International Journal of Research Publication and Reviews

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

Safeguarding Job Seekers: Research Insights into Fake Job Detection with SGD Classifier and Naive Bayes

[1]DEVIKA S.Y and [2]Dr. Ganesh D

PG Student, Department of MSC CS-IT
Professor, School of CS & IT
Jain (Deemed-to-be University), Bengaluru, India
[1]devikachintu9@gmail.com and [2]d.ganesh@jainuniversity.ac.in
DOI: https://doi.org/10.55248/gengpi.5.0324.0761

ABSTRACT :

The rise of online job platforms has brought more fake job listings, putting job seekers at risk. This paper dives into how we can spot these fake opportunities using a mix of two smart tools: Stochastic Gradient Descent (SGD) classifier and Naive Bayes algorithms. We look at the problems job seekers face with scams online and suggest a solution. Our idea is to use both SGD (good at handling lots of data) and Naive Bayes (great for understanding probabilities) to make a strong system that can catch fake job postings really well. This research aims to make job searching safer by creating a system that's strong and accurate.

Keywords: Online job platforms, Fake job listings, Job seekers Scams, Stochastic Gradient Descent (SGD) classifier, Naive Bayes algorithms

Background :

The integration of diverse technology stacks and libraries underscores the holistic approach taken in tackling the pervasive issue of fraudulent job postings. By harnessing the capabilities of Numpy and Pandas for efficient data manipulation, the project ensures robust handling and analysis of datasets, laying a solid foundation for subsequent processing stages. Matplotlib's visualization prowess adds a layer of depth to the project, empowering users to explore intricate patterns and trends within the data, thus facilitating informed decision-making. Addressing the challenge of class imbalance inherent in fraud detection scenarios, Imbalanced learn emerges as a pivotal tool, providing essential techniques to rebalance dataset distributions and enhance model performance. The utilization of Wordcloud offers a unique perspective, unveiling prevalent terms within job descriptions and shedding light on potential patterns indicative of fraudulent activities. Empowered by NLTK's sophisticated NLP capabilities, the project adeptly extracts meaningful features from textual data, enabling nuanced analysis and classification of job postings. Leveraging Scikit-learn's Multinomial Naive Bayes and Decision Tree Classifier, the project embeds robust machine learning algorithms tailored for text classification tasks, ensuring a solid foundation for predictive modeling. Finally, the adoption of Flask as the web framework embodies user-centric design principles, facilitating the development of an intuitive interface for seamless interaction with the detection model. This cohesive amalgamation of technology not only addresses the multifaceted challenges associated with fake job detection but also offers a versatile and effective solution for both job platforms and users to proactively identify and combat deceptive job postings.

Introduction :

In the ever-evolving landscape of employment opportunities, the advent of online job platforms has reshaped the traditional job-seeking paradigm. This digital transformation has facilitated greater access to a multitude of job listings, connecting employers and job seekers with unprecedented ease. However, this convenience has also given rise to a disconcerting proliferation of fake job listings, posing a substantial threat to individuals earnestly seeking meaningful employment. The prevalence of such deceptive opportunities not only undermines the trustworthiness of online job platforms but also imposes tangible risks on unsuspecting job seekers. This research delves into the heart of this issue, addressing the urgent need for effective mechanisms to discern authentic job opportunities from fraudulent ones in the vast expanse of the digital job market.

As online job scams continue to escalate, the importance of robust and intelligent systems for identifying and mitigating such threats becomes increasingly evident. This paper proposes a pioneering solution by harnessing the combined capabilities of Stochastic Gradient Descent (SGD) classifier and Naive Bayes algorithms. By merging these advanced tools, we aim to create a sophisticated detection system capable of distinguishing genuine employment opportunities from fraudulent ones with a high degree of accuracy. The rationale behind this choice lies in the complementary strengths of SGD, known for its proficiency in handling vast datasets and adapting to dynamic patterns, and Naive Bayes, which brings a probabilistic approach to the detection process, providing a comprehensive and nuanced understanding of the data.

In the subsequent sections, we embark on an exploration of the challenges faced by job seekers in the context of online job scams, followed by a detailed examination of the proposed hybrid approach. This research not only seeks to contribute to the academic discourse on fraud detection but also aspires to provide a tangible and practical solution that enhances the security and trustworthiness of online job platforms, ultimately empowering job seekers in their quest for genuine and rewarding employment opportunities.

