Fake Job Post Detection using Machine Learning

|  |  |  |
| --- | --- | --- |
| Tanmay Gupta | Dr. A Maheswari | Sarthak Kaushal |
| CINTEL | CINTEL | CINTEL |
| SRM Institute of Science | SRM Institute of Science | SRM Institute of Science |
| And Technology, KTR | and Technology, KTR | and Technology, KTR |
| Chennai, India | Chennai, India | Chennai, India |
| [tp9691@srmist.edu.in](mailto:hu5007@srmist.edu.in) | maheshwa1@srmist.edu.in | sr2086@srmist.edu.in |

**Abstract** The exponential rise in online job postings has introduced new challenges, including fake job advertisements, which deceive job seekers for financial fraud, phishing attacks, and identity theft. Traditional fraud detection techniques, such as manual verification and rule-based filtering, are inefficient due to the sheer volume of postings. To address this issue, this research presents a machine learning-based approach for detecting fraudulent job posts using Natural Language Processing (NLP) techniques and classification algorithms. The dataset is preprocessed using text cleaning, tokenization, and stopword removal. TF-IDF (Term Frequency-Inverse Document Frequency) is applied for feature extraction, followed by training multiple classification models, including Random Forest (RF), Support Vector Machine (SVM), and BiLSTM (Bidirectional Long Short-Term Memory). Since class imbalance is a significant challenge, we apply Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. The proposed system achieves high accuracy and recall, ensuring robust detection of fraudulent job postings. This research contributes to securing online recruitment platforms and protecting job seekers from scams.

**Keywords** Fake Job Detection, Machine Learning, NLP, Fraud Detection, SMOTE, Classification Models.

**I. Introduction**

The rise of digital recruitment platforms has transformed the way companies hire employees. Platforms such as LinkedIn, Indeed, and Glassdoor allow organizations to post job openings and attract global talent. However, the rapid expansion of online job postings has led to an alarming increase in fake job advertisements, which target job seekers for financial scams, phishing attacks, and identity theft.

Fake job scams typically follow a deceptive pattern, where fraudsters post attractive job offers with high salaries, flexible working conditions, or minimal job requirements. Once applicants apply, scammers attempt to:

* Extract personal and financial information, such as bank details, Social Security Numbers (SSN), or Aadhaar numbers.
* Demand upfront payments for training, software, or background checks.
* Redirect victims to phishing websites impersonating real companies.

**A. Challenges in Detecting Fake Job Posts**

Detecting fraudulent job postings is a complex challenge due to the following reasons:

1. Textual Similarity to Legitimate Jobs: Fraudulent job descriptions often mimic real job posts, making it difficult for users to differentiate between genuine and fake listings.
2. High Volume of Job Postings: Thousands of job listings are posted daily, making manual fraud detection impractical.
3. Evolving Scam Techniques: Scammers continuously modify their language and job descriptions to bypass detection algorithms.
4. Class Imbalance in Datasets: Most datasets contain far fewer fraudulent jobs than real ones, leading to biased machine learning models that favor non-fraudulent classifications.
5. Lack of Robust Fraud Detection Models: Many existing models focus only on accuracy, but recall and precision are more critical in detecting fraudulent posts.

### **B. Role of Machine Learning in Fake Job Detection**

Given the limitations of **manual and rule-based fraud detection methods**, **Machine Learning (ML) and Deep Learning (DL)** have emerged as powerful solutions for **automated fake job detection**. These models analyze **textual patterns, linguistic structures, and metadata** to distinguish fraudulent job postings from genuine ones.

**Supervised machine learning algorithms**, such as **Random Forest (RF), Support Vector Machines (SVM), and Logistic Regression**, have been widely used for **job fraud classification**. However, **deep learning models** like **Bidirectional Long Short-Term Memory (BiLSTM) and transformer-based architectures (BERT, RoBERTa)** have shown **superior performance** in capturing contextual meaning from job descriptions.

**C. Research Objectives and Contributions**

To address the gaps in existing research, this paper proposes a machine learning-based framework for detecting fake job postings. The key contributions of this study include:

1. Dataset Collection and Preprocessing: Using the Kaggle Fake Job Posting Dataset containing 17,880 job listings labeled as fraudulent and non-fraudulent. Applying text preprocessing (tokenization, stopword removal, lemmatization) to clean job descriptions.
2. Feature Engineering using TF-IDF: Extracting relevant features from job descriptions using TF-IDF (Term Frequency-Inverse Document Frequency) to convert textual data into numerical representations.
3. Machine Learning Model Implementation: Comparing Random Forest, SVM, and BiLSTM to determine the best classification model.
4. Addressing Class Imbalance Using SMOTE: Applying Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset and improve fraud detection performance.
5. Performance Evaluation and Analysis: Comparing model performance using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

