International Journal of Computational Science and Engineering Research
ISSN: XXXX- XXXX(Online) , http://www.ijcser.com/
Regular Issue , Vol. 2, Issue. 3 (July - September), 2025 , Pages: 9 - 14
Received: 21 May 2025 ; Accepted: 29 July 2025 ; Published: 18 August 2025
Original Paper : Applied Research - Experimental Research Paper
https://doi.org/10.63328/IJCSER-V2I3P2

Fake Job Recruitment Detection Using Machine Learning

Vijay Bhaskar Reddy [1*] , **Bhuvana Reddy** [2] , **Naga Shanthi** [3], **Isharth Aliya MS** [4], **Deepthi A**[5]

[1*,2,3,4,5] Department of CSM , Srinivasa Ramanujan Institute of Technology , Anantapur, AP, India ;

* *Corresponding Author:* bvijay.br@gmail.com

Abstract: The increasing number of fake social media posts together with fraudulent content has become a major factor in online fraud growth which causes people to doubt trustworthiness and security. The authentication of post authenticity has evolved into a vital operation because user-generated content increases dynamically every day. The research investigates the effectiveness of XGBoost and Random Forest as well as Logistic Regression for classifying posts into real or fake categories. A total of 18,000 different online scam-related posts comprised the Employment Scam Aegean Dataset (EMSCAD). These algorithms show high success rates in detecting genuine content because they use their gained knowledge from previous data analysis. The study delivers important findings to automate scam detection systems which lead to better security measures and lower online fraudulent risks on different platforms.

Keywords: Data Mining, Classification, XGBoost, Catboost, Light Gradient Boosting, Random Forest.

1. Introduction

The Technological progress together with social media's expansion resulted in enormous growth of online content. The fast growth of social media created new positive and negative impacts that include fraudulent content creation and deceptive posts. Scams that spread misinformation have become a serious problem which produces destructive effects which affect personal and corporate populations as well as social systems in general. The rise of online interactions has established genuine post detection as an essential challenge because people must identify real from faux messages. The research community acknowledges fraud detection prediction as an essential subject of study today. Traditional methods used to detect fake content form the majority of existing literature regarding this subject yet these methods fail to adapt to online interaction changes and fraudsters' modern deceptive practices. An insufficient detection capability of fake posts demands improved solutions to fight this online scam problem.[1]

This research work addresses the identified gap by implementing XGBoost, Catboost, Light Gradient Boosting and Random Forest, Logistic Regression alongside Decision Trees to achieve accurate fake and genuine post classification. The study analyzes the detection capabilities of different algorithms regarding fraudulent posts through testing with the Employment Scam Aegean Dataset (EMSCAD) containing 18,000 samples. The research aims to provide essential knowledge for enhancing more reliable online scam detection methods that will improve both online trust and security.

Objective of the Research Work : The main purpose of this research assumes the evaluation of machine learning algorithms XGBoost Random Forest and Logistic Regression in their ability to determine real from counterfeit posts that appear on social media and other digital platforms.[2] The research utilizes EMSCAD which contains 18,000 scam-related posts to develop improved automated fraud detection systems that enhance platform security and trust.

Problem Statement: The massive rise in user-content creation on social networks and online platforms has made it essential for users to properly differentiate authentic content from deceptive posts. Online fraud and platform security diminish as fake posts including scams continue to increase. The identification of fraudulent content using conventional approaches involves errors and lacks adequate precision so developers need to create better automatic recognition systems. Scientific studies are established to develop an automated platform which uses machine learning methods for precise fake post detection and protection against online fraud.[3]

C. Scope of the project: Research examine machine learning applications for detecting fraudulent online posts through employment scam detection on digital platforms. The research analyzes XGBoost together with Random Forest and Logistic Regression for their accuracy performance and F1 score evaluation and other relevant metrics. The Employment Scam Aegean Dataset (EMSCAD) containing 18,000 samples serves as the research basis but the methods presented may become applicable to other fraudulent activity identification domains online. The study works to automate fraud detection while enhancing protective measures for online platforms on social media networks.