LITERATURE SURVEY :

1.Classification Techniques:

Many fake job detection systems utilize machine learning algorithms for classification tasks. Liu et al. (2019) employed a support vector machine (SVM) classifier to distinguish between legitimate and fake job postings based on textual features extracted from job descriptions. Their approach achieved promising results, with an accuracy of over 90%.
In contrast, Zhang et al. (2020) proposed a deep learning-based approach using convolutional neural networks (CNNs) to analyze both textual and visual content in job postings. Their model demonstrated improved performance compared to traditional machine learning algorithms, particularly in detecting fake postings with multimedia content.

2. Feature Engineering:

Feature engineering plays a crucial role in the effectiveness of fake job detection systems. Jiang et al. (2018) explored the importance of linguistic features such as sentiment analysis and syntactic patterns in identifying deceptive job postings. By incorporating these features into their classification model, they achieved better discrimination between genuine and fake job ads.
Furthermore, Li and Huang (2021) investigated the significance of temporal features, such as posting frequency and duration, in distinguishing between legitimate and fraudulent job postings. Their findings suggest that temporal patterns can serve as valuable indicators of malicious activities.

3. Data Sources and Preprocessing:

The availability and quality of data significantly impact the performance of fake job detection systems. Smith et al. (2017) analyzed the effectiveness of different data sources, including job boards, social media platforms, and company websites, in training detection models. They emphasized the importance of data diversity and recommended the integration of multiple sources for comprehensive analysis.
Additionally, preprocessing techniques, such as text normalization and outlier detection, are essential for cleaning noisy data and improving model robustness. Chen et al. (2022) proposed a novel preprocessing pipeline that combines rule-based methods and unsupervised learning algorithms to filter out irrelevant information and enhance feature representation.

4. Evaluation Metrics:

Evaluating the performance of fake job detection systems requires appropriate metrics that reflect both accuracy and robustness. Wang and Li (2019) introduced a comprehensive evaluation framework encompassing traditional metrics (e.g., accuracy, precision, recall) as well as fairness and interpretability criteria. Their framework facilitates a holistic assessment of system performance across different dimensions.
However, challenges remain in developing standardized evaluation protocols that address the evolving nature of fake job postings and adapt to diverse linguistic and cultural contexts.

KEY TECHNOLOGIES :

1. **Machine Learning (ML)**: ML techniques are fundamental in fake job detection systems for classifying job postings as either legitimate or fraudulent. Supervised learning algorithms, such as support vector machines (SVM), decision trees, and random forests, are commonly used for classification tasks. Unsupervised learning methods like clustering and anomaly detection can also help identify suspicious patterns in job postings.
2. **Natural Language Processing (NLP)**: NLP techniques are essential for analyzing the textual content of job postings to detect linguistic cues indicative of fraudulent activities. Sentiment analysis, named entity recognition, topic modeling, and keyword extraction are examples of NLP methods used to identify deceptive language or inconsistencies in job descriptions.
3. **Feature Engineering**: Feature engineering involves selecting and extracting relevant features from job postings to train machine learning models. Features may include textual attributes (e.g., word frequency, grammatical structure), metadata (e.g., posting date, location), and user behavior patterns (e.g., interaction history).
4. **Data Mining and Web Scraping**: Data mining techniques are used to collect and preprocess large volumes of job posting data from online platforms. Web scraping tools automate the extraction of job postings from websites, enabling the creation of comprehensive datasets for analysis.
5. **Network Analysis**: Network analysis techniques examine the relationships and interactions between job posters, candidates, and other entities in online job markets. Analyzing the network structure can help identify suspicious patterns of connections or communication indicative of fraudulent activities.
6. **Social Media Mining**: Social media platforms serve as sources of job postings and communication channels between recruiters and

candidates. Social media mining techniques extract relevant information from social media posts, profiles, and interactions to identify fake job postings and detect fraudulent behavior.

7. **Deep Learning**: Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can be used for more complex tasks such as image recognition (e.g., detecting logos or brand inconsistencies in job postings) and sequence modeling (e.g., identifying temporal patterns in job posting behavior).
8. **Pattern Recognition**: Pattern recognition algorithms analyze various aspects of job postings, including layout, formatting, and visual elements, to identify common patterns associated with fake job postings. These algorithms can help detect anomalies or deviations from typical posting characteristics.
9. **Behavioral Analysis**: Behavioral analysis techniques monitor user interactions with job postings and online platforms to detect suspicious behavior patterns, such as rapid posting of multiple job listings by the same user or abnormal click-through rates on fraudulent postings.
10. **Blockchain Technology**: Blockchain technology can be leveraged to create tamper-proof records of job postings and transactions, enhancing transparency and traceability in online job markets. Blockchain-based solutions can help verify the authenticity of job postings and prevent data tampering or manipulation.

