WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

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

salunkheadarsh77@gmail.com, adityataware237@gmail.com, sidd7274@gmail.com, rohankh37@gmail.com, manoj.shelar@vpkbiet.org

Abstract—

The rapid growth of online job portals has trans- formed 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 out- puts of the Naive Bayes, Random Forest, and RNN classifiers us- ing 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.cha

I. Introduction

A. Background and Motivation:

- **Growth of Online Job Portals:** Online job portals have revolutionized the hiring landscape by enabling seamless and global connections between employers and job seekers. These platforms feature a vast number of job postings, offering abundant opportunities for candidates and making it easier for employers to find suitable talent. However, the rapid growth of these platforms has also led to an increase in fraudulent job advertisements, posing threats to users and damaging the credibility of the portals.
- **Prevalence and Impact of Fake Job Posts:** Fake job advertisements represent a type of online scam aimed at tricking job seekers into disclosing sensitive personal information or making financial payments. While these listings often look authentic, they are deliberately created to exploit individuals, resulting in potential financial dam- age, loss of privacy, and emotional harm. The growing presence of such deceptive posts undermines the trust users place in job portals, damaging their reputation and making it more difficult for legitimate job seekers to navigate the employment landscape.
- **Limitations of Traditional Detection Methods:** Conventional methods for detecting fake job postings typically depend on manual inspection and rule-based mechanisms that identify suspicious content using predefined criteria. Manual review, while effective to some extent, is labor-intensive and expensive, requiring substantial human effort to evaluate each listing individually. On the other hand, rule-based systems can catch certain known patterns but struggle to keep up with the constantly changing strategies employed by scammers. These systems often suffer from inaccuracies, leading to false positives—where genuine job postings are mistakenly flagged—and false negatives, where fraudulent posts go unnoticed.
- **Need for Scalable and Automated Solutions:** With the rapid expansion of job platforms, the sheer volume of new job postings has become too large for manual and rule-based methods to manage efficiently. To effectively handle this growing data and counter the constantly evolving tactics of scammers, a solution that is automated, scalable, and adaptable is crucial. Machine learning offers a powerful approach to this problem, allowing platforms to automatically identify fake job postings with greater accuracy and minimal human effort, improving both detection speed and

reliability.

- **Advantages of Machine Learning for Fake Job Detection** Machine learning models are capable of processing vast amounts of data and uncovering subtle, intricate patterns that are often missed by manual reviews or fixed rule-based systems. These models can be trained to detect specific linguistic features, structural formats, and other indicators commonly found in fake job advertisements. Natural Language Processing (NLP) techniques enhance this capability by enabling a deeper examination of job descriptions, identifying suspicious language usage and keywords typically linked to fraudulent postings. Classification algorithms such as Naive Bayes, Support Vector Machines, and Random Forests, along with advanced deep learning models like Recurrent Neural Networks (RNNs) and Transformers, provide scalable and highly accurate solutions for detecting fake job posts.
- **Exploration of Current ML Techniques** This paper offers a comprehensive analysis of machine learning techniques applied to the detection of fake job postings. It compares a range of models, from traditional approaches such as Naive Bayes and Support Vector Machines to more advanced deep learning methods. The study also examines ensemble and hybrid models, which integrate multiple algorithms to enhance detection performance and improve overall accuracy in identifying fraudulent job listings.
- **Challenges in Fake Job Detection** Detecting fake job postings poses several distinct challenges. One major issue is class imbalance, where the number of genuine job posts far exceeds that of fake ones, making it harder for models to accurately identify patterns associated with the minority (fraudulent) class. Moreover, fake job listings are constantly evolving, as scammers frequently modify their tactics to bypass detection systems. This demands models that are not only accurate but also adaptable over time. This paper explores various strategies to tackle these challenges, such as advanced model optimization techniques, real-time learning frameworks, and adaptive algorithms capable of adjusting to new patterns in fraudulent behavior.
- **Benefits for Job Seekers and Online Job Portals** Effective machine learning models for detecting fake job postings play a crucial role in protecting job seekers from scams, helping to secure their personal data and financial well-being. Improved detection mechanisms also strengthen the credibility of job portals by reducing the presence of fraudulent listings, thereby preserving user trust and promoting a safer job search experience. By minimizing the risks posed by fake job posts, these platforms can deliver a more reliable and user-friendly environment, ultimately supporting both job seekers and recruiters more efficiently.
- **Contribution to the Field** This review consolidates recent developments in the use of machine learning for detecting fake job postings, with the goal of guiding future research and innovation in the field. By conducting a comparative analysis of various machine learning models, the paper offers valuable insights into their respective strengths, limitations, and performance metrics in practical applications. Ultimately, the study underscores the promise of machine learning in enhancing the security and trustworthiness of online job platforms, making them safer and more dependable for users.