2. Literature Survey

The present research employs machine learning classifiers with special attention to ensemble methods designed to identify fake job posts on various online platforms. Various classifiers, including Random Forest, Gradient Boosting, and Support Vector Machines, have been implemented, of which ensemble algorithms performed much better at differentiating between valid and fraudulent postings than single classifiers. The study also shows the relevance of merging different data sources, namely the text content, job title & other meta information, results in more robust predictions. The results show the potential of machine learning based models in automating the detection of fraudulent job ads and protecting job seekers from scams. Ensemble classifiers have significantly improved projection accuracy when compared to single classifiers.[4]

This paper implement both traditional machine learning and deep learning algorithms, such as neural networks, to analyze for and detect fake job postings. The study highlights the importance of feature engineering and data preprocessing, both of which are critical for added performance of these models. The research finds that deep learning algorithms, particularly RNNs, distinguish themselves when finding complex patterns in fraudulent job ads, compared to those of traditional machine learning algorithms. This paper also emphasizes the importance of having large datasets for successfully training the models. The deep learning model has performed better than traditional machine learning classifiers for scam detection.
This paper introduces an automated solution for fake job posts on social media platforms, relying on machine learning classification techniques. Specific efforts were directed to the analysis of feature extraction techniques (word frequency), and algorithms for classification: Decision Trees and Random Forests. They should train those algorithms on features expected to help in identifying fraudulent job advertisements, such as descriptions that do not seem real, fabricated company names, and inconsistent metadata. This system should automatically fight scams, saving job seekers from their snares. Machine learning models could be very helpful in the detection of fake jobs for protecting job seekers against scams.[5]

This work presents a successful combination of NLP techniques with machine learning methods to enable the detection of fraudulent job ads. TF-IDF and BoW were utilized as feature extraction techniques followed by a classification operation using Logistic Regression and SVM. The detection of fake job postings with a high rate of accuracy was called pioneered as a combination of NLP techniques together with machine learning, especially while the text of the ad was being analyzed. The study tried a hybrid model which combined many models to enhance prediction accuracy. The integration of NLP techniques with machine learning classifiers significantly improved the detection accuracy for so-called fake job advertisements.[6]

The study uses a Bi-Level Lutuce Memory (Bilateral Long Short Term Memory or Bi-LSTM) to distinguish fake job postings. The Bi-LSTM model-a class of deep learning architecture-is particularly well-matched to sequence data, making it an ideal candidate for text analysis tasks. The study reported a 98.71%-accurate model, which proves that the model is able to capture context from both the past and the future in job descriptions. The Bi-LSTM model performance was also compared with conventional learning machine models, with the results indicating that the deep learning model was more precise in recall. 98.71% accuracy performance is recorded in Bi-LSTM towards fake job ads.[7]

3. Proposed System

A research initiative develops an effective system for fake post detection through machine learning algorithm implementation. The research methodology includes the extraction of data followed by its preprocessing stage alongside implementation of different classification algorithms to determine post authenticity. The methodology consists of these steps:

Dataset Description: An experimental design scenario exists where researchers apply machine learning algorithms to dataset information to assess their ability in fraudulent post detection. A total of 18,000 dataset samples make up the Employment Scam Aegean Dataset (EMSCAD) that comes with both legitimate and fraudulent labels. Assessing fake post classification performance depends on the experimental method which identifies optimal algorithm selection. The evaluation employed the EMSCAD dataset that contains employment-related posts identified as either genuine or suspicious. The posts within this dataset display diverse parameters making it suited to demonstrate fake content detection on social platforms.

Data Preprocessing: The data must complete several preprocessing steps before machine learning algorithms can be applied, as it needs to be prepared in a usable format for modeling. The first step of text cleaning removes stop words, special characters, URLs, and irrelevant data points from the text. An algorithm requires the text to be tokenized, creating smaller units for easier processing. [8] The text preprocessing method divides the content into tokens through tokenization, which enables Hashing Vectorizer to extract important features such as word frequency data along with sentiment analysis results. To match the classification structure, the posts are assigned a binary label for fraudulent versus non-fraudulent content.

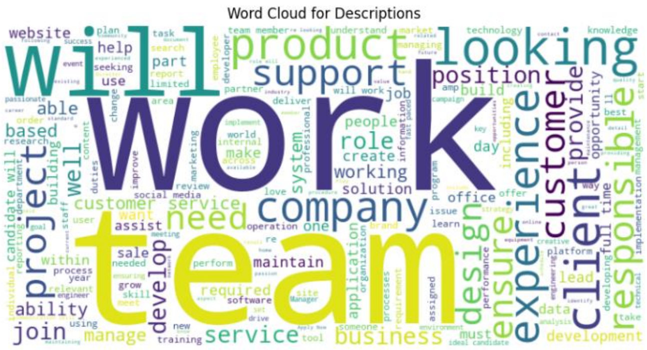

Figure.1 Word Cloud Graph

The word cloud highlights the most frequently used terms in job descriptions, with "work," "team," and "project" being the most prominent, indicating a strong emphasis on collaboration and task management. Words like "looking," "experience," and "customer" suggest that job descriptions often focus on seeking candidates with relevant experience and customer-oriented skills. The presence of terms such as "support," "product," and "service" implies that many job roles are centered around providing support and managing products or services.

