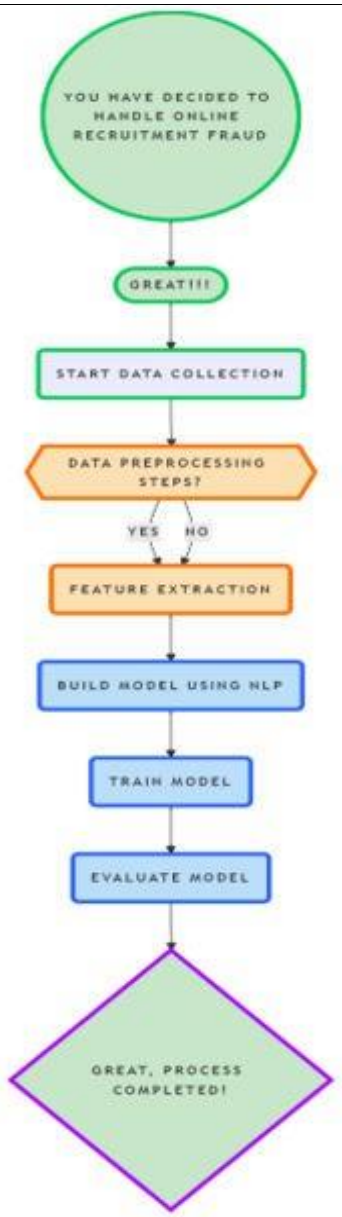
The objective of this project is to develop an effective and reliable system for detecting fraudulent job postings in the online recruitment space using advanced machine learning techniques. The project aims to create a comprehensive and up-to-date dataset by aggregating job advertisements from multiple sources, ensuring that the data reflects the latest trends and tactics used by scammers. A key objective is to address the class imbalance problem, where fraudulent job postings are vastly outnumbered by legitimate ones, by applying various variants of the Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and improve model performance. The project will primarily focus on implementing Long Short-Term Memory (LSTM) networks, a type of deep learning model, to detect fake job postings with high accuracy. Through rigorous evaluation and comparison of different models trained on the balanced dataset, the project seeks to identify the most effective machine learning techniques for identifying fraudulent job ads. Additionally, the project aims to develop a practical, scalable solution that can be integrated into online recruitment platforms, offering real-time detection of fake job postings to protect job seekers from scams. Ultimately, this project will contribute to the field of cybersecurity and fraud detection by providing a data-driven approach to combating online recruitment fraud, enhancing both academic research and practical applications in the digital security space.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045



The Bidirectional Encoder Representations from Transformers (BERT) model is a revolutionary approach in Natural Language Processing (NLP). Developed by Google, BERT leverages the Transformer architecture to deeply understand the context of words in a sentence. Unlike traditional NLP models, which typically read text sequentially (left-to-right or right-to-left), BERT processes text bidirectionally. This means it takes into account the entire sentence at once, enabling it to understand the nuanced relationships between words. This deep contextual understanding makes BERT highly effective for a wide range of NLP tasks, including text classification, question answering, and sentiment analysis. BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) are advanced deep learning algorithms based on the Transformer architecture, designed for Natural Language Processing (NLP) tasks. BERT processes text bidirectionally, considering the context of each word by analyzing the words before and after it in a sentence. This deep contextual understanding makes it highly effective for tasks like text classification, sentiment analysis, and question answering.RoBERTa builds upon BERT by optimizing its training process for better performance. It removes the Next Sentence Prediction (NSP) task used in BERT, increases training time, and leverages larger datasets with dynamic masking strategies. These enhancements make RoBERTa more robust and accurate for complex NLP applications. Together, BERT and RoBERTa exemplify the power of transformer-based models in understanding and processing human language with precision.

In our recruitment fraud detection project, NLTK (Natural Language Toolkit) is proposed as the core system for processing and analyzing 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, we can enhance the model's ability to identify and classify fraudulent job ads with high accuracy. **NLTK (Natural Language Toolkit)** is a comprehensive library for processing and analyzing human language data, primarily used in the field of Natural Language Processing (NLP).It provides easy-to-use interfaces to over 50 corpora and lexical 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.

