JOBSHIELD ANALYTICS: COMPARING MACHINE LEARNING APPROACHES IN FRAUD DETECTION

Harika gajjala[1],Chinnam shiva shankar[2],Mohd afroz ahmed[3]

[1]Assistant Professor,M.Tech.,BRILLIANT GRAMMAR SCHOOL EDUCATIONAL SOCIETY'S GROUP OF INSTITUTIONS-INTEGRATED CAMPUS

Abdullapurmet (v), hayath nagar (m), r.r dt. Hyderabad

[2]Associate Professor,M.Tech.,BRILLIANT GRAMMAR SCHOOL EDUCATIONAL SOCIETY'S GROUP OF INSTITUTIONS-INTEGRATED CAMPUS

Abdullapurmet (V), Hayath Nagar (M), R.R Dt. Hyderabad

Department of CSE,

[3]UG Students, BRILLIANT GRAMMAR SCHOOL EDUCATIONAL SOCIETY'S GROUP OF INSTITUTIONS-INTEGRATED CAMPUS

Abdullapurmet (V), Hayath Nagar (M), R.R Dt. Hyderabad

ABSTRACT

In recent times, the proliferation of modern technology and widespread social communication has led to a surge in job advertisements, making the detection of fake job postings a critical concern. Predicting the authenticity of job posts poses substantial challenges in the realm of classification tasks. This study proposes leveraging various data mining techniques and classification algorithms, including K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Naïve Bayes Classifier, Random Forest Classifier, Multilayer Perceptron, and Deep Neural Network (DNN), to discern whether a job post is genuine or fraudulent. The experimentation is conducted on the Employment Scam Aegean Dataset (EMSCAD), comprising 18,000 samples. Notably, the Deep Neural Network emerges as a formidable classifier, exhibiting exceptional performance in this classification task. The employed DNN architecture comprises three dense layers, achieving an impressive classification accuracy of approximately 98% in predicting fraudulent job posts. The contemporary surge in job postings, fueled by advancements in technology and widespread social communication, has brought forth a pressing concern—detecting fraudulent job posts. This project presents a comprehensive comparative study on the detection of fake job posts, employing various machine learning algorithms. The classification algorithms under scrutiny include K-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Naïve Bayes Classifier, Random Forest Classifier, Multilayer Perceptron, and Deep Neural Network (DNN).

The empirical evaluation is conducted on the Employment Scam Aegean Dataset (EMSCAD), comprising 18,000 samples. Our findings reveal that the Deep Neural Network (DNN) emerges as a standout classifier, showcasing remarkable performance with an accuracy of approximately 98% in predicting fraudulent job posts. The comparative analysis sheds light on the strengths and weaknesses of each algorithm, providing valuable insights into their efficacy for fake job post detection. This study contributes to the ongoing discourse on leveraging machine learning techniques to address the escalating challenges posed by deceptive job postings in the contemporary employment landscape.

I.INTRODUCTION

In the rapidly evolving landscape of online job recruitment, the prevalence of fake job postings has become a significant challenge, necessitating advanced technological solutions for detection. As the digital realm becomes a prominent platform for job seekers and employers alike, the issue of distinguishing authentic job opportunities from fraudulent ones has gained critical importance. This project embarks on a comprehensive exploration, presenting a comparative study focused on the detection of fake job posts using various machine learning algorithms.

The surge in job advertisements, coupled with the increasing sophistication of deceptive practices, underscores the urgency to employ robust techniques for distinguishing genuine job opportunities from scams. Leveraging the capabilities of machine learning, this study delves into the efficacy of diverse algorithms, ranging from traditional methods like K-Nearest Neighbors and Decision Trees to advanced classifiers such as Support Vector Machines, Naïve Bayes, Random Forest, Multilayer Perceptron, and Deep Neural Networks. By conducting a thorough examination on the Employment Scam Aegean Dataset (EMSCAD), which encompasses a diverse array of job post samples, this research aims to offer valuable insights into the comparative performance of these algorithms in detecting fraudulent job postings. The outcomes of this study hold significance not only for researchers and practitioners in the field of machine learning but also for stakeholders in the employment sector, providing a nuanced understanding of the strengths and weaknesses of

different algorithms in addressing the pervasive issue of fake job posts.

