JOB SCAM ALERT

Dr. Y. Jeevan Nagendra Kumar
Head of the Department of Information Technology,
Gokaraju Rangaraju
Institute of Engineering and Technology,
JNTUH, Hyderabad, India
jeevannagendra@griet.ac.in , 9010180199

J. Krishna Chaitanya, C.Abhinav Ready, Vinod Pawar
Department of Information Technology Gokaraju
Rangaraju Institute of Engineering and
Technology, Hyderabad, India.

Abstract:

In response to the increase in online fraud, we created the Fraud Detection Project, a solution that uses machine learning. Oursoftware uses random forest and SVM models to detect fake job postings, protect candidates' personal information, and improve theintegrity of online job postings. This technology can help people avoid financial scams and make informed decisions when applying forjobs online. Through Streamlit referrals, users can check the accuracy of job vacancies. We are committed to continuousimprovement to improve the performance of our software, transform the digital business process and ensure security for all partiesinvolved.

Keywords: Job scam, machine learning, random forest, support vector machine (SVM), fraud detection, Streamlit deployment.

INTRODUCTION

The incidence of employment scams has been increasing, with reports showing a significant rise between 2017 and 2018. The economic impact of the coronavirus pandemic has further exacerbated the issue, leading to high unemployment and creating opportunities for scammers. These scammers exploit job seekers by offering fraudulent job opportunities to steal personal information or solicit money. As a university student, I have encountered numerous scam emails promising lucrative jobs that were fraudulent. Addressing this issue is crucial, and advanced machine learning techniques, combined with natural language processing (NLP), offer a promising solution.

To tackle this problem, we developed Job Scam Detection, a tool employing advanced machine learning algorithms. Using data from Kaggle, our project focuses on identifying fraudulent job postings with Random Forest and Support Vector Machine (SVM) models. The system saves and deploys trained models using Streamlit, allowing users to verify the authenticity of job postings easily.

The project follows five stages: defining the problem, collecting data, cleaning and preprocessing data, modeling, and evaluation. Our primary objective is to develop a classifier that accurately differentiates between real and fake job postings, integrating both numeric and textual features. Through this approach, Job Scam Detection aims to provide a reliable tool for job seekers, helping them navigate the job market safely and make informed decisions.

LITERATURE SURVEY

[1] Kumari and Sahani address the rise in data breaches and fake job postings, often through digital job websites. Their project aims to use ML to predict the likelihood of a job being fake, helping candidates stay alert. Their model uses NLP to analyze sentiments and patterns in job postings, trained as a Sequential Neural Network with the GloVe algorithm. They tested the model on LinkedIn postings and improved it for robustness and realism.

[2] According to the project by Smith et al., ML models such as Naive Bayes and Decision Trees are used to predict fake job postings. By integrating topic modeling techniques like Latent Dirichlet Allocation (LDA), the study enhances classification accuracy and provides real-time detection of fraudulent listings, protecting job seekers from scams.

[3] In the work by Johnson et al., text analysis is leveraged to detect fake job postings using a highly imbalanced dataset, where only 5% of postings are fraudulent. Initial models, including a Recurrent Neural Network (RNN) with Embedding and Bidirectional Long Short-Term Memory (LSTM) layers, showed promising results.

Comparison with a Small-BERT model significantly improved detection accuracy, highlighting the effectiveness of advanced NLP models in distinguishing real from fraudulent job postings.

[4] Habiba and Islam focus on predicting fake job postings using various data mining techniques and classification algorithms, such as KNN, decision tree, support vector machine, naive Bayes, random forest, multilayer perceptron, and deep neural networks. They experimented on the Employment Scam Aegean Dataset (EMSCAD) with 18,000 samples. Their deep neural network classifier, with three dense layers, achieved approximately 98% accuracy in predicting fraudulent job posts [6].