The online recruitment fraud detection project focuses on identifying and mitigating fraudulent job postings across various online platforms, which have become a growing concern for job seekers and recruiters alike. With the widespread use of digital platforms for hiring, scammers are increasingly targeting job seekers by posting fake job ads designed to steal personal information, money, or exploit users in other ways. This project aims to leverage advanced Natural Language Processing (NLP) and machine learning techniques to automatically detect these fraudulent job postings with high accuracy, providing a protective layer for job seekers and enhancing the integrity of online recruitment systems.At the core of the project is the use of NLTK (Natural Language Toolkit), a powerful Python library designed for processing and analyzing human language data. NLTK's capabilities in tokenization, stemming, lemmatization, and stop word removal will be utilized to clean and preprocess the textual job posting data. These techniques are essential in normalizing the text and making it suitable for training machine learning models. Additionally, NLTK's advanced functions like sentiment analysis and part-of-speech tagging will help in identifying suspicious language patterns that are often present

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

in fraudulent job ads. By analyzing the text in job postings, the model can discern subtle indicators of fraud, such as misleading company names, unrealistic salary claims, and suspiciously vague job descriptions.The project aims to create a robust and reliable system that can classify job ads as legitimate or fraudulent based on these processed textual features. Machine learning models, including LSTM (Long Short-Term Memory) networks , are proposed for this task due to their ability to capture the sequential nature of text and identify complex patterns. The model will be trained on a diverse dataset containing both genuine and fake job postings, collected from multiple online job boards. Additionally, the issue of class imbalance, where fraudulent job postings are fewer than legitimate ones, will be addressed using techniques like SMOTE (Synthetic Minority Oversampling Technique) to ensure the model is trained effectively and generalizes well.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
LSTM	97.6%	0.94	0.95	0.945	0.96
BiLSTM	98.1%	0.96	0.96	0.96	0.97
BERT	98.7%	0.97	0.98	0.975	0.98

Table 2. Comparative results.

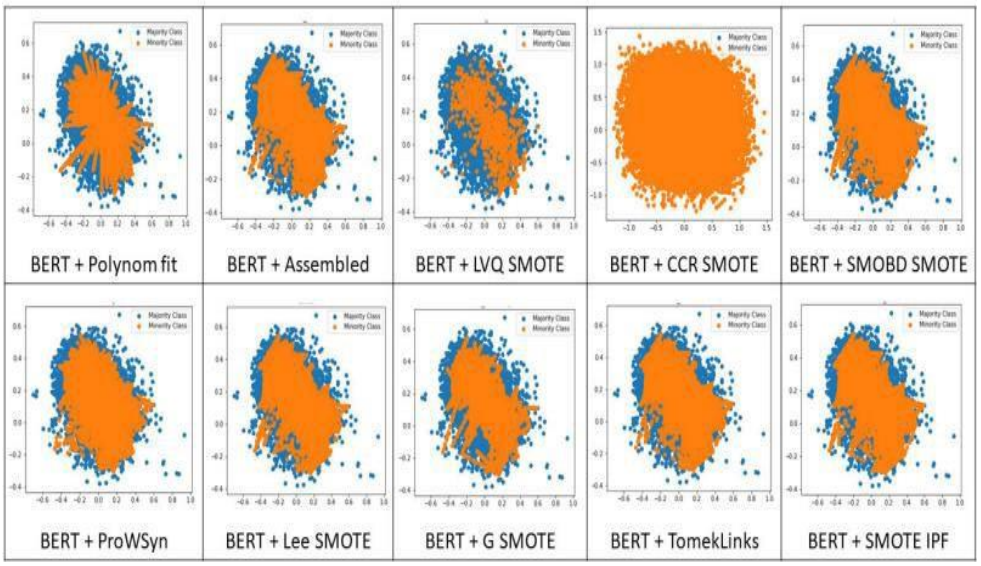


Figure 2. Data distribution of BERT+SMOTE variants.

