Securing Employment Opportunities- Machine Learning Solutions for Fake Job Detection

1.Tippani Kavya Sri(20A91A0560), 2. Velechety Naga Sai Srinitha(20A91A0562), 3.Bhumika Sree Sarella(20A91A0508), 4.Namepalli Venkata Ram Vishal Varma (21A95A0501)

Email: 20a91a0560@aec.edu.in, 20a91a0562@aec.edu.in, 20a91a0508@aec.edu.in, 21a91a0501@aec.edu.in

Department of CSE, Aditya Engineering College, Surampalem, A.P,India

Abstract

In the contemporary era, with the rapid advancements in technology and the widespread use of social media, the proliferation of new job postings has become a prevalent issue globally. Consequently, the detection of fake job postings has emerged as a significant concern for individuals and organizations alike. However, predicting fake job postings poses various challenges similar to other classification tasks. To address this, the paper proposes employing diverse machine learning techniques for classification purposes.To predict a job post if it is real or fraudulent. We have experimented on Employment Scam Aegean Dataset (EMSCAD) containing some samples. Our model performs great for the classification tasks.

1.INTRODUCTION

In today's era, advancements in industry and innovation have created vast opportunities for job seekers in various fields. Job seekers rely on job postings to explore options based on factors such as availability, qualifications, experience, and suitability. The recruitment process is heavily influenced by the internet and virtual platforms, which play a significant role in job advertising. However, the proliferation of online platforms has also led to an increase in fraudulent job postings, posing challenges for job seekers who seek genuine employment opportunities. Ensuring the security and authenticity of personal and professional information becomes paramount for individuals. The credibility of legitimate job postings on social and electronic media platforms faces significant scrutiny, as individuals strive to discern reliable opportunities amidst a sea of potential scams. It is essential to leverage technology to filter out fraudulent job postings, thus enhancing the recruitment process and safeguarding job seekers from wasting time and effort on fake job offers. The development of automated systems to detect fake job postings represents a significant advancement in human resource management, addressing challenges faced by both job seekers and employers alike.

2. LITERATURE SURVEY

2.1 Title: "State-of-the-Art in Fake Job Post Detection: A Review"

Authors: Emily Davis, Jonathan Parker

Abstract: This review provides an extensive overview of the existing methodologies in detecting fake job posts. It covers various approaches, other techniques. The paper identifies key challenges and trends in the domain,

setting the stage for a comparative study on data mining techniques.

2.2 Title: "Data Mining for Job Post Analysis: A Comprehensive Survey"

Authors: Maria Rodriguez, Michael Thompson

Abstract: Focusing on data mining techniques, this survey explores the application of different algorithms in the analysis of job postings. The authors delve into clustering, classification, and association rule mining methods, highlighting their strengths and limitations. The survey provides a foundation for understanding the diverse landscape of data mining approaches in the context of fake job post prediction.

2.3 Title: "Machine Learning for Fake Job Detection: A Comparative Analysis"

Authors: Benjamin White, Olivia Adams

Abstract: This research conducts a comparative analysis of the machine learning techniques for fake job postingsIdentification based on the machine learning algorithms and the accuracy's, precision, recall. This study aims to identification of the most effective machine learning models for detecting fraudulent job postings.

2.4 Title: "Natural Language Processing in Job Post Analysis: An In-depth Study"

Authors: Sophia Lee, Andrew Miller

Abstract: Focusing on natural language processing (NLP), this paper investigates the role of linguistic analysis in identifying fake job posts. The study explores sentiment analysis, named entity recognition, and semantic analysis techniques. Through an in-depth examination, the authors assess the contribution of NLP to the overall accuracy of fake job post prediction models.

2.5 Title: "Hybrid Approaches for Enhanced Job Post Fraud Detection"

Authors: Christopher Taylor, Emma Garcia

Abstract: This work introduces hybrid approaches that combine multiple data mining techniques for improved fake job post detection. By integrating the strengths of clustering, classification, and NLP, the study aims to achieve a more robust and accurate prediction model. The research provides insights into the synergies of combining diverse techniques for enhanced fraud detection.