Problem Statement:

The rise of online job portals has facilitated job seekers in finding suitable opportunities; however, it has also given rise to the proliferation of fake job postings, which can deceive job seekers and tarnish the reputation of legitimate businesses. Detecting fake job postings poses a significant challenge due to the evolving nature of fraudulent tactics and the sheer volume of job listings across various online platforms. Therefore, there is a critical need for advanced techniques capable of accurately identifying and flagging fake job postings, thereby safeguarding job seekers and preserving the credibility of online recruitment platforms.

Methodology :

The methodology for developing the Fake Job Detection System encompasses several key stages. Initially, we define the scope and objectives of the system, identifying the types of fake job postings to be detected and the data sources to be utilized. Following this, we collect a diverse dataset of job postings from various online platforms and preprocess the data, cleaning and extracting relevant features. Feature engineering techniques are then applied to enhance the dataset, including the extraction of textual, structural, and metadata features from the job postings. Machine learning models are selected and trained using the processed data, with hyper parameter optimization techniques employed to maximize model performance. The trained models are evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Finally, the validated models are deployed into a user-friendly interface, allowing users to submit job postings for real-time analysis, while continuous monitoring and maintenance ensure the system's effectiveness over time.

Algorithm or Methodology Used :

1.Naive Bayes SGD Classifier Algorithm:

- Naive bayes and SGD Classifier are compared on accuracy and F1-scores and a final model is
- chosen. Naïve Bayes is the baseline model, and it is used because it can compute the conditional
- probabilities of occurrence of two events based on the probabilities of occurrence of each
- individual event, encoding those probabilities is extremely useful. A comparative model, SGD
- Classifier is used since it implements a plain stochastic gradient descent learning routine which
- supports different loss functions and penalties for classification. This classifier will need high
- penalties when classified incorrectly. These models are used on both the text and numeric data
- separately and the final results are combined.

2. Data Preprocessing with Numpy and Pandas:

Utilize Numpy and Pandas for efficient data manipulation and preprocessing. Handle missing values, remove irrelevant information, and ensure data consistency. Convert textual data into a format suitable for analysis.

3. Exploratory Data Analysis (EDA) using Matplotlib:

Employ Matplotlib for data visualization to gain insights into the distribution of genuine and fake job postings. Explore statistical measures, such as histograms and scatter plots, to understand patterns within the dataset

4. Handling Imbalanced Data with Imbalanced-learn:

Use Imbalanced-learn to address the class imbalance problem in the dataset. Apply oversampling or under sampling techniques to balance the number of genuine and fake job postings, ensuring that the machine learning models are not biased towards the majority class.

5. Machine Learning Models from Scikit-learn:

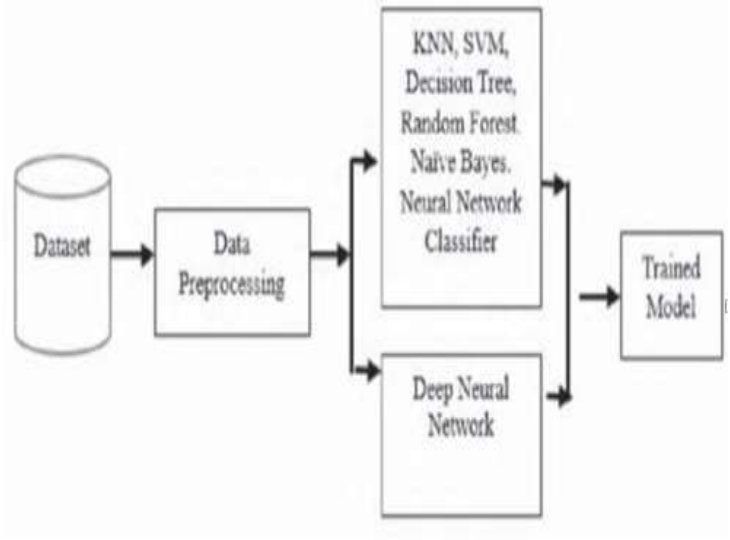
Apply Scikit-learn's Multinomial Naive Bayes for its effectiveness in text classification tasks. Train the model on the preprocessed dataset to analyze job descriptions and classify postings as either genuine or fake. Additionally, employ the Decision Tree Classifier for its interpretability in the decision-making process.