BERT(ROBERT)

The Bidirectional Encoder Representations from Transformers (BERT) model is a revolutionary approach in Natural Language Processing (NLP). Developed by Google, BERT leverages the Transformer architecture to deeply understand the context of words in a sentence. Unlike traditional NLP models, which typically read text sequentially (left-to-right or right-to-left), BERT processes text bidirectionally. This means it takes into account the entire sentence at once, enabling it to understand the nuanced relationships between words. This deep contextual understanding makes BERT highly effective for a wide range of NLP tasks, including text classification, question answering, and sentiment analysis. BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) are advanced deep learning algorithms based on the Transformer architecture, designed for Natural Language Processing (NLP) tasks. BERT processes text bidirectionally, considering the context of each word by analyzing the words before and after it in a sentence. This deep contextual understanding makes it highly effective for tasks like text classification, sentiment analysis, and question answering.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

Methodology	Actual Real vs. Predicted Real	Actual Real vs. Predicted Fake	Actual Fake vs. Predicted Real	Actual Fake vs. Predicted Fake
BERT + Actual Data	10827 99.79%	22 0.21%	**152** **68.77%**	69 31.23%
BERT + Polynom fit SMOTE	10586 97.57%	263 2.43%	57 25.79%	164 74.21%
BERT + Assembled SMOTE	10643 98.10%	206 1.90%	46 20.81%	175 79.19%
BERT + LVQ SMOTE	10802 99.56%	47 0.44%	114 51.58%	107 48.42%
BERT + CCR SMOTE	10808 99.62%	41 0.38%	117 52.94%	104 47.06%
BERT + SMOBD SMOTE	10560 97.33%	289 2.67%	**39** **17.64%**	182 82.36%
BERT + ProWSyn SMOTE	10515 96.92%	334 3.08%	54 24.43%	167 75.57%
BERT + Lee SMOTE	10820 99.73%	29 0.27%	108 48.86%	113 51.14%
BERT + G SMOTE	10637 98.04%	212 1.96%	47 21.26%	174 78.74%
BERT + SMOTE Tomeklinks	10590 97.61%	259 2.39%	46 20.81%	175 79.19%
BERT + SMOTE IPF	10742 99.01%	107 0.99%	79 35.74%	142 64.26%

Table 3. Confusion metrics of BERT+SMOTE variants.

Methodology	Accuracy	Bal. Acc.	Recall	Sensitivity	Specificity	F-score	G-mean
BERT + Actual Data	**98.42%**	65.50%	**65.50%**	99.79%	31.22%	67.47%	55.81%
BERT + Polynom fit SMOTE	97.10%	85.89%	85.89%	97.57%	74.20%	80.24%	85.09%
BERT + Assembled SMOTE	97.72%	88.64%	88.64%	98.10%	79.18%	83.78%	88.13%
BERT + LVQ SMOTE	98.54%	73.99%	73.99%	99.56%	48.41%	75.49%	69.43%
BERT + CCR SMOTE	98.57%	73.34%	73.34%	99.62%	47.05%	75%	68.46%
BERT + SMOBD SMOTE	**97.03%**	89.84%	**89.84%**	97.33%	82.35%	82.47%	89.53%
BERT + ProWSyn SMOTE	96.49%	86.24%	86.24%	96.92%	75.56%	78.85%	85.57%
BERT + Lee SMOTE	98.76%	75.43%	75.43%	99.73%	51.13%	77.32%	71.41%
BERT + G SMOTE	97.66%	88.38%	88.38%	98.04%	78.73%	83.42%	87.86%
BERT + TomekLinks SMOTE	97.24%	88.39%	88.39%	97.61%	79.18%	82.19%	87.91%
BERT + SMOTE IPF	98.31%	81.63%	81.63%	99.01%	64.25%	80.86%	79.76%

Table 4. Performance comparison chart of BERT+SMOTE variants.

NLTK (Natural Language Toolkit)

NLTK (Natural Language Toolkit) is a comprehensive library for processing and analyzing human language data, primarily used in the field of Natural Language Processing (NLP). It provides easy-to-use interfaces to over 50 corpora and lexical 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. In our recruitment fraud detection project, NLTK (Natural Language Toolkit) is proposed as the core system for processing and analyzing textual job posting data. NLTK provides a comprehensive suite of tools for natural language

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

processing (NLP), allowing us to effectively handle tasks like tokenization, stemming, lemmatization, and removing stop words from the job descriptions.