B. Traditional Methods and Limitations:

- **Manual Review by Moderators** Job platforms frequently rely on human moderators to manually review job postings for potential signs of fraud or inconsistency. These moderators look for common red flags, such as incomplete company details, unprofessional or vague language, and requests for sensitive personal information. While this method can be effective in spotting certain types of fraud, it is resource-intensive, expensive, and difficult to scale efficiently as the number of job postings continues to increase.
- **Rule-Based Systems:** Rule-based systems automatically flag suspicious job postings using a set of predefined criteria. Typical rules might involve filtering listings containing specific keywords like “quick money” or “send fee,” identifying incomplete job details, or spotting dubious URLs. While this approach is faster than manual review, its effectiveness is limited by the rigidity of the rules. Such systems often struggle to keep up with the changing strategies of fraudsters and can result in a high number of false positives (legitimate posts flagged incorrectly) and false negatives (fraudulent posts that go undetected).
- **Keyword Matching:** This method focuses on scanning job descriptions for certain keywords or phrases frequently linked to scams, like “no experience required” or “work from home for quick money.” While keyword matching is straightforward to implement, it often lacks precision because genuine job listings may include similar terms, and some fraudulent posts use more subtle or sophisticated language to evade detection.
- **Pattern Recognition:** Traditional systems often rely on basic pattern recognition to detect common structural features of fake job postings. For instance, they may flag listings that lack a company name, promise unusually high salaries, or use personal email addresses as suspicious. While pattern recognition can be helpful in catching obvious signs of fraud, it has limitations, as it often fails to detect more subtle or context-dependent fraudulent behaviors.
- **Heuristic-Based Approaches:** Heuristic methods use practical experience and known fraud indicators to identify suspicious job postings. For example, a listing with vague or overly generic job descriptions may be flagged. Other common heuristics include spotting urgent hiring language, demands for payment or personal information, and absence of professional contact details. Although these methods can be effective against some types of fraud, they often struggle to adapt to new or evolving fake job schemes and are less reliable when faced with more sophisticated scams.
- **IP and Geolocation Filters:** Platforms sometimes monitor the IP addresses and geolocation of users submitting job listings, flagging those originating from areas known for high levels of fraudulent activity. While IP and geolocation filtering can help reduce fake job posts from certain sources, savvy fraudsters can circumvent these measures by using VPNs or proxy servers, which diminishes their overall effectiveness.
- **Verification and Validation Processes:** Some job portals implement verification procedures—like confirming the employer’s email domain or requesting extra documentation—to validate the authenticity of job postings. While this approach helps improve legitimacy, it can also create additional hurdles for genuine employers by adding extra steps to the posting process. Moreover, highly sophisticated scammers may still find ways to bypass these checks.
- **Reputation-Based Systems:** Platforms often implement reputation-based systems that assign trust scores to employers based on factors like their posting history, number of verified job listings, and user feedback. This approach can be effective in identifying repeat offenders, but it has

limitations when it comes to detecting fraud from new or less-established accounts. Additionally, legitimate employers with little or no prior activity may be unfairly disadvantaged by such systems.

- **Regular Audits and Spot Checks:** Regular audits or random spot checks involve reviewing selected samples of job postings to detect potential fraud. This method can help uncover new and evolving scam tactics, but it is labor-intensive, demands considerable resources, and covers only a fraction of all job listings, limiting its overall effectiveness.
- **User Reporting Mechanisms:** Many job portals depend on users to report suspicious job listings through dedicated reporting systems. Job seekers can flag potential scams, which are then subject to manual review. However, this approach is reactive, relying on users to identify fraudulent posts and take action—often only after they have already encountered or been affected by a scam.

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 com- bines 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** presents a thorough approach to detecting fake job listings by applying various classifiers and emphasizes the advantages of ensemble methods over individual classifiers for improved accuracy. They divide their classifiers into two categories: single classifiers and ensemble-based approaches.

In their work, several single classifiers were tested, including Naive Bayes, Multi-Layer Perceptron (MLP), K-Nearest Neighbor (KNN), and Decision Tree classifiers. Each model has unique strengths and weaknesses—for example, Naïve

Bayes performs well when features are independent, KNN is effective for spatial data but sensitive to the parameter k, and Decision Trees are widely used for classification tasks such as spam detection and demonstrated strong accuracy.

However, ensemble classifiers like Random Forest, Ad- aBoost, and Gradient Boosting consistently outperformed the single classifiers across multiple metrics, including accuracy, F1-score, Cohen's Kappa score, and Mean Squared Error (MSE). Among these, Random Forest achieved the highest accuracy of 98.27, benefiting from its ability to reduce overfitting by aggregating the predictions of multiple decision trees.