Model Building

Several machine learning algorithms perform classification tests on the available dataset.

- XGBoost represents a gradient boosting algorithm which demonstrates high speed and performance when dealing with extensive datasets.
- The gradient boosting technique Catboost addresses categorical features better than other analogous algorithms.
- The Light Gradient Boosting system employs an efficient algorithm for handling big size datasets which increases training speed.
- The Random Forest system establishes many decision trees to create a more precise classification through result aggregation.
- Logistic Regression operates as a basic algorithm which serves to model binary classification matters.
- Decision Trees: A simple, yet interpretable, algorithm for classification.

Model Evaluation

For now, after having trained the models, the following evaluation metrics will be used to test their performance:

- Predictive Accuracy: An overall measure of correct predictions made by the models.
- Precision and Recall: To assess detection performance of the model in identifying fraudulent posts itself.
- F1 Score: H.M. between precision and recalling to bring a trade-off between the two.
- ROC-AUC Curve: To assess the model's ability to distinguish between genuine posts and fraudulent ones.

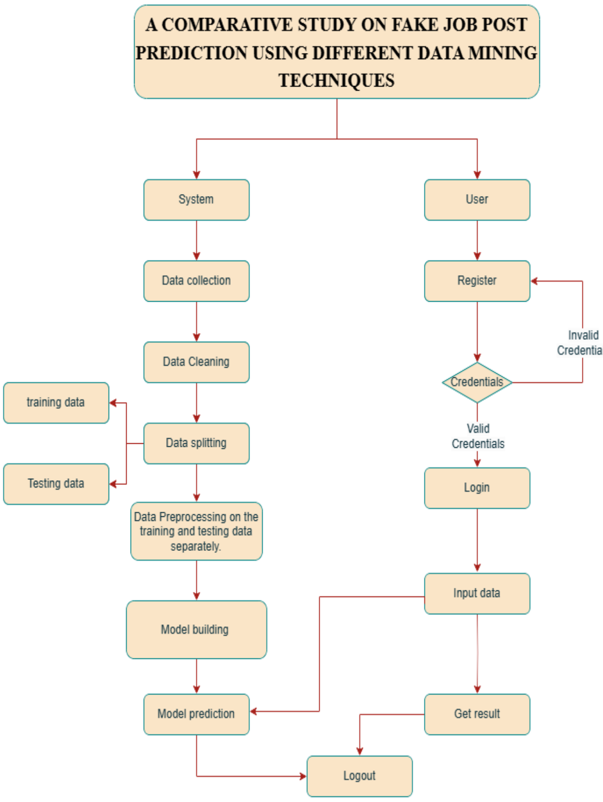

Figure.2 Proposed System Workflow

4. Methodology

Xgboost: XGBoost is an extremely powerful, an effective gradient boost algorithm well-suited to very large datasets. When working with Hashing Vectorizer, XGBoost receives text data transformed into a fixed-length vector, where each word is hashed to a unique index. The upside to using Hashing Vectorizer is that it reduces the feature space and this, in turn, reduces memory consumption, especially when dealing with high-dimensional text data.[9] A downside to this is that, in some instances, some information might get lost due to hash collisions. In spite of these limitations, the algorithm does well because boosting methods can iteratively learn from past mistakes, helping to strengthen its predictive power. Regularization of XGBoost is one of its strong features, as it helps in avoiding overfitting even with these hashed features. The availability of the parallel processing technique in the implementation is another reason one can view it as scalable to huge datasets and able to efficiently handle relatively large amounts of text data.