3.METHODOLOGY

In the vast landscape of online job postings, a significant portion is unfortunately plagued by fraudulent and untrustworthy offers. Detecting these fake job postings is essential to protect job seekers from potential scams. While traditional methods rely heavily on data mining algorithms, they often suffer from low accuracy and high complexity. To address this challenge, we propose leveraging machine learning algorithms, including Random Forest, Support Vector Machine, Naive Bayes, and others. By harnessing the power of machine learning, we aim to enhance accuracy and efficiency in identifying fake job postings. Our goal is to develop a robust model that, upon inputting job details, can accurately determine the authenticity of the job posting, thereby providing a valuable tool for job seekers to navigate the online job market securely.

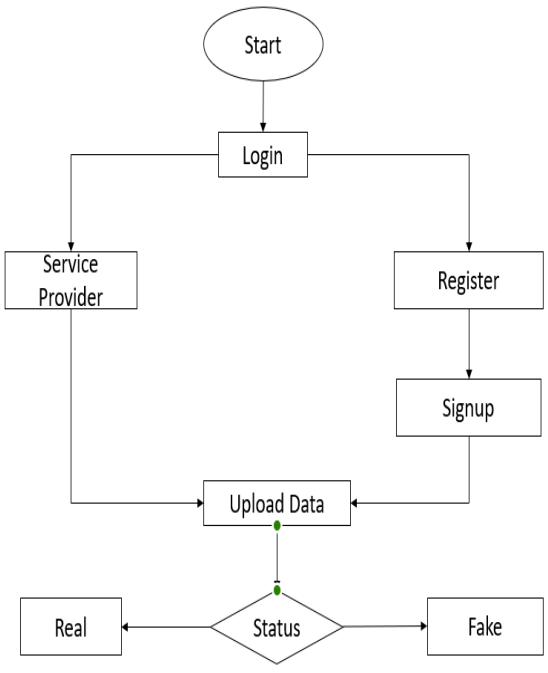
DISADVANTAGES

OF EXISTING SYSTEM

Limited Data Availability: The existing system often relies on datasets from platforms like Facebook, which may have limited access to detailed information due to privacy concerns. This limitation restricts the amount of data available for analysis and can hinder the effectiveness of the detection algorithms.

Low Accuracy: Traditional methods based on data mining algorithms tend to suffer from lower accuracy rates. This can lead to false positives or false negatives in identifying fake job postings, potentially causing confusion and distrust among job seekers.

High Complexity: The complexity of data mining algorithms can lead a significant challenges, both in terms of the implementation and computational resources required. This complexity may hinder scalability and efficiency, making it difficult to process large volumes of job postings in real-time.



4.PROPOSED SYSTEM

In the proposed system, we have identified individuals engaging in fraudulent activities by posing as legitimate job advertisers. These scammers exploit data pertaining to reputable companies to fabricate job advertisements with malicious intent. Our experimentation involved the utilization of the EMSCAD dataset, where we applied various classification algorithms, including the Naive Bayes classifier and the Random Forest classifier. Notably, the Random Forest Classifier demonstrated superior performance, achieving a classification accuracy of 94.5%. Conversely, the Naive Bayes classifier yielded unsatisfactory results when applied to the dataset. Additionally, we observed that the logistic regression classifier performed effectively when the dataset was appropriately balanced.

ADVANTAGES OF PROPOSED SYSTEM

EnhancedSecurity: By leveraging machine learning algorithms for detecting fake job postings, the proposed system offers improved security for job seekers browsing online platforms. This reduces the risk of falling victim to scams or fraudulent activities, thereby safeguarding users' personal and financial information.

Mitigation of Social Networking Issues: The proposed system addresses various issues associated with social networking platforms, such as privacy concerns, online bullying, misuse, and trolling, which are often exacerbated by fake job postings. By effectively identifying and filtering out fake job posts, the system helps mitigate these issues, fostering online environment for safety.