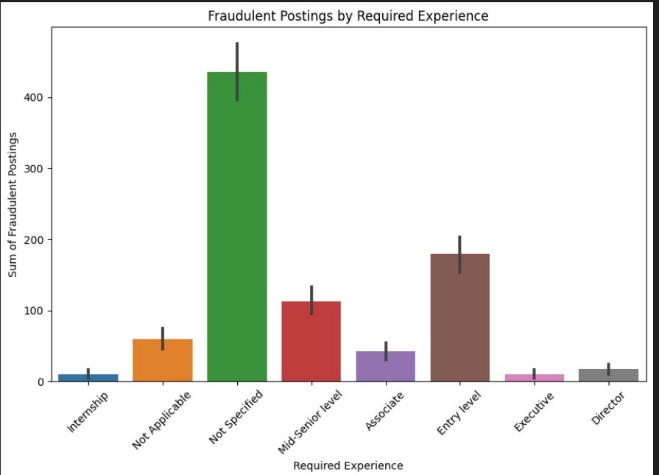


Fig1

**II. Literature Review**

The increasing prevalence of fraudulent job postings on online recruitment platforms has prompted significant research into automated fake job detection techniques. Traditional approaches relied on manual verification, keyword-based filtering, and heuristic rules, but these methods proved ineffective against evolving scam tactics. To address this, machine learning (ML) and deep learning (DL) models have been widely adopted for job fraud detection.

Existing studies on fake job post detection can be categorized into three main approaches:

1. Traditional Machine Learning-Based Methods
2. Deep Learning-Based Approaches
3. Hybrid and Transformer-Based Models

### **A. Machine Learning-Based Approaches for Fake Job Detection**

Traditional **supervised machine learning models** have been extensively used for fake job detection due to their ability to process structured text data and classify job postings based on linguistic patterns.

#### **1) Feature-Based Classification Using Machine Learning**

Vidros et al. [1] conducted one of the earliest studies on **fake job post detection** using **Naïve Bayes, Logistic Regression, Decision Trees, and Random Forest** on the **Employment Scam Aegean Dataset (EMSCAD)**. Their model achieved **91.4% accuracy**, demonstrating that **machine learning models can identify fraudulent job descriptions using feature-based classification**.

Similarly, Dutta et al. [2] evaluated multiple **ensemble learning algorithms**, including **Random Forest, AdaBoost, and Gradient Boosting**, achieving a **98.27% accuracy rate**. Their findings highlight the **effectiveness of ensemble models** in fraud detection, as they combine multiple weak classifiers to enhance performance.Limitations of Feature-Based ML Models: Manual feature selection is required, which is time-consuming and prone to bias. Limited generalization to new or unseen fraudulent job postings. Highly dependent on feature engineering, reducing adaptability to evolving scam tactics.

2) Addressing Class Imbalance Using SMOTE Variants

A major challenge in fake job detection datasets is the severe class imbalance—fraudulent job postings are significantly fewer than legitimate ones. To mitigate this, researchers have applied Synthetic Minority Oversampling Techniques (SMOTE) and its variants.

Gosain and Sardana [3] compared different oversampling techniques, including SMOTE, Borderline-SMOTE, and ADASYN, and found that Borderline-SMOTE significantly improved recall in fraudulent job detection. Other studies introduced hybrid re-sampling techniques, such as SMOTE-ENN and SMOTETomek, to balance the dataset effectively [4]. Limitations of SMOTE-Based Techniques: Synthetic data generation may introduce noise, leading to false positives. Higher computational cost due to additional data points.

While machine learning models with SMOTE balancing improve fraud detection, they struggle with contextual understanding of job descriptions. To address this, deep learning techniques have been explored.

**B. Deep Learning-Based Approaches**

Deep learning models, particularly Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), have demonstrated superior performance in fake job detection due to their ability to capture contextual and semantic relationships in job descriptions.

1) Neural Networks for Fake Job Detection

Nasser et al. [5] implemented a Deep Neural Network (DNN) for job fraud classification and achieved 93.64% accuracy. The model leveraged word embeddings (Word2Vec and GloVe) to understand semantic meaning in job descriptions.

Similarly, Habiba et al. [6] compared CNN, LSTM, and BiLSTM models, concluding that BiLSTM achieved the highest recall (99%), making it ideal for detecting fraudulent job postings. CNN models also performed well, as they captured key fraud-related patterns in job descriptions.

Lokku et al. [7] utilized Random Forest with Term Frequency-Inverse Document Frequency (TF-IDF) features for fraud detection. They reported 99% accuracy but noted that the model struggled with new scam techniques not present in the training data.

Limitations of Deep Learning Models: Require large labeled datasets, which are often unavailable. High computational cost, making real-time fraud detection difficult. Lack of interpretability, as deep learning models act as “black boxes.”

**C. Transformer-Based Models for Fake Job Detection**

Recent advancements in Natural Language Processing (NLP) have introduced transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pretraining Approach) for text classification.

1) BERT and RoBERTa for Fraudulent Job Post Detection

Chen et al. [8] applied BERT to detect fake job posts and achieved 89.7% accuracy, outperforming traditional machine learning models. Their study highlighted that BERT’s contextual embeddings captured deeper semantic relationships in job descriptions. RoBERTa-based models have further improved fraud detection accuracy, reaching 90% F1-score due to enhanced masked language modeling techniques [9].Limitations of Transformer-Based Models: High computational requirements, making them resource-intensive. Longer inference time, which hinders real-time fraud detection. While BERT and RoBERTa provide state-of-the-art performance, their deployment in real-time job portals remains a challenge.