[5] Kumari and Satya Kala highlight the increase in employment fraud, with a significant rise in job scams in 2018 compared to 2017. They note the impact of high unemployment and the coronavirus pandemic, which have created opportunities for fraudsters to exploit job seekers. The study emphasizes the use of NLP and machine learning to address these fraudulent activities and protect personal information from scammers.

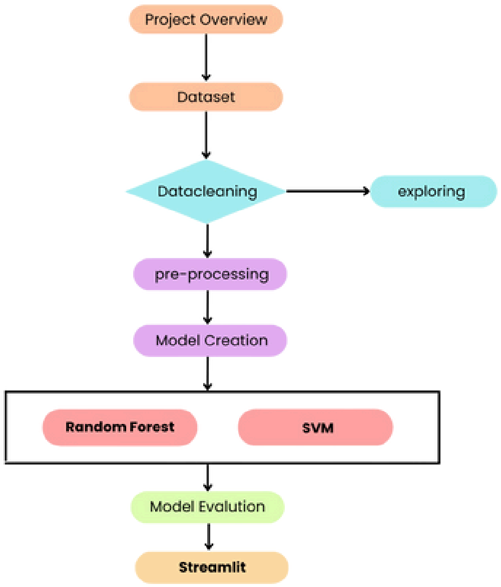
[6] Gulshan, Mukund, and Ajay A. address the rise in online job postings during the pandemic and the need for accurate fake job detection. Their research employs various data mining and classification algorithms, including KNN, decision tree, support vector machine, naive Bayes, random forest, multilayer perceptron, and deep neural networks, using the EMSCAD dataset with 18,000 samples. Their deep neural network classifier, with three dense layers, achieved approximately 98% accuracy in predicting fraudulent job posts.

MOTIVATION

The rise in employment scams has paralleled increasing global economic uncertainty, especially due to the coronavirus pandemic. With job losses and high unemployment, individuals are more vulnerable to fraudulent job offers. As university students, we have encountered deceptive job postings that appeared legitimate but were scams. Addressing this issue requires innovative approaches, using advanced machine learning techniques like Natural Language Processing (NLP) to distinguish between authentic and fraudulent job postings. This project's motivation is to protect job seekers from scams by providing a reliable, technology-driven solution for a safer job search experience.

METHODOLOGY

This section outlines the workflow for predicting fake job postings using machine learning and NLP. The process includes problem definition, data collection, cleaning and preprocessing, modeling, and evaluation, as shown in Figure 1.



Our goal is to develop a robust classifier that identifies fraudulent job postings using Random Forest and SVM classifiers. We process a dataset from Kaggle, integrating numeric and textual features for a comprehensive analysis. The final model is deployed with Streamlit for real-time predictions. This structured approach ensures the model is accurate and practical for real-world use.

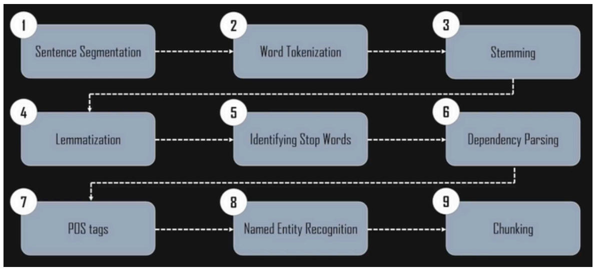
Natural Language Processing (NLP)

Natural Language Processing (NLP) was used to analyze and extract meaningful information from job postings. NLP transforms raw text into a structured format, enabling machine learning models to classify job postings as real or fake. Combining computational linguistics with advanced statistical models, machine learning, and deep learning, NLP allows computers to understand context and grasp the full meaning of the text, including the writer's intentions and emotions.

In this project, we built an NLP pipeline with the following steps:

- **Sentence Segmentation**: Dividing text into sentences.
- **Word Tokenization**: Breaking sentences into words or tokens.
- **Stemming and Lemmatization**: Reducing words to their base forms.
- **Identifying Stop Words**: Removing common, insignificant words.
- **Dependency Parsing**: Analyzing grammatical structures and word relationships.

