International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 577-585 www.ijaem.net ISSN: 2395-5252

Fake Job Post Detection using Machine Learning and Deep Learning

Adarsh Salunkhe, Aditya Taware, Siddhant Golande, Rohan Khandagale, Manoj D.Shelar

Computer Engineering Vpkbiet, Baramati Baramati, India
Computer Engineering Vpkbiet, Baramati Baramati, India
Computer Engineering Vpkbiet, Baramati Baramati, India
Computer Engineering Vpkbiet, Baramati Baramati, India
Computer Engineering Vpkbiet, Baramati Baramati, India

Date of Submission: 05-07-2025 Date of Acceptance: 15-07-2025

ABSTRACT: The rapid growth of online job portals has transformed the recruitment landscape, providing job seekers with easy access to employment opportunities. However, this digital shift has also led to an alarming increase in fraudulent job postings, which can deceive applicants, cause financial loss, and compromise personal data. Detecting such fake job advertisements has become a critical challenge in ensuring safe and trustworthy online hiring platforms. In this research, we propose a robust framework for fake job post detection using both machine learning and deep learning techniques, evaluated on the EMSCAD dataset—a publicly available collection of real and fake job listings sourced from online platforms such as Naukri.com. Our approach begins with thorough data preprocessing, including text cleaning, normalization, and feature extraction. For classical machine learning models such as Naive Bayes and Random Forest, we utilize Term Frequency-Inverse Document Frequency (TF-IDF) to convert textual content into meaningful numerical features. For deep learning, we employ a Recurrent Neural Network (RNN) model, where the input sequences are tokenized, padded, and passed through an embedding layer initialized with pre-trained GloVe vectors for better semantic representation.

We also implement an ensemble model that integrates the outputs of the Naive Bayes, Random Forest, and RNN classifiers using a soft voting mechanism. This ensemble approach is designed to combine the strengths of each individual model—capturing both shallow linguistic patterns and deeper contextual semantics—to enhance overall prediction accuracy.

Our experiments demonstrate that while each model performs well individually, the ensemble model achieves superior performance with an accuracy of 94.3, an F1-score of 92.2, and a ROC-AUC of 0.97. These results highlight the effectiveness of combining machine learning and deep learning approaches for the task of fake job detection. The proposed methodology can serve as a foundation for building automated, scalable, and reliable fake job post detection systems for online recruitment platforms, thereby improving user safety and trust.

I. INTRODUCTION

A. Background and Motivation:

- **Growth of Online Job Portals:** Online job portals have transformed the hiring process by allowing employers and job seekers to connect globally and more efficiently. These platforms host thousands of job listings, providing a wealth of opportunities for job seekers while simplifying the recruitment process for employers. However, with the rapid expansion of these platforms, there has also been a rise in fake job postings, which create risks for users and negatively impact the reputation of these platforms.
- **Prevalence and Impact of Fake Job Posts**: Fake job posts are a form of online fraud, often designed to deceive job seekers into sharing personal data or making monetary payments. These posts may appear legitimate but are crafted with the intent to exploit unsuspecting users, leading to financial loss, privacy breaches, and emotional distress. As fake job posts increase, they erode user trust in job portals, harming the credibility of these platforms and complicating the experience for genuine job seekers.
- **Limitations of Traditional Detection Methods:** Traditional fake job detection

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 577-585 www.ijaem.net ISSN: 2395-5252

approaches rely on manual reviews and rule-based systems that flag suspicious posts based on fixed criteria. Manual review is time-consuming and costly, as it requires a significant investment in human resources to check each post. Rule-based systems can detect some patterns but often lack the flexibility to adapt to evolving tactics used by scammers. These systems can generate false positives, flagging legitimate posts as fraudulent, and false negatives, letting some fake posts slip through undetected.