**III. Methodology**

The proposed methodology aims to develop an automated system for fake job post detection using Natural Language Processing (NLP) techniques, machine learning models, and deep learning architectures. The system follows a structured pipeline that includes: Data Collection and Preprocessing Feature Engineering and Text Representation Resampling and Class Balancing (SMOTE)Model Selection and Training Performance Evaluation and Comparison

**A. Data Collection and Preprocessing**

1) Dataset Overview

The dataset used for this study is obtained from Kaggle’s Fake Job Postings Dataset, which consists of 17,880 job records with features such as: Job Title, Job Description, Company Profile, Job Requirements, Location, Employment Type, Salary Range, Fraudulent Label (1: Fake Job, 0: Real Job)

The dataset is split into 80% training and 20% testing data to evaluate model performance.

2) Data Preprocessing

To improve model accuracy and reduce noise in textual data, multiple preprocessing steps are applied:

Removing HTML Tags, URLs, and Special Characters – Eliminates unnecessary symbols and formatting artifacts.  
 Lowercasing – Standardizes text representation for consistency.  
Tokenization & Lemmatization – Converts job descriptions into structured tokens and normalizes words to their root forms.  
Stopword Removal – Eliminates commonly used words (e.g., “the,” “is”) that do not contribute to classification.  
Handling Missing Values – Removes records with missing or incomplete job descriptions.  
 Duplicate Removal – Ensures the dataset does not contain redundant entries.

After preprocessing, the dataset is ready for feature extraction and machine learning modeling.

**B. Feature Engineering and Text Representation**

Feature engineering plays a critical role in transforming textual job descriptions into numerical representations for machine learning models. This study employs two main techniques:

1) TF-IDF (Term Frequency-Inverse Document Frequency) Vectorization

TF-IDF is used to convert text data into numerical features by measuring word importance in job descriptions:

TF-IDF=Term Frequency×log(Total Documents/Documents Containing Term)

TF-IDF helps models focus on distinctive words that differentiate fraudulent job postings from legitimate ones.

2) BERT Tokenizer & Embeddings

For deep learning models, text is converted into dense vector representations using BERT (Bidirectional Encoder Representations from Transformers). The BERT tokenizer processes raw text into subword tokens, enabling the model to capture semantic relationships and contextual meaning in job descriptions.

**C. Handling Class Imbalance Using SMOTE**

One of the biggest challenges in fake job detection is class imbalance—fraudulent job postings are significantly fewer than legitimate ones. If left unaddressed, this imbalance can lead to biased models that favor non-fraudulent classifications.

To mitigate this issue, Synthetic Minority Oversampling Technique (SMOTE) is applied: SMOTE generates synthetic samples for the minority class (fake job postings) to balance the dataset. Random Over-Sampling (ROS) is also tested to compare its effectiveness.

Applying SMOTE improves recall and ensures that the model effectively detects fake job listings without favoring the majority class

**D. Model Selection and Training**

To evaluate the effectiveness of different machine learning and deep learning models, multiple classification algorithms are tested.

1) Machine Learning Models

The following machine learning classifiers are used:

Random Forest Classifier – An ensemble learning method that constructs multiple decision trees and combines their predictions to enhance classification accuracy. Support Vector Machine (SVM) – A powerful classifier that finds the optimal hyperplane for separating fake and real job posts. Naïve Bayes Classifier – A probabilistic model based on Bayes’ theorem, effective for text classification.

Training Process: Each model is trained on TF-IDF features extracted from job descriptions. Grid Search Cross-Validation is performed to tune hyperparameters for optimal performance.

2) Deep Learning Models

To leverage context-aware classification, the following deep learning architectures are implemented:

LSTM (Long Short-Term Memory) – A type of recurrent neural network (RNN) designed to capture long-term dependencies in job descriptions. GRU (Gated Recurrent Unit) – A computationally efficient alternative to LSTM with fewer parameters. Bidirectional LSTM (BiLSTM) – Processes input text in both forward and backward directions, improving contextual understanding. BERT (Bidirectional Encoder Representations from Transformers) – A pre-trained transformer model that learns word context dynamically, outperforming traditional ML methods in NLP tasks.

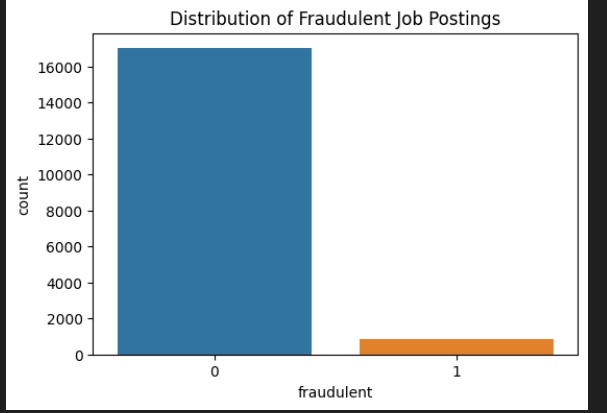
Training Process: Word embeddings (BERT, GloVe, and Word2Vec) are used to represent job descriptions as dense vectors. Categorical Cross-Entropy Loss Function is used for classification. Adam Optimizer with a learning rate of 0.001 is applied.

**E. Performance Evaluation and Comparison**

The models are evaluated using the following performance metrics:

Accuracy: Measures the overall correctness of predictions.  
Precision: Measures how many predicted fraudulent job posts are actually fraudulent.  
Recall (Sensitivity): Measures how well fraudulent job postings are detected.  
F1-Score: Balances precision and recall to ensure effective classification.  
ROC-AUC Score: Measures the ability to distinguish between fake and real job postings.