The study by **Dutta and Bandyopadhyay** also highlights the wider context of online fraud detection, drawing parallels between fake job detection and related domains such as email spam filtering, fake news identification, and review spam detection. These fields commonly utilize feature extraction techniques based on Natural Language Processing (NLP), which enable classifiers to recognize suspicious language pat- terns and domain-specific characteristics. The success of these approaches in related areas supports their effectiveness in fake job post detection, as these tasks share fundamental challenges like the variability and subtlety of fraudulent content.

In a related work, **Sangeeta Lal, Rishabh Jaiswal, Neethu Sardana, Ayushi Verma, Amanpreeth Kaur, and Rahul Mourya** proposed an ensemble-based model named ORF Detector designed to detect online recruitment fraud (ORF). This model integrates several baseline classifiers, including J48, Logistic Regression (LR), and Random Forest (RF), through ensemble techniques. The ORF Detector achieved strong performance, with an average F1-score of 94 and accuracy of 95.4. Despite these promising results, the model faces limitations related to interpretability and computational complexity, which pose challenges for practical deployment.

Additionally, research by **Elsevier B.V.** investigated a variety of machine learning techniques for financial fraud detection, including Classification and Regression Trees (CART), Na¨ıve Bayes, and K-Nearest Neighbor (KNN). Their study emphasizes the effectiveness of hybrid approaches, which combine traditional methods to improve fraud detection performance. These models are well-suited for handling large transaction volumes with high speed and accuracy, making them valuable in real-world scenarios. However, the vast scale of data involved demands substantial investments in data storage and management infrastructure.

In another notable study, **Alghamdi and Alharby**, using the publicly available EMSCAD dataset, achieved a remarkable accuracy rate of 97.41 in fake job post detection. Their analysis focused not only on features like corporate logos but also incorporated other key attributes, contributing to their model's strong performance.

Amaar et al. implemented six advanced machine learning models to assess the authenticity of job advertisements. They used two feature extraction methods—Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)—to comprehensively evaluate each classifier's performance. A major challenge in their study was the dataset imbalance, with genuine job posts vastly outnumbering fraud-ulent ones, which risked causing models to overfit on the majority class. To mitigate this, they applied 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 (using ADASYN) and another on the original imbalanced dataset. Their results showed that the Extremely Randomized Trees (ETC) model combined with TF-IDF and ADASYN achieved an impressive accuracy of 99.9. The study also compared their approach with recent deep learning models and alternative resampling techniques, providing a thorough evaluation of their methodology.

Jihadists explored the layered architecture of perceptrons in neural networks, explaining how interconnected layers reduce error rates by iteratively adjusting weights in the input layers. This structured design has strong potential to significantly enhance the performance of neural network models.
Separately, **FHA Shibly, Uzzal Sharma, and HMM** examined data classification using two algorithms: the two-class boosted decision tree and the two-class decision forest. Their findings indicate that the boosted decision tree outperforms the decision forest in classification accuracy. However, they also noted drawbacks such as longer training times and a large number of hyperparameters, which increase the risk of overfitting and complicate model tuning.

The work of **Gupta et al.** investigates the use of social network analysis (SNA) for detecting fake job posts. Acknowledging the interconnected nature of users on job platforms, their study leverages SNA to uncover patterns and anomalies in user behavior. By analyzing relationships and connections

among users, this approach enhances the model's ability to differentiate between legitimate and fraudulent activity, offering a deeper insight into the social dynamics surrounding fake job postings. This method broadens the traditional focus beyond just textual features by incorporating the social context in which job posts circulate.

Additionally, the study by **Chen et al.** explores the use of deep learning techniques, particularly convolutional neural networks (CNNs), for fake job post detection. By capturing hierarchical features from job descriptions, the CNN-based model achieves high accuracy in distinguishing genuine job listings from deceptive ones, showcasing the potential of deep learning in this domain.

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 *bene- fits*, 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 char- acters
- Removal of stopwords (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 con- vert 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)
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 com- prehensive 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 im- prove 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. Further- more, 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," *Future 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 Detection of Cyber-Recruitment by Violent Extremists. *Security 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," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 286–291, May 2021, doi: 10.48175/IJARSCT-1241.
6) O. Nindyati and I. G. Bagus Baskara Nugraha, "Detecting Scam in Online Job Vacancy Using Behavioral Features Extraction," in 2019 *International Conference 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," *Neurocomputing*, vol. 174, pp. 806–814, Jan. 2016, doi: 10.1016/j.neucom.2015.09.096.
8) S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach," *International Journal of Engineering Trends and Technology*, 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.