- **Need for Scalable and Automated Solutions:** As job platforms grow, the volume of data generated by new job postings becomes overwhelming for manual and rule based methods to handle effectively. An automated, scalable, and adaptable solution is essential to keep pace with the vast amount of data and evolving techniques used by fraudsters. Machine learning presents an efficient solution for this challenge, enabling platforms to detect fake job posts automatically with reduced human intervention and increased accuracy.
- **Advantages of Machine Learning for Fake Job Detection:** Machine learning models can analyze large datasets and identify complex, nuanced patterns that would be difficult to detect manually or through rule-based approaches. These models can be trained to recognize linguistic cues, structure patterns, and other indicators specific to fake job listings. NLP techniques allow for the detailed analysis of job description content, reveal ing language patterns or keywords commonly associated with fake job posts. Classification algorithms like Naive Bayes, Support Vector Machines, Random Forests, and deep learning methods like Recurrent Neural Networks (RNNs) and Transformers offer scalable solutions for accurate classification.
- **Exploration of Current ML Techniques:** This review paper provides an in-depth exploration of machine learning techniques used for fake job post detection. Various machine learning models are compared, covering both traditional approaches like Naive Bayes and Support Vec tor Machines and advanced methods like deep learning. Ensemble models and hybrid methods, which combine multiple algorithms, are also discussed for their ability to improve detection accuracy.
- **Challenges in Fake Job Detection:** Detecting fake job posts presents unique challenges, such as the issue of class imbalance, where legitimate job posts vastly out number fraudulent ones, making it difficult for models to recognize minority class patterns. Additionally, fake job posts are dynamic and can change frequently as fraudsters adjust their tactics, requiring models that can adapt over time. This paper examines strategies to address these challenges, including advanced model tuning, real-time learning, and adaptive algorithms.
- **Benefits for Job Seekers and Online Job Portals**: Robust machine learning models for fake job detection can help protect job seekers from falling victim to scams, safeguarding their personal information and finances. Enhanced detection systems also improve the reliability of job portals, maintaining user trust and ensuring a safer job search environment. By mitigating risks associated with fake job posts, job portals can offer a more trust worthy platform, fostering a positive user experience and supporting the recruitment process more effectively.
- **Contribution to the Field:** This review synthesizes the latest advancements in machine learning applications for fake job post detection, aiming to inform future research and development in this area. Through a comparative analysis of different ML models, this paper provides insights into the trade-offs and performance metrics relevant to real-world deployment. Overall, this paper highlights the potential of machine learning to make online job platforms safer and more reliable for users.

B. Traditional Methods and Limitations:

- **Manual Review by Moderators:** Job platforms often employ human moderators to review job posts manually, looking for signs of fraud or inconsistency. Moderators check for red flags such as missing company information, unprofessional language, or requests for sensitive information. This approach, while effective in identifying some fraudulent patterns, is labor-intensive, costly, and challenging to scale as the volume of job posts grows.
- **Rule-Based Systems:** Rule-based systems use predefined criteria to flag suspicious job posts automatically. Common rules may include filtering posts based on keywords (e.g., ”quick money,” ”send fee”), checking for incomplete job details, or detecting suspicious URLs. Although this method is faster than manual review, it is limited by the fixed nature of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 577-585 www.ijaem.net ISSN: 2395-5252

rules. These systems are often unable to adapt to evolving tactics used by fraudsters and can produce high false-positive or false-negative rates.

- **Keyword Matching**: This approach involves searching job descriptions for specific keywords or phrases commonly associated with scams, such as "no experience required" or "work from home for quick money." Although easy to implement, keyword matching can lack accuracy, as legitimate job posts can also contain similar language, and some fake posts may use sophisticated language to avoid detection.
- **Pattern Recognition:** Traditional systems may use simple pattern recognition techniques to identify common structural elements of fake job posts. For example, pat terns such as lack of a company name, unusually high pay, and personal email addresses are often flagged as suspicious. Pattern recognition is useful but limited, as it often cannot identify more subtle or context-dependent fraudulent patterns.
- **Heuristic-Based Approaches**: Heuristic methods rely on empirical knowledge or common indicators of fraud. For instance, a post that lacks detailed job descriptions or provides overly generic descriptions might be flagged. Heuristics can also include red flags like urgent hiring language, requests for payment or personal information, and lack of professional contact details. While effective for certain types of fraud, heuristic-based methods often fail to generalize to new types of fake job posts and are not robust against sophisticated fraud tactics.
- **IP and Geolocation Filters:** Platforms may track the IP addresses and geolocations of users posting job listings, flagging those from regions with high instances of fraudulent activity. IP and geolocation filters can help reduce fake job posts from specific sources, but they may be bypassed by fraudsters using VPNs or proxy servers, limiting their effectiveness.
- **Verification and Validation Processes:** Some job portals use verification processes, such as verifying the employer's email domain or requiring additional documentation, to ensure legitimacy. This approach is useful but can increase friction for genuine employers, as it requires more steps in the job posting process and is often bypassed by highly sophisticated scammers.
- **Reputation-Based Systems:** Platforms sometimes use reputation-based systems, assigning scores or trust levels to employers based on their history, the number of legitimate posts, and user feedback. While helpful in identifying repeat fraudsters, this method struggles to detect fraud from new or untested accounts and may unfairly penalize legitimate employers with limited posting history.
- **Regular Audits and Spot Checks:** Regular audits or random spot checks are conducted to review samples of job posts for signs of fraud. This approach helps identify emerging scam patterns but is time-consuming, requires significant resources, and only provides partial coverage of job postings.
- **User Reporting Mechanisms:** Many job portals rely on users to report suspicious job postings. A reporting system allows job seekers to flag potential scams, which can then be reviewed manually. This is a reactive approach, dependent on users recognizing fraud and taking action, which often happens only after a scam has affected users.