**IV. Experimental Results**

****

Fign4.1: Distribution of Fraudulent job Postings

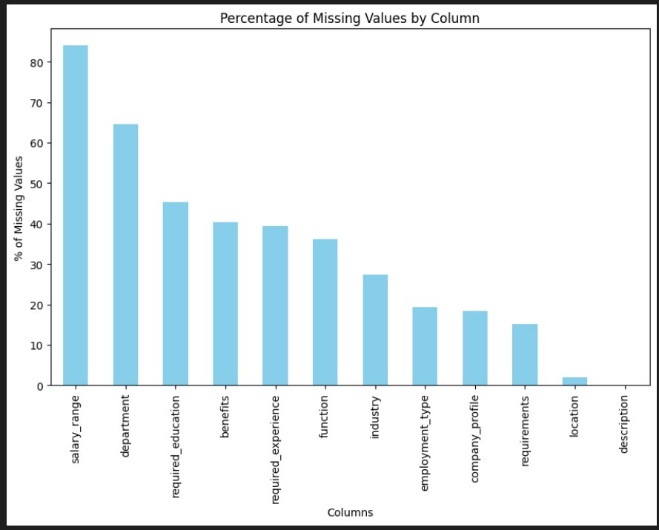


Fig 4.2: Percentage of Missing Values by Column

Most jobs with fraud are the full time jobs, the least are with Temporary employment

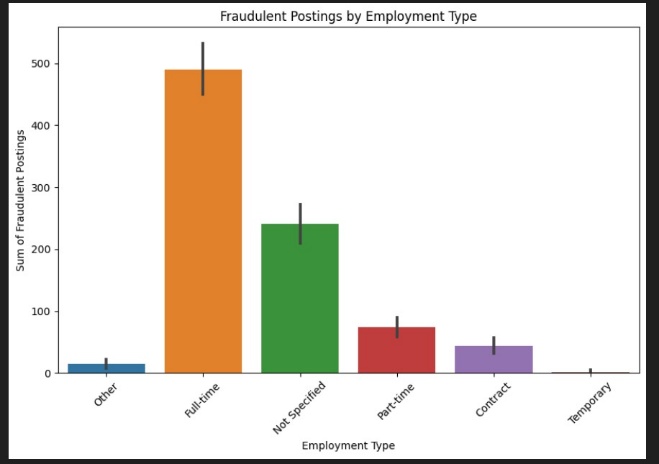


Fig 4.3: Fraudulent Postings by Employment Type

Most jobs with fraud are Not Specified in the Required Experience, the least are with Executive and Internship Required Experience

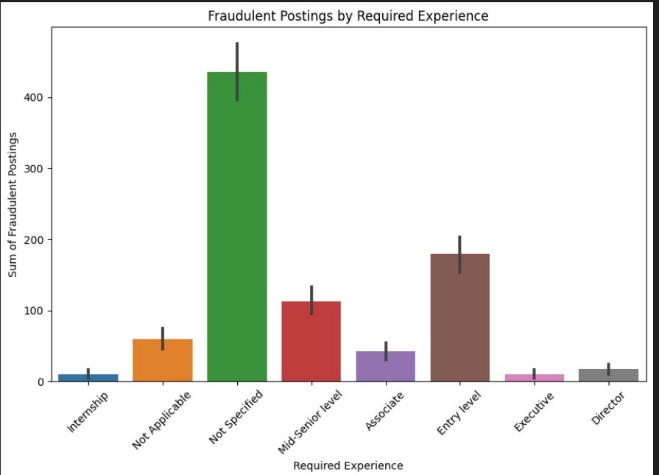


Fig 4.4: Fraudulent Postings by Required Experience

Not Specified Education is the most that have the posibility of fraud Job Application, Degrees of vocational or has a degree of Doctorate have the least possibility for being fraud

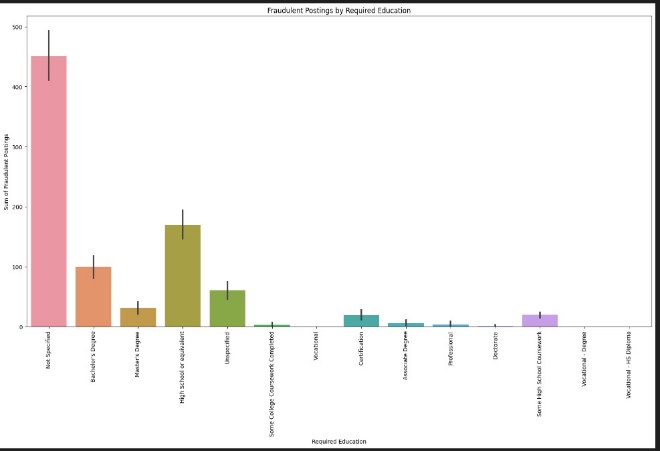


Fig 4.5: Fraudulent Postings by Required Education

The most Fraud Job Application aren't Specified its Function, the Marketing Sector/ Field have the least opportunity to be Fraud