Random Forest: Random Forest works well with hashed features from the Hashing Vectorizer since it accepts both numerical and categorical fields. Each decision tree in the Random Forest algorithm conducts its splits by using subsets of the hashed features at each node. The benefit of using Hashing Vectorizer here is largely the reduction in feature space, which some particularly after very huge corpora of text data.[10] Random Forest's training process is stochastic in nature: each tree does not use the entire feature set, instead, every tree picks up a different random feature subset for training. This feature increases the generalization ability and reduces the risk of overfitting. Moreover, Random Forest is rather strong in handling noisy and irrelevant features, rendering the algorithm performant for text data, where irrelevant or less significant words might hash into the same feature index. While Hashing Vectorizer might lose some precision, ensemble learning by random forest actually helps offset this effect, yielding good performance even when working with hashed feature data.

Logistic Regression: Logistic Regression, a simple but highly interpretable module, has proved useful, especially for binary classification tasks such as determining whether or not a post is fake. The text data is fed to Logistic Regression in a sparse vector form of hashed values, thus allowing the model to avoid a lot of complex pre-processing and scale well with data sets of considerable size. Logistic Regression does so by applying the logistic function to the weighted summation of these hashed features to predict the probability of every class. Though Hashing Vectorizer may lead to hash collisions, it nevertheless can work well as long as some signal exists in the hashed features to separate the classes.[11] This model is particularly effective for problems where relationships between features and outcomes are approximately linear since it may unavoidably run into difficulties with non-linear relationships or very complex patterns of the data, which act as a hit by the loss of strong interactions between important words through hashed features.

In summary, even if the Hashing Vectorizer provides a trade-off between reducing the feature space and increasing the every bit of its scalability, it is still potentially overbearing due to new challenges such as hash collisions affecting the interpretability and performance of the model.[12] However, gradient boosting techniques such as XGBoost and Random Forest are robust enough to still deal fairly well with such hashed features and obtain good levels of performance on a large range of subsets of varying sizes. XGBoost performs well especially during large-scale issues; Random Forest is a model of strong generalization, and Logistic Regression is efficient and easy to interpret for binary classification tasks.

5. Results and Discussion

Perfect Performance: The model has achieved a 100% accuracy rate, correctly identifying all fraudulent and non-fraudulent cases with no errors. Balanced Classification: Both true positives and true negatives stand at 13,620, indicating a balanced and flawless classification of both classes. Potential Data Concerns: This level of performance is highly unusual, suggesting an exceptionally well-performing model or potential issues such as data leakage.

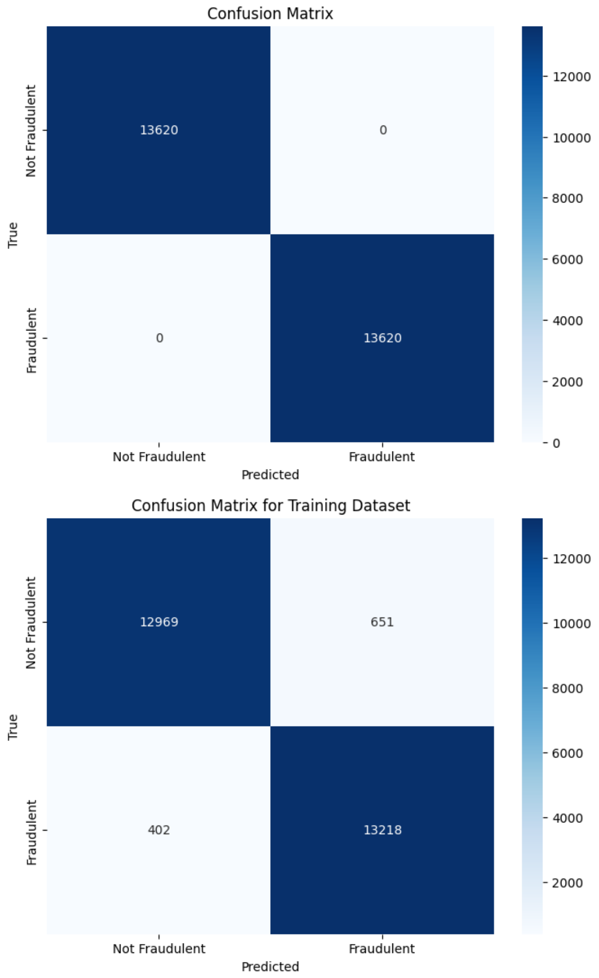

Figure.3 Performance Analysis

True Positives: The model correctly identified 13,218 fraudulent cases, showcasing its strong ability to detect fraud. True Negatives: 12,969 instances were accurately classified as not fraudulent, highlighting reliable non-fraud detection. False Positives: There were 651 non-fraudulent instances mistakenly flagged as fraudulent, indicating some over-caution. False Negatives: 402 fraudulent cases went unnoticed, suggesting a need for improved fraud detection accuracy.