II. SCOPE AND OBJECTIVES

A. Scope of the Study

With the increasing digitization of recruitment processes, online job portals have become the primary medium for connecting job seekers with potential employers. However, this convenience has also led to the proliferation of fake job postings, where fraudsters exploit job seekers by offering misleading or non-existent employment opportunities. These scams can result in significant financial loss, emotional stress, and the misuse of personal information. Therefore, there is a growing need for intelligent systems that can automatically identify and filter out such fraudulent job advertisements.

This research aims to address this problem by developing an automated fake job post detection system using machine learning and deep learning techniques. The study is based on the EMSCAD dataset, which contains labeled job postings collected from real online recruitment platforms like Naukri.com. The dataset includes both legitimate and fake job posts, offering a reliable benchmark for training and evaluating classification models.

The scope of this research includes:

- **Textual Analysis:** Analyzing the linguistic and semantic patterns in job postings to differentiate between real and fake listings.

- **Data Preprocessing:** Performing comprehensive data cleaning, normalization, and feature engineering on raw textual data.
- **Feature Extraction:** Using methods such as Term Frequency-Inverse Document Frequency (TF-IDF) for machine learning models and word embeddings (e.g., GloVe) for deep learning models to transform text into numerical representations.
- **Model Development:** Implementing and training multiple classification models—Naive Bayes, Random Forest, and Recurrent Neural Networks (RNN)—to assess their ability to detect fake job posts.
- **Ensemble Learning:** Building a hybrid model that combines the predictions of individual classifiers using a soft voting technique to improve overall prediction accuracy.
- **Evaluation:** Testing the performance of each model using metrics like accuracy, precision, recall, F1-score, and ROC-AUC to determine their effectiveness.
- **Deployment Readiness:** Exploring the potential for integrating the proposed system into real-world job portals to prevent the spread of fraudulent job advertisements.

This study is limited to analyzing text-based English language job posts and does not consider multimedia job advertisements (e.g., images, videos, PDFs) or multilingual data. The findings are, however, generalizable to similar datasets and can serve as a foundation for future enhancements.

B. Objectives of the Study

The primary goal of this research is to build an intelligent and reliable system that can automatically identify and classify fake job postings from real ones using advanced computational techniques. The specific objectives are outlined below:

- To investigate the problem of fake job postings and understand how fraudulent advertisements differ in content and structure from genuine ones.
- To perform exploratory data analysis (EDA) on the EM SCAD dataset to identify patterns, trends, and anomalies that may indicate fraudulent behavior in job postings.
- To preprocess the textual data by removing noise, normalizing text, handling missing values, and converting unstructured text into structured formats suitable for analysis.
- To apply feature extraction techniques such as TF-IDF for classical machine learning models and tokenization with word embeddings for deep learning models to convert raw text into meaningful numerical vectors.
- To design and train individual models:
  - **Naive Bayes Classifier** for fast and interpretable classification using probabilistic reasoning.
  - **Random Forest Classifier** to handle complex feature interactions and provide robust predictions.
  - **Recurrent Neural Network (RNN)** to capture sequential and contextual relationships in job description texts.
- To develop an ensemble model that integrates the strengths of the above classifiers using soft voting, aiming to enhance the reliability and performance of the fake job post detection system.
- To evaluate the models thoroughly using appropriate performance metrics (accuracy, precision, recall, F1-score, and ROC-AUC) to determine which model or combination of models provides the most effective solution.
- To provide practical insights and recommendations for deploying the proposed solution in real-world online job portals, with the ultimate goal of protecting users from employment fraud.
- To identify limitations and propose directions for future work, including the integration of more diverse features, multilingual support, and explainability tools.