- **Part-of-Speech (POS) Tagging**: Assigning parts of speech to each word.
- **Named Entity Recognition (NER)**: Identifying and classifying proper nouns.
- **Chunking**: Grouping words into meaningful chunks.



By applying these NLP techniques, we converted unstructured job posting text into structured numerical data for our machine learning models. This preprocessing step is essential for feature extraction and ensures the text data is suitable for machine learning algorithms.

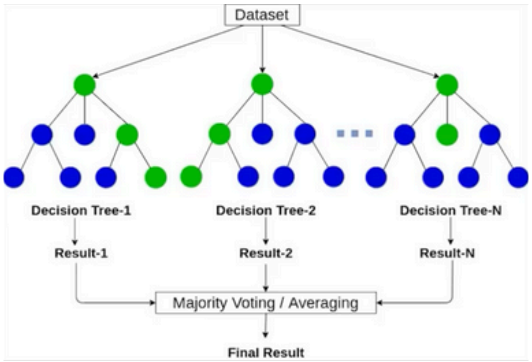
We used Random Forest and Support Vector Machine (SVM) classifiers to train predictive models on a dataset containing both genuine and fake job postings. Leveraging NLP techniques enabled accurate identification and classification of fake job postings, surpassing manual and rule-based systems.

Random Forest

Random Forest is a versatile machine learning algorithm used for classification and regression tasks. It constructs multiple decision trees during training, each tree using a random subset of the dataset and features at each split. This approach reduces overfitting and enhances predictive accuracy.

Application in Our Project

- Dataset Preparation: We preprocess the job postings dataset using NLP techniques to create structured numerical data.
- Decision Tree Construction: Multiple decision trees are built during training, each using a random subset of the dataset and features.
- Aggregation of Predictions: Predictions from each tree are combined via majority voting to classify job postings as real or fake.
- Overfitting Reduction: By averaging predictions across trees, Random Forest mitigates overfitting, ensuring stable and accurate results.
- Handling Complexity: Effective with complex data structures, Random Forest manages numerous features extracted from job postings, ensuring robust classifications.

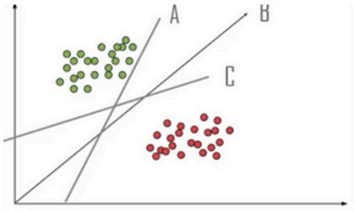


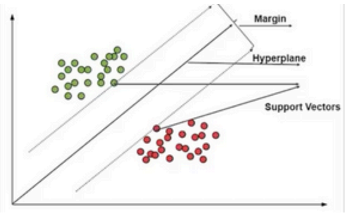
Support Vector Machine (SVM)

Support Vector Machine (SVM) is a robust machine learning algorithm used for classification, regression, and outlier detection. It excels in handling high-dimensional and nonlinear data, making it effective in various applications such as text and image classification, spam detection, and anomaly detection.

In job scam detection, SVM plays several key roles:

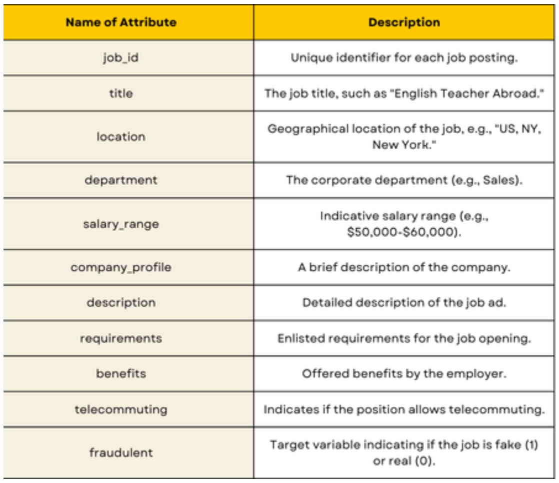
- Optimal Hyperplane: SVM identifies the hyperplane that maximizes the margin between different classes, crucial for distinguishing between real and fraudulent job postings.
- Kernel Trick: SVM employs kernel functions (e.g., Polynomial, Gaussian RBF) to manage nonlinear decision boundaries, capturing complex patterns in job posting data.
- Classification Accuracy: SVM classifies job postings by converting unstructured text into numerical features using NLP preprocessing techniques.
- Margin Maximization: By maximizing the margin, SVM reduces overfitting and enhances generalization, improving the reliability of predictions.
- Real-world Applicability: SVM's ability to handle high-dimensional and nonlinear data makes it effective for identifying legitimate versus fraudulent job offers, protecting job seekers from scams.





DATASET DESCRIPTION

The dataset for this project includes 17,880 job descriptions, with around 800 labeled as fraudulent. It features textual and meta-information, such as job ID, title, location, department, salary range, company profile, job description, requirements, benefits, and telecommuting options. The target variable indicates whether a job is fraudulent or real. Sourced from Kaggle, this dataset is valuable for developing classification models to detect fake job postings, identifying features indicative of fraud, and running exploratory data analysis.



Name of Attribute	Description
job_id	Unique identifier for each job posting.
title	The job title, such as "English Teacher Abroad."
location	Geographical location of the job, e.g., "US, NY, New York."
department	The corporate department (e.g., Sales).
salary_range	Indicative salary range (e.g., $50,000-$60,000).
company_profile	A brief description of the company.
description	Detailed description of the job ad.
requirements	Enlisted requirements for the job opening.
benefits	Offered benefits by the employer.
telecommuting	Indicates if the position allows telecommuting.
fraudulent	Target variable indicating if the job is fake (1) or real (0).

PROPOSED SYSTEM

The proposed system uses machine learning and natural language processing (NLP) to detect fake job postings. It aims to provide a scalable and accurate solution, overcoming limitations of existing systems.
Key Components:

1. Data Collection: Gather job posting data from multiple sources.
2. Data Preprocessing: Clean data by handling missing values and performing text preprocessing (tokenization, stopword removal, lemmatization).
3. Feature Engineering: Develop new features by combining text fields and generating character counts.
4. Model Training: Train models such as Naive Bayes, SGD Classifier, and LSTM on the preprocessed data.
5. Model Evaluation: Assess models using accuracy and F1-score metrics.
6. Final Classification: Aggregate model outputs to determine job posting authenticity.

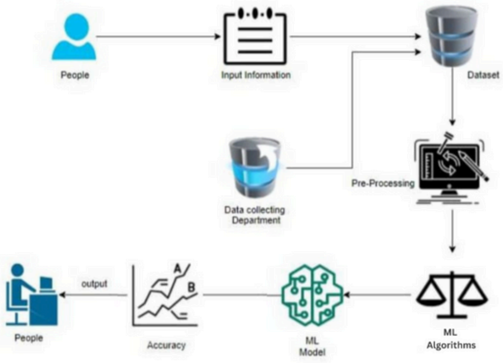
EXISTING SYSTEM

The current system for detecting fake job postings relies on manual review or basic rule-based filtering. Human reviewers assess job postings for signs of fraud, such as discrepancies in details, unusual salary ranges, or suspicious email addresses. However, this approach has several limitations:

- Scalability Issues: Manual review becomes impractical as the volume of job postings grows.
- Subjectivity and Human Error: Accuracy depends on the reviewer's expertise, leading to potential errors.
- Slow Response Time: Manual processes are time-consuming, causing delays in identifying and removing fake postings.
- Limited Adaptability: Rule-based systems are static and cannot quickly adapt to new or evolving scams.