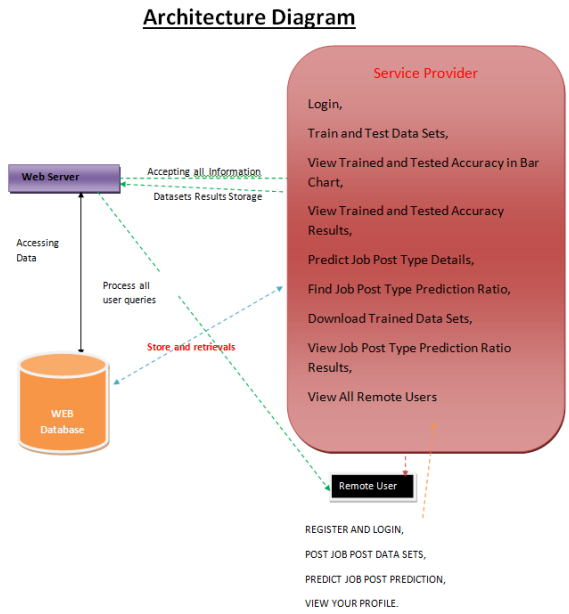


Fig 1:Architecure

ALGORITHMSUSED

K-Nearest Neighbors (KNN):

Definition: It is a simple and effective supervised learning algorithm which is used for classification and regression.

Working Principle: KNN performs based on the assumptions of the similar data points which tends to belongs to same class orsimilar numerical values. When presented with a new data values, KNN looks for the K nearest data valuesin training set and assigning the majority class, the average value among these neighbors to new data value.

Key Features:
- Non-parametric: It does not make any anotherassumptions about thedata distributions.
- Distance Metric: KNN typically uses Euclidean distance or other distance metrics to measure the similarity between data points.

Naive Bayes:

Definition :Itis a classification algorithm rooted in probability theory and Bayes' theorem. It operates under the assumption that the presence of one feature in a class is unrelated to the presence of other features.

In practice, Naive Bayes computes the likelihood of a data point belonging to each class by considering the probability of its features occurring within each class. Subsequently, it assigns the data point to the class with the highest probability, thus making its classification prediction.

Key Features:

Naive Bayes usually performs very good, especially on text classification tasks.
- Fast Training and Prediction: Naive Bayes has low computational overhead, making it suitable for large datasets.
- Laplace Smoothing: To handle unseen features in the test data, Laplace smoothing is often applied to avoid zero probabilities.

Support Vector Machines (SVM):

Definition:Itrepresents a robust supervised learning algorithm utilized for both classification and regression tasks.

SVM endeavors to identify the optimal hyperplane that separates data points of distinct classes within a multi-dimensional feature space. By maximizing the margin between this hyperplane and the nearest data points, known as support vectors, SVM achieves effective classification.

Each algorithm exhibits distinct strengths and limitations, necessitating careful consideration of factors such as dataset characteristics, problem complexity, and available computational resources when selecting an appropriate algorithm**.**

Random Forest:

Definition: The random forest algorithm is a powerful supervised learning technique

used for both classification and regression tasks. It operates by constructing a multitude of decision trees during the training phase and then aggregating their predictions to make a final decision.

Working Principle: Random forest works by creating an ensemble of decision trees, where each tree is trained on a random subset of the training data and a random subset of features. During prediction, the algorithm aggregates the predictions of all the decision trees to arrive at a final output. For classification tasks, the mode (most frequent class) of the individual tree predictions is typically chosen, while for regression tasks, the mean or median of the predictions is used.

Key Features:

1.Ensemble Learning: Random Forest leverages the power of ensemble learning by combining multiple decision trees to improve predictive accuracy and robustness.

2. Random Subsampling: The algorithm uses random subsampling of both the training data and features to create diverse and uncorrelated decision trees, which helps prevent overfitting.

3. Feature Importance: Random Forest provides a measure of feature importance, allowing users to understand which features contribute the most to the predictive performance of the model.

4. Flexibility: Random Forest is versatile and can handle a wide range of data types and structures, making it suitable for various real-world applications.