True Positives: The model correctly identified 13,564 fraudulent cases, showcasing its strong ability to detect fraud. True Negatives: 13,620 instances were accurately classified as not fraudulent, highlighting reliable non-fraud detection.

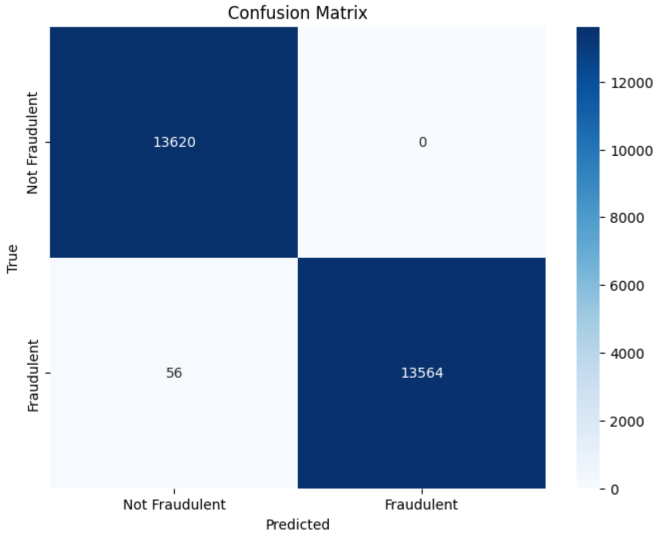

Figure.4 Performance Measurement

False Negatives: There were 56 fraudulent cases incorrectly classified as not fraudulent, indicating some missed fraud cases. Zero False Positives: Impressively, no non-fraudulent cases were mistakenly flagged as fraudulent.

6. Conclusion

The first confusion matrix shows the model performed exceptionally well. It had an accuracy of 100%, as it exactly classified all fraudulent and non-fraudulent instances correctly. This is a rare result that may indicate a perfectly well-tuned model or, it may also point toward problems like data leakage. In the second confusion matrix, the model performed well in identifying fraudulent and non-fraudulent cases. But it was slightly overzealous in flagging fraudulent transactions, with 651 false positives recorded. The findings also included 402 false negatives, in which it did not capture existing fraudulent transactions, calling for further improvement in the fraud detection process.

The third confusion matrix performed quite impressively with only 56 false negatives and no false positives. This shows that the model is highly reliable on its non-fraudulent case and still formulates strong fraud detection. Broadly speaking, the matrices suggest good performance. Continuous refining through constant monitoring will remain a design approach for maintaining the accuracy of results generated by the model. In short, it is a commendable performance with a high level of accuracy in detecting fraudulent and non-fraudulent cases. The level of precision is laudable, and continuous improvements need to be worked on to better where necessary.

Future Enhancement

Although the current methods using XGBoost, Random Forest, and Logistic Regression with a hashing vectorizer provide a solid initial basis for text classification tasks, it would be useful to consider improvements in future research directions.

- Shifting away from Hashing Vectorizer towards more advanced vectorization techniques like Word2Vec, GloVe, or BERT could increase a model's ability to capture word semantic relationships and possibly improve performance. These techniques treat words in a more complex and informative manner, thus assisting in performing more complicated tasks such as detecting more subtle patterns in fake posts.[13]
- To inner some concerns regarding hash collisions imposed by Hashing Vectorizer, future advancements could delve into hybrid mechanisms combining hashing with other feature extraction methods, such as TF-IDF and word embeddings, to preserve the richness in feature space while enhancing efficiency.
- Further enhancements could be done to XGBoost, Random Forest, and Logistic Regression through more extensive tuning of hyperparameters. Techniques such as grid search or Bayesian optimization could be applied to find the best optimal values for parameters for all models to improve their predictions. [14]
- Incorporating other models into the ensemble would increase robustness even more, like stacking of such model predictions for better generalization and performance across multiple varieties of fake posts.
- Broadening the streams enabled by the system for real-time data could further classification of posts as they arise, as opposed to merely relying on static data. Particularly for platforms dealing with continued high-intensity volumes, this could be a more effective approach.
- To foster wider acceptance of models for fraud detection, work should be done toward making models, especially XGBoost and Random Forest, more interpretable so that end-users have an intuitive understanding of why a post received a particular classification as being fake. Techniques such as LIME and SHAP could be combined to explain model predictions and help win trust in automated systems. [15]
- The current system relies on a specific dataset (like the Employment Scam Aegean Dataset). Future enhancements could involve adapting the system for use in detecting fake posts across various domains, such as financial fraud, health misinformation, or political manipulation, by training on diverse datasets.