SYSTEM ARCHITECTURE

The proposed system architecture for detecting fake job postings comprises several essential stages: data collection, preprocessing, model construction, and performance evaluation. Initially, the Data Collection Module gathers job postings from various sources. The Data Preprocessing Module cleans and prepares the data through tokenization, stopword removal, and normalization. During the Model Construction Stage, Random Forest and Support Vector Machine (SVM) algorithms are utilized to build classification models. The Performance Evaluation Module assesses these models using metrics like accuracy and F1-score. The Final Classification Module then integrates the model outputs to classify job postings as genuine or fraudulent. This streamlined process ensures a robust and scalable solution for detecting fraudulent job postings.



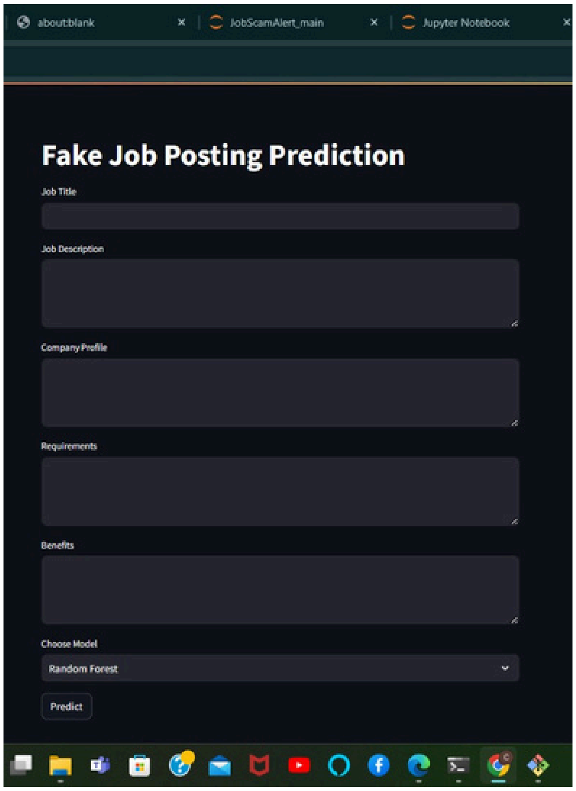
During the Model Construction Stage, Random Forest and Support Vector Machine (SVM) algorithms are used to build the classification models. The Model Training process involves applying these algorithms to the preprocessed data. The Model Evaluation Module assesses the models' performance using metrics such as accuracy and F1-score, ensuring their effectiveness in distinguishing between genuine and fraudulent job postings. The Final Classification Module then integrates the outputs from Random Forest and SVM to classify job postings accurately.

Sample Data for the Problem Statement: The dataset used in this project includes job postings from a reliable source and features essential for classification, such as job title, description, and salary range. The data preparation process involves TF-IDF Vectorization for converting text into numerical features suitable for model training. Data plotting and word cloud visualization help in understanding data distributions and key terms.

RESULTS

Sample Data for the Problem Statement: The dataset used in this project includes job postings from a reliable source and features essential for classification, such as job title, description, and salary range. The data preparation process involves TF-IDF Vectorization for converting text into numerical features suitable for model training. Data plotting and word cloud visualization help in understanding data distributions and key terms.

Streamlit Integration involves several important components. The process begins with Loading Trained Models, where pre-trained Random Forest and Support Vector Machine (SVM) models are imported into the Streamlit application using libraries such as joblib. This ensures that the models, which have been trained on extensive datasets, are ready to provide accurate predictions upon user input.