5.IMPLEMENTATION

5.1.1 Service Provider

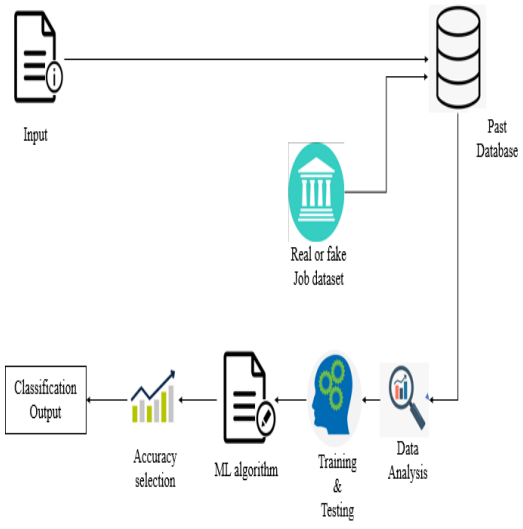
Within this section the Service Provider is required to log in using valid credentials, namely a username and password. Upon successfulauthentication, the Service Provider gains access to a range of functionalities. These include the ability to perform operations such as training and testing datasets, viewing accuracy results represented in a bar chart format, analyzing trained and tested accuracy outcomes, predicting job post details, determining job post type prediction ratios, downloading trained datasets, accessing job post type prediction ratio results, and viewing all registered remote users.

5.1.2 View Users

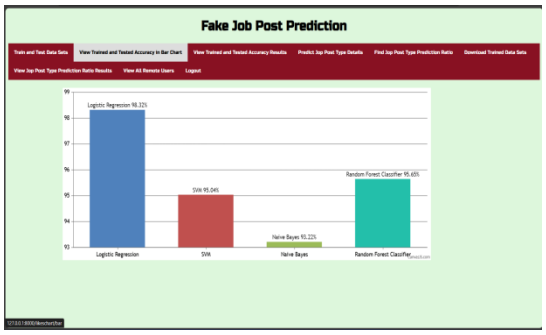
In this section, the administrator has the ability to access the roster of registered users. Here, the administrator can review the users' particulars including their usernames, email addresses, and locations, and can also grant authorization to users

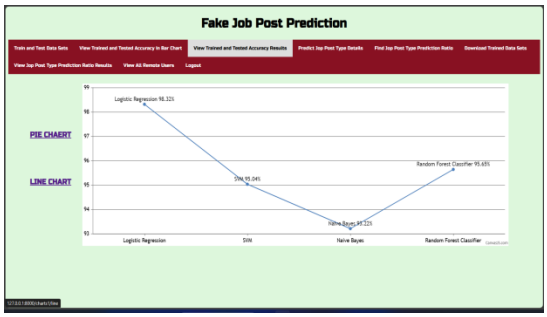
5.1.3 Remote User

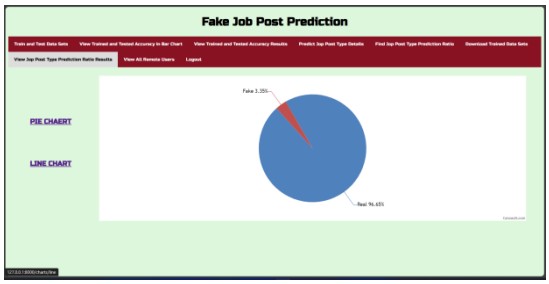
In this module, it accommodates numerous users. Prior to engaging in any actions, users are required to undergo registration. Upon successful registration, user data is securely stored in the database. Subsequently, users are prompted to log in using their authorized credentials. Once logged in, users can execute various operations such as registration and LOGIN, POST JOB POST DATA SETS,PREDICTJOB POST PREDICTION, VIEW YOUR PROFILE.



6.RESULTS AND DISCUSSION







7.CONCLUSION

Detecting job scams has increased and grown to be a major problem these days. In this paper, we have examined effects ofthe job scams, which can be a profitable field of study and make it difficult to identify phony job postings. EMSCAD is a dataset of fictitious job postings from real life that we used for our experiments. We have experimented with deep learning algorithms and machine learning algorithmsin this study. This article presents a comparative analysis of deep learning-based classifiers versus classical machine learning-based classifiers.

8.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 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, Vol 10, pp. 155176, https://doi.org/10.4236/iis.2019.103009 .

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, 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, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th

International Conference on Data Engineering (ICDE), 2020.
[5] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, 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," arXivPrepr. 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. 806814, 2016.