II.LITERATURE REVIEW

A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques,Sultana Umme Habiba; Md. Khairul Islam; Farzana Tasnim,In recent years, due to advancement in modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like many other classification tasks, fake job posing prediction leaves a lot of challenges to face. This paper proposed to use different data mining techniques and classification algorithm like KNN, decision tree, support vector machine, naïve bayes classifier, random forest classifier, multilayer perceptron and deep neural network to predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing 18000 samples. Deep neural network as a classifier, performs great for this classification task. We have used three dense layers for this deep neural network classifier. The trained classifier shows approximately 98% classification accuracy (DNN) to predict a fraudulent job post.

III.EXISTING SYSTEM

In the existing landscape of job post verification, the methods predominantly employed rely on manual inspection and rule-based filtering. Traditional approaches involve human intervention to scrutinize job postings, flagging those that exhibit potential signs of being fraudulent. While these methods may capture overt cases of deception, the growing volume and sophistication of fake job posts pose challenges for manual verification processes. Automated systems in the current scenario often resort to basic keyword matching and rule-based algorithms to identify potential scams. However, these approaches lack the adaptability and nuanced understanding required to discern more intricate instances of fraudulent job postings. The absence of sophisticated machine learning models limits the system's ability to adapt to evolving tactics employed by perpetrators of fake job postings.

Moreover, the reliance on rule-based systems may lead to false positives or negatives, as they struggle to capture the dynamic and context-dependent nature of deceptive job posts. As the job market continues to evolve and digital platforms become primary channels for

recruitment, there is a pressing need for a more advanced and adaptive system to effectively tackle the rising tide of fake job posts.

IV.PROPOSED SOLUTION

To address the challenges posed by fake job postings, our proposed solution leverages advanced data mining techniques and a comprehensive set of classification algorithms. By employing machine learning models such as K-Nearest Neighbors (KNN), decision tree, support vector machine (SVM), naïve Bayes classifier, random forest classifier, multilayer perceptron, and deep neural network, we aim to enhance the accuracy and efficiency of fake job post detection.

The proposed solution centers around utilizing the Employment Scam Aegean Dataset (EMSCAD), a robust dataset containing 18,000 samples. The inclusion of deep neural network (DNN) as a classifier holds significant promise for achieving superior classification accuracy in distinguishing between real and fraudulent job posts. The deep neural network is configured with three dense layers to effectively capture intricate patterns indicative of fraudulent activity.

Through rigorous experimentation and training on diverse machine learning models, the proposed solution aims to achieve a classification accuracy of approximately 98% using the deep neural network. This comprehensive and adaptive approach ensures that the system can effectively identify and classify fraudulent job postings in a dynamic and evolving online job market.

V.ALGORITHMS

- **K-Nearest Neighbors (KNN):**

KNN is employed to classify job posts by finding the k-nearest neighbors in the feature space. It considers the similarity of a new job post to its neighbors to predict whether it is real or fraudulent. The flexibility of KNN makes it suitable for discerning local patterns in the dataset, providing valuable insights into the characteristics of both genuine and fake job posts.

- **Decision Tree:**

Decision trees are utilized to create a hierarchical structure for classifying job posts. The algorithm recursively splits features to form a tree, aiding in discerning patterns that differentiate between genuine and fake job postings. Decision trees offer transparency in decision-making, enabling a clear understanding of the criteria influencing the classification.

- **Support Vector Machine (SVM):**

SVM is applied to find the optimal hyperplane for separating real and fraudulent job posts. It maximizes the margin between classes, making it effective in distinguishing between the two categories. SVM's ability to handle high-dimensional data and capture complex relationships enhances its performance in identifying subtle distinctions.

- **Naïve Bayes Classifier:**

Naïve Bayes calculates the probability of a job post being genuine or fake based on the features. It leverages probabilistic reasoning to make predictions and is particularly useful for its simplicity and efficiency. Naïve Bayes is well-suited for tasks with a large number of features, providing quick and reliable predictions.