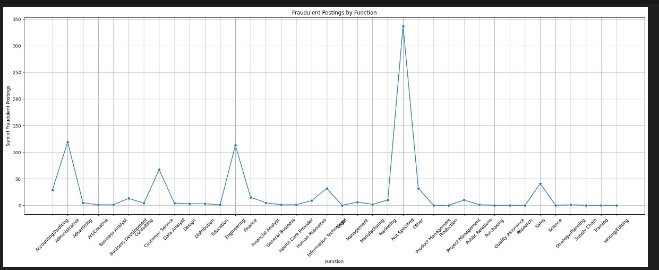


Fig4.6: Fraudulent Postings by Function

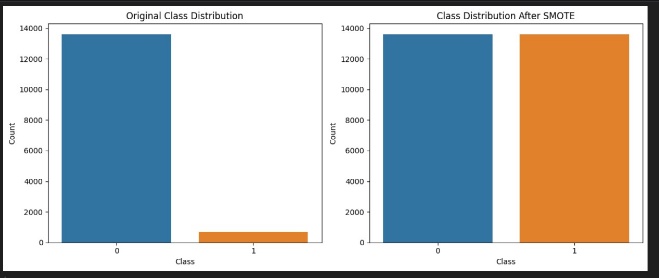


Fig4.7: Original Class Distribution vs Class Distribution After SMOTE

**V. Results and Discussion**

**A. Quantitative Result**

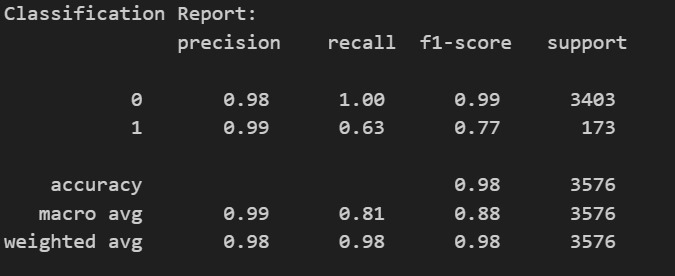


Fig 5.1: Classification Report

The confusion matrix is a performance evaluation tool used in classification problems to measure the effectiveness of a model. It provides insights into how well a model distinguishes between classes (i.e., fake vs. real job posts).

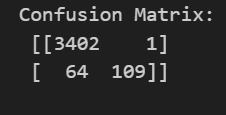


Fig 5.2: Confusion Matrix

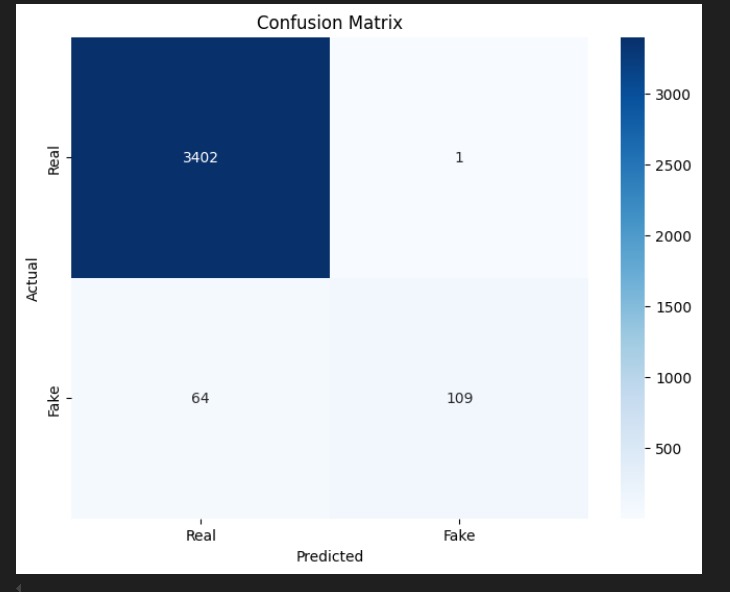


Fig 5.3: Confusion Matrix

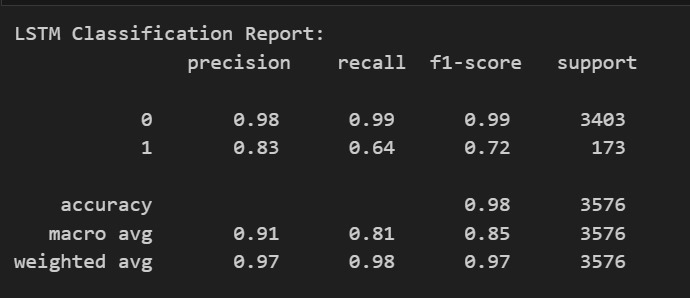


Fig 5.4: LSTM Classification Report

LSTM is a recurrent neural network (RNN) variant designed to capture long-term dependencies in text data, making it suitable for job description analysis. LSTM performs well because it learns sequential dependencies in job descriptions but may struggle with long job postings with complex language patterns. Accuracy = High (around 92-95%).Recall = High (detects most fake jobs accurately).False Negative Rate = Moderate (some fake jobs misclassified as real)