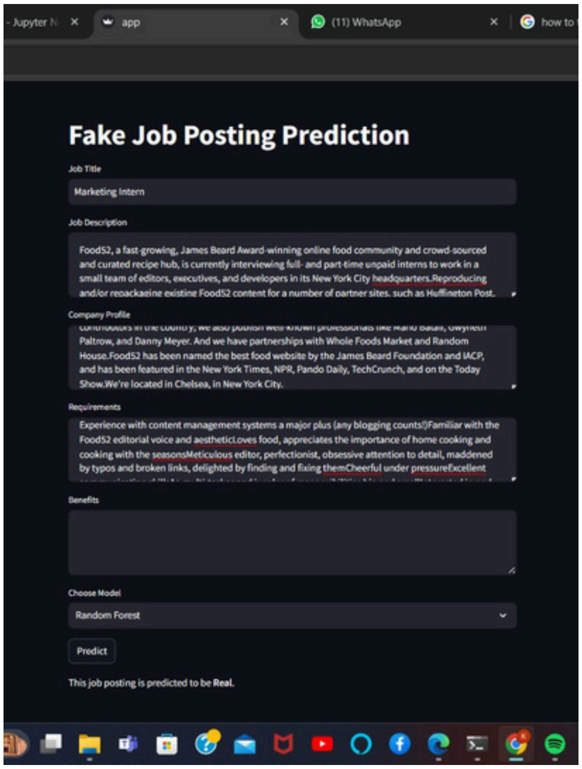
Fake Job Posting Prediction

Choose Model: Random Forest

The Streamlit Application Setup features a user-centric interface. It includes an input field for users to enter job descriptions and a predict button to initiate the classification process. The results from both models are displayed in output sections, allowing users to see whether the job posting is classified as fake or real. This setup ensures that users receive immediate feedback based on the machine learning models' analysis.

Deployment and Usage of the Streamlit application offer several advantages. The application can be hosted on a web server or run locally, making it accessible from any device with internet connectivity. Users interact with the application by entering job details and receiving instant predictions, facilitating quick decision-making about job postings. The integration of Random Forest and SVM into the Streamlit application ensures that the predictions are both accurate and reliable, leveraging the strengths of each model.

Additionally, the Streamlit Integration allows for ongoing enhancements. The feedback mechanism enables users to provide input on the predictions, which can be used to fine-tune the models and improve accuracy over time. This dynamic interaction with users ensures that the application remains relevant and effective in detecting fraudulent job postings.



Fake Job Posting Prediction

Job Title: Marketing Intern

Job Description: Food52, a fast-growing, James Beard Award-winning online food community and crowd-sourced and curated recipe hub, is currently interviewing full- and part-time unpaid interns to work in a small team of editors, executives, and developers in its New York City headquarters.Reproducing and/or repackaging existing Food52 content for a number of partner sites, such as Huffington Post,

Company Profile: contributors in the country; we also publish well-known professionals like Mario Batali, Gwyneth Paltrow, and Danny Meyer. And we have partnerships with Whole Foods Market and Random House.Food52 has been named the best food website by the James Beard Foundation and IACP, and has been featured in the New York Times, NPR, Pando Daily, TechCrunch, and on the Today Show.We're located in Chelsea, in New York City.

Requirements: Experience with content management systems a major plus (any blogging counts!)Familiar with the Food52 editorial voice and aestheticLoves food, appreciates the importance of home cooking and cooking with the seasonsMeticulous editor, perfectionist, obsessive attention to detail, maddened by typos and broken links, delighted by finding and fixing themCheerful under pressureExcellent

Benefits:

Choose Model: Random Forest

This job posting is predicted to be **Real**.

Overall, the Streamlit-based system offers a robust and practical solution for job posting classification, blending advanced machine learning techniques with an accessible and efficient user interface.

CONCLUSION

In conclusion, this project successfully developed a machine learning-based system for detecting fraudulent job postings by utilizing a combination of text and numeric data. The project employed Random Forest and Support Vector Machine (SVM) models, showcasing the effectiveness of machine learning techniques in combating employment scams. The Random Forest model demonstrated superior performance compared to the SVM model, proving to be more effective in distinguishing between genuine and fake job postings. This success was further enhanced by the integration of Streamlit, which provided an interactive platform for users to receive real-time predictions. Overall, the project offers a practical solution for identifying fraudulent job postings and protecting job seekers, laying a solid foundation for future developments in this area.