III. LITERATURE REVIEW WITH BENEFITS AND LIMITATIONS

A study by **Dutta and Bandyopadhyay** provides a comprehensive approach by applying different classifiers to detect fake job listings, highlighting the benefits of ensemble classifiers over single classifiers for better accuracy in fake job detection. They categorize their classifiers into two main types: single classifiers and ensemble-based approaches.

In their approach, several single classifiers were explored, including the Naive Bayes Classifier, Multi-Layer Perceptron, K-Nearest Neighbor, and Decision Tree classifiers. Each of these models has distinct strengths and limitations. For instance, Naive Bayes is effective with independent features, while K-Nearest Neighbor excels in handling spatial data but is sensitive to the choice of k. Decision Trees, commonly used for spam detection and other classification problems, also performed well, achieving a high level of accuracy. However, ensemble classifiers—specifically Random Forest, AdaBoost, and

Gradient Boosting—showed superior performance across various evaluation metrics like accuracy, F1-score, Cohen Kappa score, and Mean Squared Error (MSE). Random Forest outperformed the other models, achieving an accuracy of 98.27, due to its ability to reduce overfitting by combining multiple decision trees and averaging their results.

The study by Dutta and Bandyopadhyay also draws attention to the broader landscape of online fraud detection, comparing job fraud detection to related tasks such as email spam filtering, fake news detection, and review spam detection. Techniques in these areas commonly involve feature extraction through Natural Language Processing (NLP), enabling classifiers to identify suspicious language patterns and content specific characteristics. The success of these techniques in related fields informs and validates their application to fake job detection, as the fundamental challenges—such as the variability and subtlety of fraudulent content—are shared across domains.

Sangeeta Lal, Rishabh Jaiswal, Neethu Sardana, Ayushi Verma, Amanpreeth Kaur, and Rahul Mourya developed an ensemble-based model called ORF Detector for detecting online recruitment fraud (ORF). The model combines baseline classifiers such as J48, Logistic Regression (LR), and Random Forest (RF) using ensemble methods, achieving an impressive average F1-score and accuracy of 94 and 95.4, respectively. However, the model faces challenges with interpretability and computational complexity.

Additionally**, research by Elsevier B.V.** explored various machine learning techniques for financial fraud detection, including Classification and Regression Trees (CART), Na¨ıve Bayes, and K-Nearest Neighbor (KNN) methods. Their study highlights the effectiveness of hybrid techniques, which blend traditional methods to enhance fraud detection capabilities. These models are capable of processing large volumes of transactions with high speed and accuracy, making them highly effective in practical applications. However, the scale of data involved requires significant investments in data storage and management resources.

Alghamdi and Alharby, leveraging the openly accessible EMSCAD dataset, have achieved an impressive culmination with a 97.41 success rate. The focal points of their scrutiny encompass not only corporate logos but also other pivotal attributes.

Amaar et al. implemented six advanced machine learning models to evaluate the authenticity of job advertisements. To compare the models, they used two feature extraction techniques: Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), providing a comprehensive analysis of each classifier's performance. A key challenge in their study was the imbalance in the dataset, where genuine job postings significantly outnumbered fraudulent ones, leading to overfitting on the majority class. To address this, they employed the Adaptive Synthetic Sampling (ADASYN) technique, which generates synthetic samples for the minority class to balance the dataset. They conducted two experiments—one on a balanced dataset and another on the original imbalanced dataset. Their analysis revealed that the Extremely Ran domized Trees (ETC) model, using TF-IDF and ADASYN, achieved a high accuracy of 99.9%. Additionally, the study compared their approach with recent deep learning models and other re-sampling methods, offering a thorough evaluation of the proposed methodology.