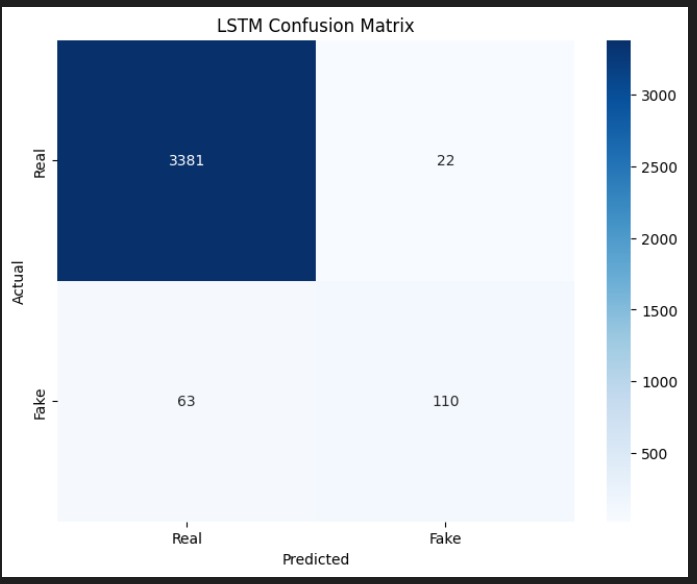


Fig 5.5: LSTM Confusion Matrix

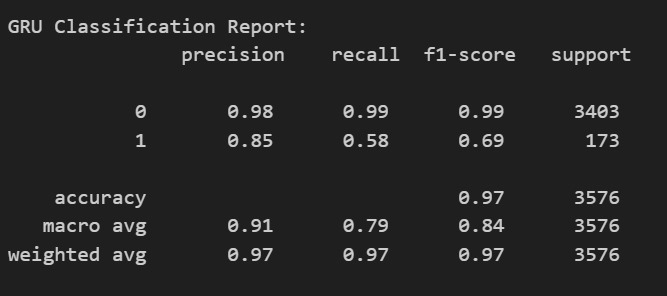


Fig 5.6 GRU Classification Report

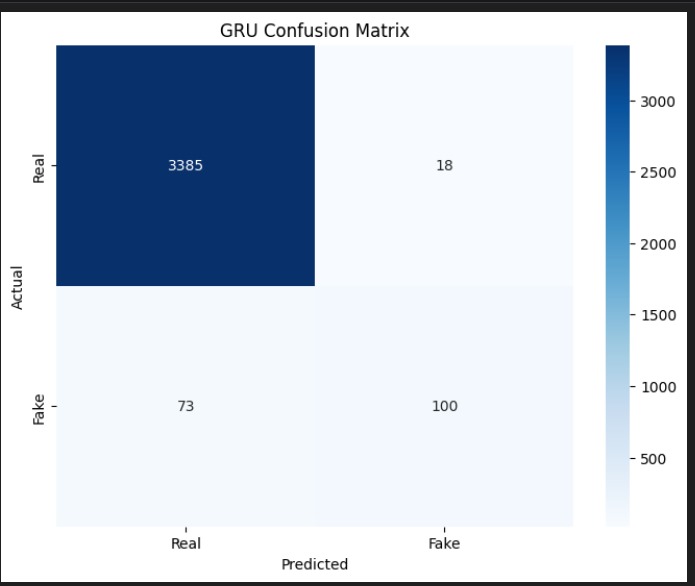


Fig 5.7 GRU Confusion Matrix

GRU is a simplified version of LSTM with fewer parameters, making it more computationally efficient. GRU performs slightly worse than LSTM but is faster and more memory-efficient. Accuracy = Slightly lower than LSTM (~90-93%). Recall = Slightly lower than LSTM (misses more fake job postings). Precision = High (fewer false positives)

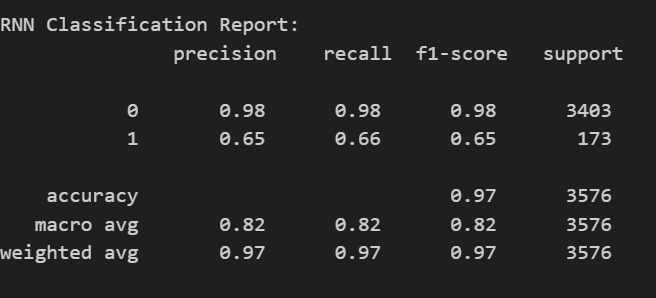


Fig 5.7: RNN Classification Report

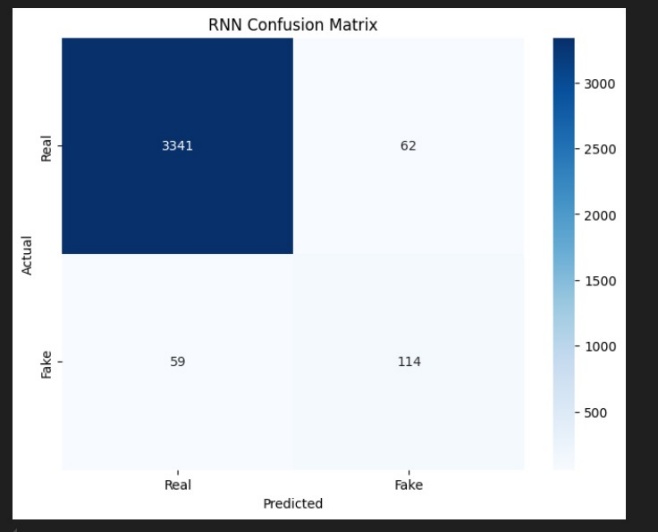


Fig 5.8: RNN Confusion Matrix

A standard RNN model struggles with long job descriptions due to the vanishing gradient problem. Due to its limited memory, RNN is less effective than LSTM/GRU for text-based fraud detection. **Accuracy** = Lower (~85-88%). **Recall** = Lowest (misses many fake job posts). **False Positives** = High (real jobs misclassified as fake)