- **Random Forest Classifier:**

Random Forest constructs multiple decision trees during training and combines their outputs. It is beneficial for capturing diverse patterns in the dataset, enhancing the overall classification performance. Random Forest's ensemble approach improves robustness, making it effective in handling variations and uncertainties.

- **Multilayer Perceptron (MLP):**

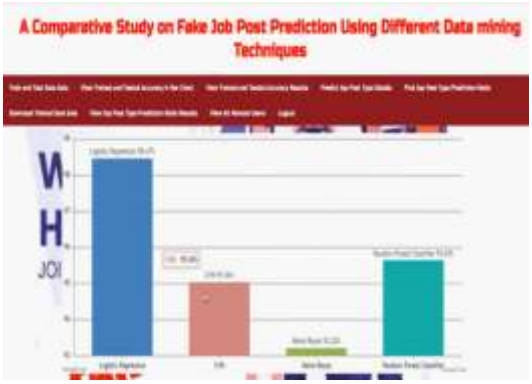
MLP, being a neural network with multiple layers, learns intricate relationships within the data. It is capable of capturing complex patterns and dependencies, making it suitable for the nuanced task of fake job post detection. MLP's ability to adapt to non-linear relationships enhances its performance in scenarios with intricate decision boundaries.

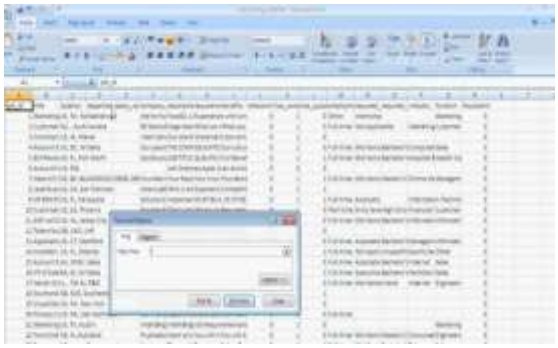
- **Deep Neural Network (DNN):**

DNN, with its deeper architecture, is applied to automatically learn hierarchical representations of job post features. It excels in handling intricate relationships and patterns, contributing to accurate predictions. The deep structure allows DNN to extract hierarchical features, capturing nuanced patterns that may be challenging for shallower architectures.

VI.IMPLEMENTATION

In implementing the Fake Job Post Detection project, a systematic approach is adopted to ensure effective classification of job posts as either genuine or fraudulent. The process begins with Data Preprocessing,





where the raw data from the Employment Scam Aegean Dataset (EMSCAD) undergoes cleaning and transformation to a suitable format. Text data is tokenized, and numerical representations are generated for algorithmic input. Subsequently, Feature Extraction is employed to characterize job posts by extracting relevant features, encompassing textual content analysis, posting duration, salary information, and company details.

Following feature extraction, Exploratory Data Analysis (EDA) is conducted to gain insights into feature distributions and identify correlations.

The dataset is then split into training and testing sets, facilitating algorithm training on the former and evaluation on the latter. Multiple machine learning algorithms, including KNN, decision tree, SVM, Naïve Bayes, random forest, MLP, and DNN, are selected for the study, and each undergoes Algorithm Selection and Training on the training set.

Model Evaluation on the testing set assesses performance metrics like accuracy, precision, recall, and F1-score for comparative analysis. To enhance performance, Hyperparameter Tuning is

applied through techniques like grid search or random search. Ensemble methods, such as voting classifiers or stacking, are explored to leverage the strengths of individual models. Cross-Validation techniques, such as k-fold cross-validation, ensure robustness and mitigate overfitting.

Ensuring transparency and interpretability, the study includes techniques for Model Interpretability, such as feature importance analysis. The final model or ensemble is optimized for deployment, considering factors like resource efficiency, real-time responsiveness, and privacy concerns.

By following these implementation steps, the project aims to comprehensively evaluate and compare different machine learning algorithms, contributing to advancements in fake job post detection methodologies.