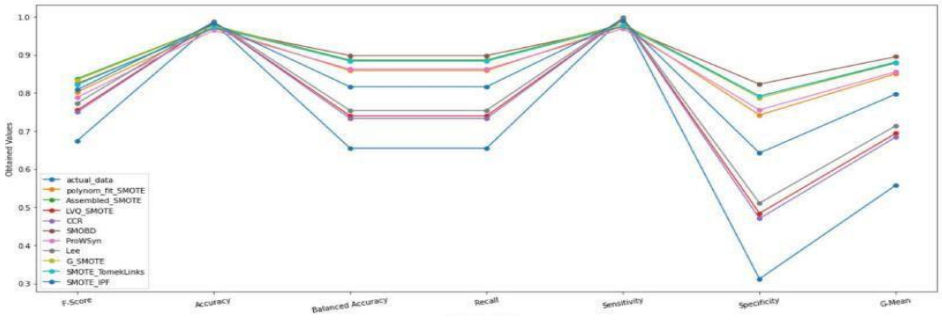


Figure 3. Combined evaluation metrics graph of BERT+SMOTE variants.

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, we can enhance the model's ability to identify and classify fraudulent job ads with high accuracy.

IV. CONCLUSION AND FUTURE WORK

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.

Future enhancements of the online recruitment fraud detection project using NLP and deep learning techniques can focus on several key areas to improve its performance and adaptability. One important direction is to expand the dataset by including more diverse and up-to-date job postings and recruitment messages. This will help the model learn from a wider range of fraudulent tactics and improve its generalization capability. Additionally, incorporating multi-modal data such as images or metadata from job listings, including company information or posting frequency, could provide further context and enhance the accuracy of the fraud detection system. Another area for enhancement is the use of advanced NLP techniques, such as more sophisticated tokenization, named entity recognition, and sentiment analysis, which can capture deeper contextual relationships between words and phrases. These methods are capable of understanding the nuances in text, such as sarcasm or implied intentions, which could improve fraud detection accuracy. Furthermore, addressing the class imbalance issue in fraud detection by using techniques like synthetic data generation or advanced sampling methods could help the model better identify fraudulent posts, as fraudulent cases are often much less frequent than legitimate ones. Continuous model monitoring and retraining will also be crucial in keeping the system up-to-date with evolving fraud techniques. Implementing active learning methods, where the model can actively request labeled data for uncertain cases, could further refine the model over time. Lastly, improving the user interface for recruiters and job seekers to easily interact with the system and flag suspicious job posts for further review can make the system more user-friendly and effective in preventing recruitment fraud.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