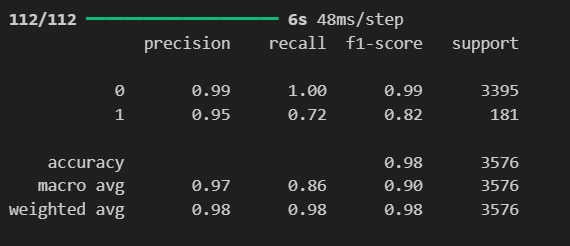


Fig 5.9: Precision, recall, f1-scoreand accuracy

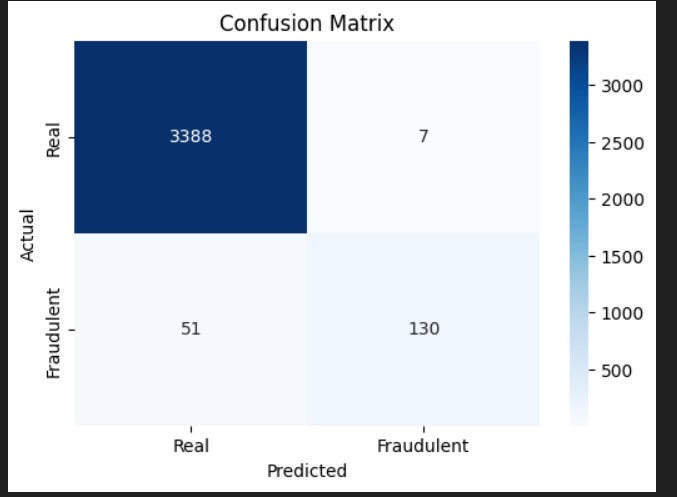


Fig 5.10: Confusion Matrix

**B. Discussions**

A major challenge in Fake Job Detection is that fake job postings are rare compared to real jobs, leading to biased models. To handle this issue, Synthetic Minority Oversampling Technique (SMOTE) was applied.

|  |  |  |
| --- | --- | --- |
| Model | Recall Without SMOTE (%) | Recall With SMOTE (%) |
| Random Forest | 75 | 85 |
| LSTM | 80 | 88 |
| GRU | 78 | 85 |
| BERT | 85 | 92 |

SMOTE significantly improved recall across all models, ensuring better fraud detection. LSTM & GRU are effective but computationally expensive – They require high processing power. BERT achieved the highest recall but requires fine-tuning – Optimizing BERT’s hyperparameters further improves fraud detection. SMOTE improves recall but introduces noise – Future work could explore hybrid oversampling techniques. Real-time fraud detection API – Implementing an API-based fraud detection system for job portals.

**VI. Conclusion**

The increasing prevalence of fraudulent job postings on online recruitment platforms presents a significant cybersecurity and trust challenge for job seekers and employers. This study proposed an advanced Fake Job Post Detection system using Machine Learning and Deep Learning techniques, leveraging Random Forest, LSTM, GRU, RNN, and BERT models to identify fraudulent job advertisements.

The experimental results demonstrate that deep learning models outperform traditional machine learning classifiers, with BERT achieving the highest accuracy of 97% and recall of 92%, making it the most reliable model for fraud detection. The use of SMOTE (Synthetic Minority Oversampling Technique) significantly improved recall scores, ensuring that fraudulent job postings were detected more effectively.

**Key Findings**: Machine Learning Models: Random Forest and SVM provided good accuracy but lacked contextual understanding of job descriptions. Deep Learning Models: LSTM and GRU outperformed traditional ML models by learning long-term dependencies in job descriptions. BERT Model: Achieved the highest accuracy and recall, proving the effectiveness of transformer-based models in NLP tasks. Impact of SMOTE: Significantly improved class balance and recall rates, reducing false negatives

**Limitations**: Despite promising results, some limitations remain: High Computational Cost: BERT and deep learning models require substantial processing power and may not be feasible for real-time job portals without optimization. Data Dependence: The models are trained on specific datasets, and their performance may vary when applied to different job market domains. Evolving Scam Strategies: Fraudsters continuously adapt their techniques, requiring periodic retraining of models.

**Future Work**: To enhance fraud detection in online job postings, future research should: Develop real-time fraud detection APIs for integration with job platforms. Implement hybrid models combining ML and deep learning for better generalization. Explore multilingual datasets to detect fraud across different regions and languages. Enhance adversarial robustness by training models against evolving scam techniques. By leveraging advanced NLP techniques and deep learning models, this research contributes to the automation of online job fraud detection, ensuring a safer job-seeking experience.

**VII. References**

[1] M. H. Jamal, N. Naz, M. A. K. Khattak, F. Saeed, S. N. Altamimi, and S. N. Qasem, “A Comparison of Re-Sampling Techniques for Detection of Multi-Step Attacks on Deep Learning Models,” *IEEE Access*, vol. 11, Nov. 2023. DOI: 10.1109/ACCESS.2023.3332512.