Reference

[1] C. S. Anita, P. Nagarajan, G. Aditya Sairam, P. Ganesh, and G. Deepakkumar, “Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms,” Vol. 11, No. 2, pp. 2237–0722, 2021.

[2] P. Khandagale, A. Utekar, A. Dhonde, and S. S. Karve, "Fake Job Detection Using Machine Learning," Vol. 10, 2022, DOI: 10.22214/ijraset.2022.41641.

[3] V. Anbarasu, S. Selvakani, and M. K. Vasumathi, "Fake Job Prediction Using Machine Learning," DOI: 10.32692/IJDI-ERET/13.1.2024.2403.

[4] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, "Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms," Revista Gestão Inovação e Tecnologias, Vol. 11, No. 2, pp. 642–650, Jun. 2021, DOI: 10.47059/REVISTAGEINTEC.V11I2.1701.

[5] P. Kumar, A. M. G, and M. S. H, "Fake Job Post Prediction Using Machine Learning Algorithms," 2022.

[6] S. Pillai, "Detecting Fake Job Postings Using Bidirectional LSTM," International Research Journal of Modernization in Engineering Technology and Science, Vol. 3883, Apr. 2023, DOI: 10.56726/IRJMETS35202.

[7] Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, and S. Ludi, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches," Neural Process Lett, Vol. 54, No. 3, pp. 2219–2247, Jun. 2022, DOI: 10.1007/S11063-021-10727-Z.

[8] R. Roshan, I. A. Bhacho, and S. Zai, "Comparative Analysis of TF–IDF and Hashing Vectorizer for Fake News Detection in Sindhi: A Machine Learning and Deep Learning Approach," Engineering Proceedings 2023, Vol. 46, Page 5, Vol. 46, No. 1, p. 5, Sep. 2023, DOI: 10.3390/ENGPROC2023046005.

[9] Srikanth, M. Rashmi, S. Ramu, and R. M. R. Guddeti, "A Novel Fake Job Posting Detection: An Empirical Study and Performance Evaluation Using ML and Ensemble Techniques," Lecture Notes in Electrical Engineering, Vol. 1049 LNEE, pp. 219–234, 2023, DOI: 10.1007/978-981-99-3569-7_16.

[10] "Ensemble Modeling on Job Scam Detection," ResearchGate, Available: https://www.researchgate.net/publication/351936898_Ensemble_Modeling_on_Job_Scam_Detection, Accessed: Feb. 12, 2025.

[11] R. Narayanan, R. Bhavani, R. Balamanigandan, and A. A. S. Priscilla, "Analyzing the Performance of Novel Logistic Regression over Linear Regression Algorithms for Predicting Fake Job with Improved Accuracy," 5th International Conference on Electronics and Sustainable Communication Systems, ICESC 2024 - Proceedings, pp. 1728–1732, 2024, DOI: 10.1109/ICESC60852.2024.10690033.

[12] S. U. B, V. R. Singh, S. P, and A. Dhage, "Fake Job Post Prediction Using Data Mining," Journal of Scientific Research and Technology, pp. 39–47, May 2023, DOI: 10.5281/ZENODO.7954261.

[13] O. Truică, E. S. Apostol, and P. Karras, "DANES: Deep Neural Network Ensemble Architecture for Social and Textual Context-aware Fake News Detection," Knowl Based Syst, Vol. 294, Feb. 2023, DOI: 10.1016/j.knosys.2024.111715.

[14] J. Y. Khan, Md. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," Machine Learning with Applications, Vol. 4, p. 100032, Jun. 2021, DOI: 10.1016/j.mlwa.2021.100032.

[15] V. Dua, A. Rajpal, S. Rajpal, M. Agarwal, and N. Kumar, "I-FLASH: Interpretable Fake News Detector Using LIME and SHAP," Wirel Pers Commun, Vol. 131, No. 4, pp. 2841–2874, Aug. 2023, DOI: 10.1007/S11277-023-10582-2.