Jihadists explore the layered structure of perceptrons in neural networks, describing how interconnected layers reduce error rates by adjusting the weights in input layers. This organized structure has the potential to significantly improve the effectiveness of neural network models. In a separate study, FHA Shibly, Uzzal Sharma, and HMM [8] investigated data classification using two specific algorithms: the two class boosted decision tree and the two-class decision forest. Their findings indicate that the two-class boosted decision tree outperforms the decision forest in classification accuracy. However, they note a significant drawback in the extended training time required and the numerous hyperparameters involved, which can make the model prone to overfitting.

The work of Gupta et al. explores the integration of social network analysis (SNA) in fake job post detection. Recognizing the interconnected nature of users on job platforms, the study utilizes SNA to identify patterns and anomalies in user behavior. By examining the relationships and connections among users, the research enhances the discriminatory power of the model, contributing to a more nuanced understanding of the social dynamics associated with fake job postings. This approach extends the traditional focus on textual features to include the social context in which job posts are disseminated

Additionally, **the study by Chen et al.** introduces the application of deep learning techniques, specifically convolutional neural networks (CNNs), for fake job post detection. By extracting hierarchical features from job

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 577-585 www.ijaem.net ISSN: 2395-5252

descriptions, the CNN-based model demonstrates a high level of accuracy in distinguishing between genuine and deceptive job postings.

IV. PROPOSED WORK

A. Dataset Description

For this study, we utilized the EMSCAD (Employment Scam Aware Dataset), which contains job postings sourced from different online employment platforms. The dataset comprises both legitimate and fraudulent job advertisements. Each record includes various attributes such as job title, company profile, job description, requirements, benefits, industry, function, and a binary label indicating whether the job post is fraudulent (1) or genuine (0).

The dataset provides a rich source of textual and categorical information suitable for machine learning and deep learning based classification.

B. Data Preprocessing

To prepare the dataset for modeling, the following preprocessing steps were applied:

1) **Null Value Removal:** All rows containing missing values in critical text fields (description, requirements, etc.) were removed to ensure data quality and consistency.
2) **Text Consolidation:** Multiple fields, including title, company profile, description, requirements, and benefits, were concatenated into a single text field. This provided a comprehensive representation of each job posting.
3) **Text Cleaning:** The combined text was cleaned using the following techniques:
- Conversion to lowercase
- Removal of punctuation, numbers, and special characters
- Removal of stop words (e.g., "the", "is", "at")
- Lemmatization to normalize words to their base forms
4) **Label Encoding:** The target variable fraudulent was encoded into numerical format, where 0 represented real job postings and 1 represented fake job postings.
5) **Class Imbalance Handling:** The dataset exhibited class imbalance, with significantly fewer fake job postings. To address this, oversampling was performed using the Synthetic Minority Oversampling Technique (SMOTE) to balance the distribution of classes in the training data.

C. Feature Extraction

For Machine Learning Models: To prepare the textual data for traditional machine learning algorithms, we employed the Term Frequency–Inverse Document Frequency (TF-IDF) vectorization technique. TF-IDF transforms raw text into numerical feature vectors by evaluating the importance of a term in a document relative to the entire corpus. This approach emphasizes informative terms while down-weighting frequently occurring but less meaningful words. The resulting sparse matrix representation was used as input for the Naive Bayes and Random Forest classifiers.

For Deep Learning Model (RNN): For the RNN-based model, a separate preprocessing pipeline was utilized, tailored to the needs of sequential neural networks:

- **Tokenization:** The Keras Tokenizer was used to convert the cleaned textual data into sequences of integer tokens. Each integer corresponds to a specific word in the constructed vocabulary.
- **Padding:** To ensure consistent input dimensions for the RNN, all sequences were padded to a fixed length. Padding was applied post-sequence using the Keras pad_sequences function.