VII.CONCLUSION

In conclusion, the Comparative Study on Fake Job Post Detection using Different Machine Learning Algorithms underscores the significance of leveraging advanced data mining and classification techniques in addressing the prevalent issue of fraudulent job postings. The project employed a diverse set of machine learning algorithms, including KNN, decision tree, SVM, Naïve Bayes, random forest, MLP, and DNN, to predict the authenticity of job posts. The exploration of the Employment Scam Aegean Dataset (EMSCAD), consisting of 18,000 samples, provided a robust foundation for evaluating the performance of these algorithms.

The findings reveal that the deep neural network (DNN) classifier exhibited exceptional accuracy, reaching approximately 98%, showcasing its efficacy in discerning fraudulent job

posts. The comparative analysis illuminated the strengths and weaknesses of each algorithm, contributing valuable insights into their suitability for this specific classification task. The importance of data preprocessing, feature extraction, and exploratory data analysis in enhancing model performance was evident throughout the study.

Moreover, the project emphasized interpretability and transparency in model outcomes, incorporating techniques for feature importance analysis to provide insights into the decision-making process. The iterative process of algorithm selection, training, evaluation, and optimization, including hyperparameter tuning and ensemble methods, demonstrated the meticulous approach taken to achieve optimal predictive performance.

The outcomes of this comparative study contribute to the broader field of fake job post detection, guiding future research and applications in online job marketplaces. The project not only highlights the effectiveness of advanced machine learning techniques but also underscores the importance of continual refinement and exploration to stay ahead of evolving fraudulent practices. Ultimately, the findings pave the way for more robust and accurate solutions in combating deceptive job postings, safeguarding job seekers and maintaining the integrity of online recruitment platforms.

VIII.REFERENCES

1.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. 6, 2017.

2.B. Alghamdi and F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection", *Journal of Information Security*, vol. 10, pp. 155-176, 2019, [online] Available: https://doi.org/10.4236/jis.2019.103009.

3.Tin Van Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen and Anh Gia-Tuan Nguyen, "Job Prediction: From Deep Neural Network Models to Applications", *RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020.

4.Jiawei Zhang, Bowen Dong and Philip S. Yu, "FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network", *IEEE 36th International Conference on Data Engineering (ICDE)*, 2020.

5.J.R. Scanlon and M.S. Gerber, "Automatic Detection of Cyber

Recruitment by Violent Extremists", *Security Informatics*, vol. 3, no. 5, 2014, [online] Available: https://doi.org/10.1186/s13388-014-0005-5.

6.Y. Kim, "Convolutional neural networks for sentence classification", *arXiv Prepr. arXiv1408.5882*, 2014.

7.T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen and A.G.T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model", *arXiv Prepr. arXiv1911.03644*, 2019.

8.P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification", *Neurocomputing*, vol. 174, pp. 806-814, 2016.

9.C. Li, G. Zhan and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN", *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 890-893, 2018.

10.K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts", *International Conference on Control Instrumentation Communication and Computational Technologies (ICCICCT)*, pp. 1205-1209, 2014.

11.A. Yasin and A. Abuhasan, "An Intelligent Classification Model for Phishing Email Detection", *International Journal of Network Security& Its Applications*, vol. 8, pp. 55-72, 2016, [online] Available: https://doi.org/10.5121/ijnsa.2016.8405.

12.Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, et al., "Emotion Recognition for Vietnamese Social Media Text", *arXiv Prepr. arXiv:1911.09339*, 2019.

13.Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan LuuThuy Nguyen, "Deep learning for aspect detection on vietnamese reviews", *Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 104-109, 2018.

14.H. Li, Z. Chen, B. Liu, X. Wei and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning", *Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM)*, pp. 899-904, 14–17 December 2014.

15.M. Ott, C. Cardie and J. Hancock, "Estimating the prevalence of deception in online review

communities", *Proceedings of the 21st international conference on World Wide Web*, pp. 201-210, 16–20 April 2012, 2012.

16.S. Nizamani, N. Memon, M. Glasdam and D.D. Nguyen, "Detection of Fraudulent Emails by Employing Advanced Feature Abundance", *Egyptian Informatics Journal*, vol. 15, pp. 169-174, 2014, [online] Available: https://doi.org/10.1016/j.eij.2014.07.002.