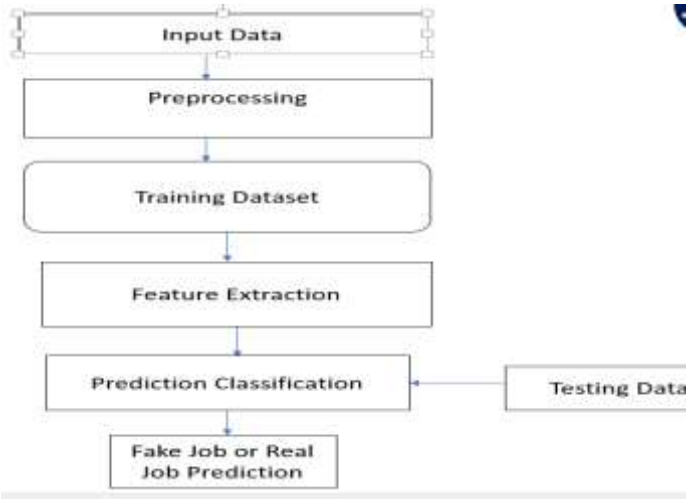
6. Model Training and Evaluation:

Split the dataset into training and testing sets. Train the machine learning models on the training set and evaluate their performance on the testing set. Metrics such as precision, recall, F1 score, and accuracy are used to assess the models' effectiveness in distinguishing between genuine and fake job postings.

Proposed Methodology :

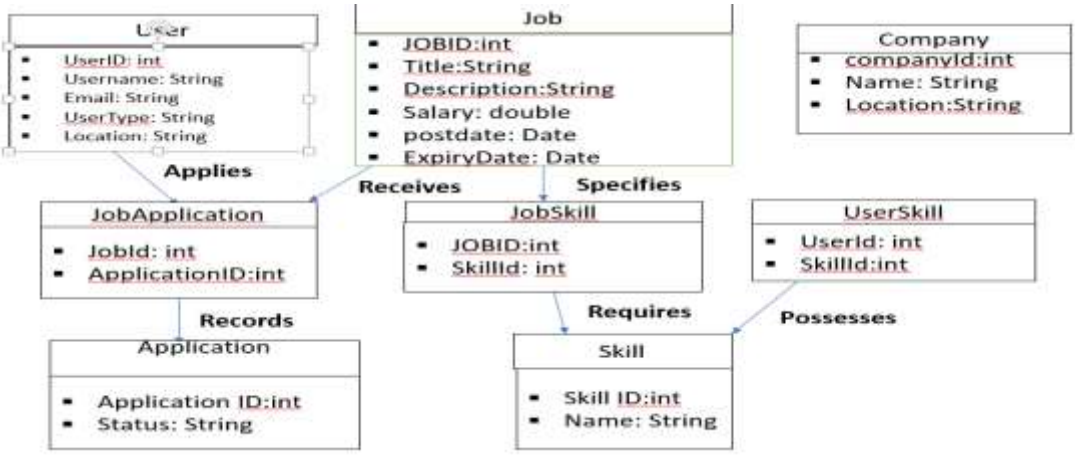
The diagram depicts a deep neural network, a type of machine learning algorithm that learns from a large amount of data. The data is fed into the network and goes through several layers of processing. With each layer, the network extracts increasingly complex features from the data. Finally, the network outputs a prediction, such as a classification (e.g., spam or not spam) or a value (e.g., image recognition)





DFD DIAGRAM

The flowchart outlines a process for predicting suitable jobs for users. First, information on user skills and job descriptions is collected and formatted. Key characteristics are then extracted, and a portion of this data is used to train a machine learning model. This model learns to classify jobs based on user profiles. After being evaluated on a separate dataset, the trained model can then predict suitable job openings for new users based on their qualifications.