D. Model Architecture

To evaluate the effectiveness of various approaches in detecting fake job postings, we implemented multiple machine learning and deep learning models, as well as an ensemble strategy. The details of each model are summarized below:

- **Naive Bayes Classifier:**
  - Utilized the Multinomial Naive Bayes algorithm, which is well-suited for discrete features such as term frequencies or TF-IDF.
  - Trained on TF-IDF feature vectors derived from the cleaned text data.
- **Random Forest Classifier:**
  - An ensemble-based model that constructs multiple decision trees using bootstrap aggregation (bagging). – Capable of capturing non-linear patterns and inter actions between features.
  - Input features were derived using TF-IDF vectorization.
- **Recurrent Neural Network (RNN):**
  - Designed to capture sequential dependencies within textual data.
  - The model architecture included:

1) Input Layer (Tokenized and Padded Sequences)
2) Dense Layer(s)

International Journal of Advances in Engineering and Management (IJAEM)
Volume 7, Issue 07 July 2025, pp: 577-585 www.ijaem.net ISSN: 2395-5252

3) Sigmoid Output Layer (for binary classification)
– The model was trained using binary cross-entropy loss and the Adam optimizer.
– Dropout regularization was included to prevent over fitting.
• **Ensemble Model:**
– Combined predictions from the Naive Bayes, Random Forest, and RNN models.
– A soft voting strategy was employed to aggregate the class probabilities from individual models.
– The ensemble aimed to leverage the complementary strengths of each model:
• Naive Bayes for high bias/low variance
• Random Forest for robustness and interpretability
• RNN for deep contextual understanding of text
– This multi-model approach was designed to improve classification performance by combining shallow and deep learning techniques, thus offering a more comprehensive detection mechanism.

E. Model Training and Evaluation

To ensure fair and consistent comparison across all models, a standardized training and evaluation framework was followed. The dataset was divided into training (80%) and testing (20%) subsets, with the training set used for model development and the testing set reserved for performance evaluation.

Training Procedure:

a) Naive Bayes & Random Forest::
- Both models were trained using TF-IDF feature vectors.
- Hyperparameters, such as the number of estimators for Random Forest, were optimized using grid search with cross-validation.
- Stratified 5-fold cross-validation was employed during training to ensure class balance and minimize overfitting.

b) Recurrent Neural Network (RNN)::
- Text sequences were tokenized and padded to a uniform length.
- The embedding layer was either randomly initialized or pre-loaded with GloVe embeddings.
- The RNN model was trained using the binary cross entropy loss function and optimized with the Adam optimizer.
- Early stopping and dropout regularization were applied to prevent overfitting.
- Batch size and number of epochs were tuned through experimentation.

c) Ensemble Model::
- Combined the probabilistic outputs of the Naive Bayes, Random Forest, and RNN classifiers using a soft voting strategy.
- The ensemble model was constructed after individual model training and evaluated on the same testing set.

Evaluation Metrics: To comprehensively assess model performance, the following metrics were used:
- **Accuracy:** The overall correctness of the model.
- **Precision:** The proportion of true fake job posts among those predicted as fake.
- **Recall (Sensitivity):** The ability of the model to correctly identify fake job posts.
- **F1-Score:** The harmonic mean of precision and recall, balancing both concerns.
- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Measures the model's ability to distinguish between classes.

These metrics provide insights into both the effectiveness and robustness of each model, especially in the presence of class imbalance, which is common in fake job detection datasets.

Table I: MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Naive Bayes	87.5%	82.3%	80.4%	81.3%	0.89
Random Forest	92.1%	90.2%	88.6%	89.4%	0.94
RNN	93.5%	91.0%	91.2%	91.1%	0.96
Ensemble (Voting)	**94.3%**	**92.5%**	**92.0%**	**92.2%**	**0.97**

V. FUTURE WORK

While this study demonstrates promising results in detecting fake job postings using a combination of machine learning and deep learning models, several avenues remain for further improvement and exploration:
- **Incorporation of More Diverse Features:** Future work can explore integrating additional

data sources such as company reviews, user feedback, or metadata like posting timestamps and recruiter profiles to enrich the feature set.

- **Advanced Deep Learning Architectures:** Experimenting with more sophisticated models such as Transformer based architectures (e.g., BERT, RoBERTa) could improve semantic understanding and classification performance.
- **Explainability and Interpretability:** Implementing model explainability techniques (e.g., SHAP, LIME) would help in understanding the key factors influencing model predictions, increasing trust and usability for real world applications.
- **Real-time Detection Systems**: Developing efficient pipelines for real-time fake job post detection on live job portals would enhance practical applicability.
- **Multilingual and Cross-platform Analysis**: Extending the model to handle job posts in multiple languages and from various international job platforms could broaden the system's utility and robustness.
- **Addressing Adversarial Attacks:** Investigating defenses against adversarial manipulation attempts by fraudulent posters can improve the model's resilience.
- **User Feedback Integration:** Incorporating human-in the-loop mechanisms for continuous model refinement based on user reports and feedback.