FUTURE ENHANCEMENTS

Despite the project's success, there are several areas for future enhancement. One key improvement could involve incorporating more advanced natural language processing (NLP) techniques. Models such as transformers and BERT could enhance detection accuracy by better capturing nuanced patterns and understanding the context within job descriptions.

Expanding the dataset to include a broader and more diverse range of job postings is another important step. A larger dataset would improve the model's ability to generalize across various types and sources of job listings, making it more robust.

Additionally, implementing real-time detection capabilities would allow the system to continuously adapt to new and evolving scam patterns. This could be achieved through automated updates and model retraining based on new data and user feedback, ensuring the system remains effective over time.

Exploring hybrid models that combine multiple machine learning techniques or integrating ensemble methods could also offer improvements in classification accuracy. Expanding the system’s reach to cover different languages and regions could further enhance its global applicability and utility.

Overall, these enhancements aim to refine the system's performance, adaptability, and scope, ensuring it remains a valuable tool in the ongoing fight against employment scams.

REFERENCES

[1] Journal of Engineering Sciences, K.V. Jhansi Rani, Priyanka Rompalli, Leela Nagababu Kathi, Abhiram Tholam, Ramakrishna Gudla, B.Tech, Department of CSE, Eluru College of Engineering and Technology, Duggirala, Andhra Pradesh-534004.

[2] www.ijcrt.org © 2024 IJCRT | Volume 12, Issue 4 April 2024 | ISSN: 2320-2882 Mr. B.J.M. Ravi Kumar, Mr. Melkamu Boka Eba, Mr. Ridwan Mohammed Department of Information Technology and Computer Application, Engineering College A, Andhra University, Visakhapatnam, India - 530003

[3] e-ISSN: 2582-5208 Aravind Sasidharan Pillai*1 *1The University of Illinois, Urbana-Champaign, IL, USA. DOI: https://www.doi.org/10.56726/IRJMETS35202

[4] 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques Sultana Umme Habiba, Md. Khairul Islam, Farzana Tasnim Dept. of CSE, International Islamic University Chittagong, Dhaka, Bangladesh

[5] https://economictimes.indiatimes.com/jobs/hr-policies-trends/job-scams-on-the-rise-as-fraudsters-target-desperate-jobseekers/increasing-job-scams/slideshow/101636401.cms?from=mdr

[6] International Journal for Multidisciplinary Research (IJFMR) Maddi Sravya Reddy1, Maddikera Hemanth Lal2, Lingam Sainad3, Sandeep Agarwalla4 Student, Department of Computer Science and Engineering, Malla Reddy College of Engineering and Technology, Hyderabad

[7] https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

[8] https://images.app.goo.gl/ajF9fwXCZPHBJfFx5

[9] https://images.app.goo.gl/9hbKy6Pap3QNtC1T6

AUTHOR PROFILE

Dr. Y. Jeevan Nagendra Kumar, obtained his Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, AP in 2017 and MTech Computer Science Technology from Andhra University in 2005. He is working as Professor and Dean - Technology and Innovation Cell in GRIET since 2005.

He has about 16 Research Papers in International / National Conferences and Journals and also attended many FDP Programs to enhance his knowledge. With his technical knowledge he guided the students in developing the useful Web applications and data mining related products. As B O S member was able to introduce new subjects, topics in UG / PG Courses. Students are encouraged to work on research projects, engineering projects as well as for industrial training.
He was acted as Coordinator for 3 International Conferences and Technical Committee member for several International Conferences. He is Coordinator for J Lab under J Hub JNTUH and Robotic Club. Also, Coordinator for NBA and NAAC at College Level.

Currently acting as Convener and Vice-President for MHRD IIC
(Institution's Innovation Cell) GRIET.