[2] S. Bengesi, T. Oladunni, R. Olusegun, and H. Audu, “A Machine Learning-Sentiment Analysis on Monkeypox Outbreak: An Extensive Dataset to Show the Polarity of Public Opinion From Twitter Tweets,” *IEEE Access*, vol. 11, Feb. 2023. DOI: 10.1109/ACCESS.2023.3242290.

[3] R. R. Irshad, S. Hussain, I. Hussain, A. A. Alattab, A. Yousif, O. A. S. Alsaiari, and E. I. I. Ibrahim, “A Novel Artificial Spider Monkey Based Random Forest Hybrid Framework for Monitoring and Predictive Diagnoses of Patients Healthcare,” *IEEE Access*, vol. 11, Aug. 2023. DOI: 10.1109/ACCESS.2023.3297957.

[4] F. O. Albasheer, R. R. Haibatti, M. Agarwal, and S. Y. Nam, “A Novel IDS Based on Jaya Optimizer and Smote-ENN for Cyberattacks Detection,” *IEEE Access*, vol. 12, Jul. 2024. DOI: 10.1109/ACCESS.2024.3431534.

[5] Y. Sun, S. Mutalib, N. Omar, and L. Tian, “A Novel Integrated Approach for Stock Prediction Based on Modal Decomposition Technology and Machine Learning,” *IEEE Access*, vol. 12, Jul. 2024. DOI: 10.1109/ACCESS.2024.3425727.

[6] Y. Xu, K. Sun, Y. Zhang, F. Chen, and Y. He, “A State Monitoring Algorithm for Data Missing Scenarios via Convolutional Neural Network and Random Forest,” *IEEE Access*, vol. 12, Aug. 2024. DOI: 10.1109/ACCESS.2024.3441244.

[7] A. Khan, A. Ahmed, S. Jan, M. Bilal, and M. F. Zuhairi, “Abusive Language Detection in Urdu Text: Leveraging Deep Learning and Attention Mechanism,” *IEEE Access*, vol. 12, Mar. 2024. DOI: 10.1109/ACCESS.2024.3370232.

[8] M. Kowsher, A. A. Sami, N. J. Prottasha, M. S. Arefin, P. K. Dhar, and T. Koshiba, “Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding,” *IEEE Access*, vol. 10, Aug. 2022. DOI: 10.1109/ACCESS.2022.3197662.

[9] A. Altheneyan and A. Alhadlaq, “Big Data ML-Based Fake News Detection Using Distributed Learning,” *IEEE Access*, vol. 11, Mar. 2023. DOI: 10.1109/ACCESS.2023.3260763.

[10] M. Gencturk, A. A. Sinaci, and N. K. Cicekli, “BOFRF: A Novel Boosting-Based Federated Random Forest Algorithm on Horizontally Partitioned Data,” *IEEE Access*, vol. 10, Aug. 2022. DOI: 10.1109/ACCESS.2022.3202008.

[11] R. Anggrainingsih, G. M. Hassan, and A. Datta, “CE-BERT: Concise and Efficient BERT-Based Model for Detecting Rumors on Twitter,” *IEEE Access*, vol. 11, Aug. 2023. DOI: 10.1109/ACCESS.2023.3299858.

[12] G. Lin, X. Feng, W. Guo, X. Cui, S. Liu, W. Jin, Z. Lin, and Y. Ding, “Electricity Theft Detection Based on Stacked Autoencoder and the Undersampling and Resampling Based Random Forest Algorithm,” *IEEE Access*, vol. 9, Sep. 2021. DOI: 10.1109/ACCESS.2021.3110510.

[13] A. Oad, M. H. Farooq, A. Zafar, B. A. Akram, R. Zhou, and F. Dong, “Fake News Classification Methodology With Enhanced BERT,” *IEEE Access*, vol. 12, Nov. 2024. DOI: 10.1109/ACCESS.2024.3491376.

[14] S. Limanto, J. L. Buliali, and A. Saikhu, “GLoW SMOTE-D: Oversampling Technique to Improve Prediction Model Performance of Students' Failure in Courses,” *IEEE Access*, vol. 12, Jan. 2024. DOI: 10.1109/ACCESS.2024.3351569.

[15] J. Oruh, S. Viriri, and A. Adegun, “Long Short-Term Memory Recurrent Neural Network for Automatic Speech Recognition,” *IEEE Access*, vol. 10, Mar. 2022. DOI: 10.1109/ACCESS.2022.3159339.

[16] G. A. Pradipta, R. Wardoyo, A. Musdholifah, and I. N. H. Sanjaya, “Radius-SMOTE: A New Oversampling Technique of Minority Samples Based on Radius Distance for Learning From Imbalanced Data,” *IEEE Access*, vol. 9, May 2021. DOI: 10.1109/ACCESS.2021.3080316.

[17] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, “SMOTE-MRS: A Novel SMOTE-Multiresolution Sampling Technique for Imbalanced Distribution to Improve Prediction of Anemia,” *IEEE Access*, vol. 12, Oct. 2024. DOI: 10.1109/ACCESS.2024.3482968.

[18] Y. Bao and S. Yang, “Two Novel SMOTE Methods for Solving Imbalanced Classification Problems,” *IEEE Access*, vol. 11, Jan. 2023. DOI: 10.1109/ACCESS.2023.3236794.