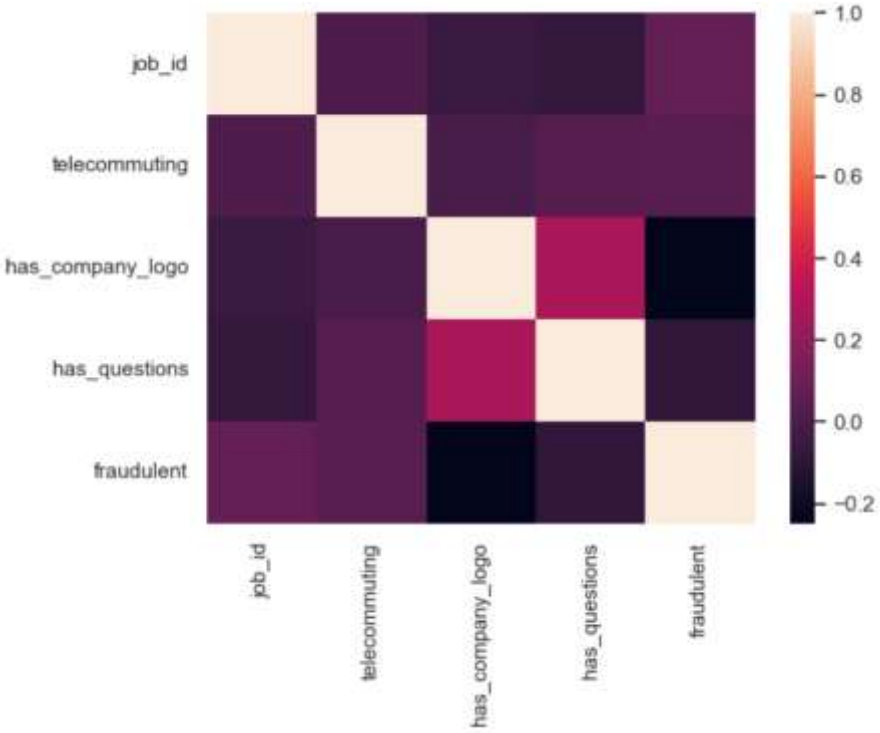


Class Diagram

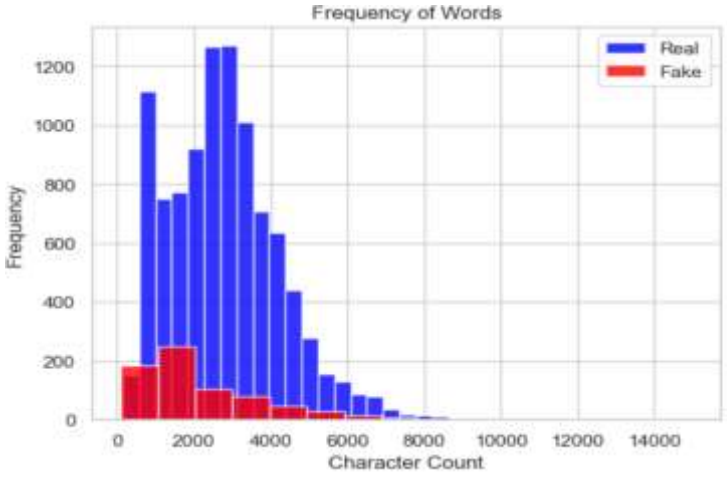
The image depicts an entity-relationship diagram (ERD) for a job listing system. It illustrates how different parts of the system relate to each other. The system tracks users (job seekers and employers), job listings, companies, skills, and applications.

Users can apply for jobs, and companies can post jobs and specify required skills. Users can also indicate their own skills. The diagram shows the relationships between these entities, allowing for data retrieval based on various criteria.

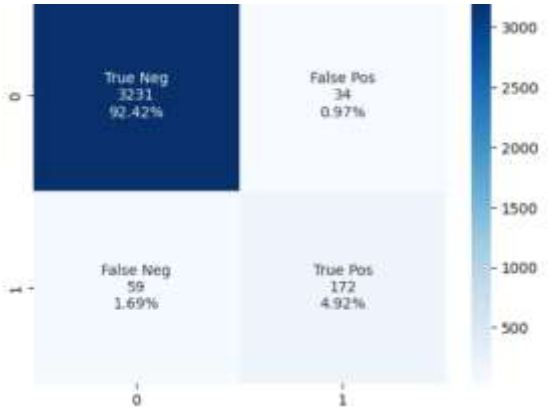
OUTPUTS :



Its an heatmap chart shows the percentage of job listings that contain various features. For instance, a high percentage of listings likely have a company logo, while a much smaller percentage might be marked as fraudulent.