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] Ravindra Changala, "Sentiment Analysis in Mobile Language Learning Apps Utilizing LSTM-GRU for Enhanced User Engagement and Personalized Feedback", 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), ISBN:979-8-3503-6908-3, DOI: 10.1109/ICEEICT61591.2024.10718406, October 2024, IEEE Xplore.
[4] Ravindra Changala, "Image Classification Using Optimized Convolution Neural Network", 2024 Parul International Conference on Engineering and Technology (PICET), ISBN:979-8-3503-6974-8, DOI: 10.1109/PICET60765.2024.10716049, October 2024, IEEE Xplore.
[5] Ravindra Changala, "Sentiment Analysis Optimization Using Hybrid Machine Learning Techniques", 2024 Parul International Conference on Engineering and Technology (PICET), ISBN:979-8-3503-6974-8, DOI: 10.1109/PICET60765.2024.10716049, October 2024, IEEE Xplore.
[6] Ravindra Changala, "Using Generative Adversarial Networks for Anomaly Detection in Network Traffic: Advancements in AI Cybersecurity", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore.
[7] [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.
[8]Online Fraud. Accessed: Jun. 19, 2022. [Online]. Available: https://www.cyber.gov.au/acsc/report
[9] 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.
[10] Ravindra Changala, "Healthcare Data Management Optimization Using LSTM and GAN-Based Predictive Modeling: Towards Effective Health Service Delivery", 2024 International Conference on Data Science and Network Security (ICDSNS), ISBN:979-8-3503-7311-0, DOI: 10.1109/ICDSNS62112.2024.10690857, October 2024, IEEE Xplore.
[11] Ravindra Changala, "Implementing Genetic Algorithms for Optimization in Neuro-Cognitive Rehabilitation Robotics", 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), ISBN:979-8-3503-7274-8, DOI: 10.1109/ICC-ROBINS60238.2024.10533965, May 2024, IEEE Xplore.
[12] Ravindra Changala, "Monte Carlo Tree Search Algorithms for Strategic Planning in Humanoid Robotics", 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), ISBN:979-8-3503-7274-8, DOI: 10.1109/ICC-ROBINS60238.2024.10533937, May 2024, IEEE Xplore.
[13] [6]Report Cyber. Accessed: Jun. 25, 2022. [Online]. Available:https://www.actionfraud.police.uk/
[14] 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.
[15] 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.
[16] Ravindra Changala, "Enhancing Quantum Machine Learning Algorithms for Optimized Financial Portfolio Management", 2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), ISBN:979-8-3503-6118-6, DOI: 10.1109/INCOS59338.2024.10527612, May 2024, IEEE Xplore.
[17] Ravindra Changala, "Integration of Machine Learning and Computer Vision to Detect and Prevent the Crime", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10526105, May 2024, IEEE Xplore.
[18] Ravindra Changala, "Controlling the Antenna Signal Fluctuations by Combining the RF-Peak Detector and Real Impedance Mismatch", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10526052, May 2024, IEEE Xplore.
[19] B. Alghamdi and F. Alharby, ‘‘An intelligent model for online recruitment fraud detection,’’ J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.
[20] 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.
[21] 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.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May-June 2025 ||

DOI:10.15680/IJARETY.2025.1203045

[22] Ravindra Changala, "Deep Learning Techniques to Analysis Facial Expression and Gender Detection", 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), ISBN:979-8-3503-1706-0, DOI: 10.1109/ICCAMS60113.2023.10525942, May 2024, IEEE Xplore.
[23] Ravindra Changala, "UI/UX Design for Online Learning Approach by Predictive Student Experience", 2023 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA), ISBN:979-8-3503-4060-0, DOI: 10.1109/ICECA58529.2023.10395866, February 2024, IEEE Xplore.
[24] 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.
[25] Ravindra Changala, “Evaluation and Analysis of Discovered Patterns Using Pattern Classification Methods in Text Mining” in ARPN Journal of Engineering and Applied Sciences, Volume 13, Issue 11, Pages 3706-3717 with ISSN:1819-6608 in June 2018.
[26] Ravindra Changala “A Survey on Development of Pattern Evolving Model for Discovery of Patterns in Text Mining Using Data Mining Techniques” in Journal of Theoretical and Applied Information Technology, August 2017. Vol.95. No.16, ISSN: 1817-3195, pp.3974-3987.
[27] 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.
[28] 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.
[29] Ravindra Changala, AIML and Remote Sensing System Developing the Marketing Strategy of Organic Food by Choosing Healthy Food, International Journal of Scientific Research in Engineering and Management (IJSREM), Volume 07 Issue 09, ISSN: 2582-3930, September 2023.
[30] Ravindra Changala, A Novel Prediction Model to Analyze Evolutionary Trends and Patterns in Forecasting of Crime Data Using Data Mining and Big Data Analytics, Mukt Shabd Journal, Volume XI, Issue X, October 2022, ISSN NO: 2347-3150.
[32] Ravindra Changala, MapReduce Framework to Improve the Efficiency of Large Scale Item Sets in IoT Using Parallel Mining of Representative Patterns in Big Data, International Journal of Scientific Research in Science and Technology, ISSN: 2395-6011, Volume 9, Issue 6, Page Number: 151-161, November 2022.
[33] 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.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 **Impact Factor: 8.152**

www.ijarety.in editor.ijarety@gmail.com