These future directions will contribute to building more accurate, robust, and deployable systems for combating fraudulent job postings.

VI. CONCLUSION

The growing reliance on online job portals has significantly improved access to employment opportunities, but it has also created new vulnerabilities—particularly in the form of fake job postings that exploit unsuspecting job seekers. In this study, we proposed and evaluated a comprehensive frame work for the detection of fake job advertisements using a combination of classical machine learning and deep learning approaches. The research utilized the EMSCAD dataset, which contains labeled job postings collected from real-world plat forms such as Naukri.com.

We implemented and compared three models—Naive Bayes, Random Forest, and Recurrent Neural Network (RNN)—each leveraging different aspects of textual data for classification. Feature extraction techniques such as TF-IDF and word embeddings were employed to convert unstructured job descriptions into structured input for the models. Furthermore, we developed an ensemble model that combines the strengths of all three approaches using a soft voting mechanism, achieving superior performance compared to individual classifiers.

Experimental results demonstrate that the ensemble model achieved an accuracy of 94.3% and an ROC-AUC score of 0.97, confirming its effectiveness in identifying fraudulent job postings. These findings indicate that hybrid approaches, which combine both statistical and deep learning models, can significantly improve the reliability and robustness of fake job post detection systems.

This work contributes not only to the field of text classification but also provides a practical solution for improving the safety and trustworthiness of online recruitment platforms. The proposed system can be integrated into real-world job portals to proactively identify and block fraudulent listings, ultimately protecting job seekers from potential harm.

REFERENCES

[1]. V Anbarasu1, Dr. S. Selvakani, Mrs. K. Vasumathi "Fake Job Prediction Using Machine Learning" INTER NATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH AND EMERGING TECH NOLOGIES Vol. 13, No. 1, 2024.

[2]. S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, "Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset," Fu ture Internet, vol. 9, no. 1, p. 6, Mar. 2017, doi: 10.3390/fi9010006.

[3]. B. Alghamdi and F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection," Journal of Information Security, vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[4]. Scanlon, J.R. and Gerber, M.S. (2014) Automatic De tection of Cyber-Recruitment by Violent Extremists. Se curity Informatics, 3, 5. https://doi.org/10.1186/s13388- 014-0005-5

[5]. R. S. Shishupal, Varsha, S. Mane, V. Singh, and D. Wasekar, "Efficient Implementation using Multinomial Naive Bayes for Prediction of Fake Job Profile," Interna tional Journal of Advanced Research in Science, Com munication and Technology, pp. 286–291, May 2021, doi: 10.48175/IJARSCT-1241.

[6]. O. Nindyati and I. G. Bagus Baskara Nugraha, "De tecting Scam in Online Job Vacancy Using Behav ioral Features Extraction," in 2019 International Con ference on ICT for Smart Society (ICISS), Ban dung, Indonesia: IEEE, Nov. 2019, pp. 1–4. doi: 10.1109/ICISS48059.2019.8969842.

[7]. P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embed ding clustering and convolutional neural network for improving short text classification," Neurocom puting, vol. 174, pp. 806–814, Jan. 2016, doi: 10.1016/j.neucom.2015.09.096.

[8]. S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruit ment Detection Using Machine Learning Approach," International Journal of Engineering Trends and Tech nology, vol. 68, no. 4, pp. 48–53, Apr. 2020,

[9]. I. M. Nasser and A. H. Alzaanin, "Machine Learning and Job Posting Classification: A Comparative Study," International Journal of Engineering and Information Systems (IJEAIS), vol. 4, no. 9, pp. 06–14, 2020.

[10]. F. Shibly, U. Sharma, and H. Naleer, "Performance Comparison of Two Class Boosted Decision Tree and Two Class Decision Forest Algorithms in Predicting Fake Job Postings," Annals of the Romanian Society for Cell Biology, pp. 2462–2472, Apr. 2021.