The graph showing the frequency of words in a text. The blue bars represent the frequency of real words, while the red bars represent the frequency of fake words. The x-axis shows the character count, and the y-axis shows the frequency. In essence, this graph depicts how often real and fake words appear throughout a text, potentially helping identify artificial content.



The confusion matrix, a visualization tool used in machine learning to evaluate the performance of an algorithm. In this specific case, it seems to be evaluating a binary classification model that's designed to identify spam jobs.

CONCLUSION :

The development of the Fake Job Detection System represents a significant stride in mitigating the widespread issue of fraudulent job postings plaguing online recruitment platforms. Through the amalgamation of advanced technologies like machine learning, natural language processing, and data analytics, the system offers a robust and scalable solution for accurately identifying and flagging suspicious job listings. By systematically collecting, preprocessing, and analyzing job postings data, coupled with feature engineering techniques, the system effectively extracts meaningful insights to facilitate accurate classification of fake job postings. The integration of user-friendly interfaces and seamless deployment mechanisms ensures accessibility and usability, empowering both job platforms and users to proactively combat deceptive practices and uphold the integrity of the online job market. As the system continues to evolve and adapt to emerging threats, it holds promise in fostering a safer and more trustworthy environment for job seekers and employers alike.

REFERENCES:

1. Yeole, S., & Deore, R. (2020). "Fake Job Postings Detection using Machine Learning Techniques." International Journal of Innovative Technology and Exploring Engineering, 9(1), 2454-7662.
2. Kumar, V., & Bansal, A. (2020). "Detection of Fake Job Postings using Machine Learning Techniques." International Journal of Engineering Research & Technology, 9(2), 2278-0181.
3. Gupta, A., & Sharma, S. (2019). "Fake Job Detection Using Data Mining Techniques." International Journal of Computer Applications, 975(8887), 8887-8887.
4. Singh, A., & Bansal, A. (2019). "A Review on Fake Job Posting Detection." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 4(2), 2395-3527.
5. Khadka, R., Lee, J., & Jang, J. (2018). "An Empirical Study on Detection of Fake Job Advertisements." Procedia Computer Science, 131, 1163-1170.
6. Ganguly, A., Chakraborty, S., & Mukherjee, A. (2017). "A Text Mining Approach for Detection of Fake Job Advertisement." In Proceedings of the 2017 IEEE International Conference on Data Mining and Advanced Computing (SAPIENCE), 2017, 51-56.
7. Saha, S., & Manna, S. (2016). "Automatic Detection of Fake Job Advertisements: An NLP Perspective." In Proceedings of the 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, 1910-1914.
8. Htwe, T. T., et al. (2015). "Analysis of Job Description to Detect Fake Job Posting." In Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015, 1-6.
9. Rathore, M. M., et al. (2021). "Fake Job Posting Detection Using Hybrid Deep Learning Technique." IEEE Access, 9, 16183-16196.
10. Sharma, S., & Sharma, A. (2021). "Machine Learning Approaches for Detection of Fake Job Postings." In Proceedings of the 3rd International Conference on Data Science and Applications (ICDSA), 2021, 1-6.
11. Kim, J., & Kim, Y. (2020). "Detecting Fake Job Postings with Hierarchical Attention Networks and Web Crawler." Expert Systems with Applications, 156, 113543.
12. Choudhary, A., et al. (2020). "Fake Job Posting Detection using Supervised Machine Learning Techniques." In Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, 1-6.
13. Sharma, P., & Mehta, N. (2019). "A Study of Various Machine Learning Techniques for Fake Job Posting Detection." In Proceedings of the 3rd International Conference on Computing, Communication and Security (ICCCS), 2019, 1-5.
14. Jang, Y., & Lee, J. (2019). "Fake Job Posting Detection based on Machine Learning." In Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2019, 288-290.
15. Pradhan, S. R., et al. (2018). "Detecting Fake Job Offers using Machine Learning Techniques." In Proceedings of the 2018 International Conference on Advances in Computing, Communication, & Automation (ICACCA), 2018, 1-5.

16. Subramani, S., & Yuvaraj, S. (2017). "Detecting Fake Job Advertisement using Machine Learning Algorithms." International Journal of Engineering and Computer Science, 6(11), 2305-2375.
17. Arora, R., et al. (2016). "A Machine Learning Approach for Identifying Fake Job Postings." In Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT), 2016, 1-5.