**JOB SCAM ALERT**

Dr. Y J Nagendra Kumar1, J. Krishna Chaitanya, C. Abhinav Reddy 3, Vinod Pawar 4

*1Head of the Department Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad.*

*2 Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad 2,3,4*

a) *chaitanyajuluri11@gmail.com*

b) *chilkuriabhinavr@gmail.com*

c) *vinodpawar4069@gmail.com*

**Abstract:** In response to the rise in online job scams, we developed Job Scam Detection, an innovative solution utilizing advanced machine learning techniques. Our software leverages Random Forest and Support Vector Machine (SVM) models to identify fraudulent job postings. This technology enables job seekers to safely navigate the digital job market, protecting them from scams and safeguarding their personal information. By improving the integrity of online job listings, Job Scam Detection fosters a more secure and trustworthy environment for both job seekers and employers. Our solution not only helps individuals avoid financial scams, such as fraudulent application fees and false promises of employment, but also empowers them to make informed decisions when applying for jobs online. The model has been saved and deployed using Streamlit, allowing users to easily check whether a job posting is fake or real. We are committed to continuously refining and optimizing our software to enhance its effectiveness, aiming to revolutionize how job seekers engage with the digital job market and providing them with confidence and peace of mind in their job search. Through ongoing research and development, we strive to stay ahead of emerging threats and ensure a seamless and secure experience for all stakeholders involved.

# Introduction

The incidence of employment scams has been steadily increasing, with CNBC reporting that the number of such scams doubled from 2017 to 2018. The current economic climate, exacerbated by the coronavirus pandemic, has led to widespread job losses and high unemployment rates, creating fertile ground for scammers. These scammers prey on desperate job seekers by luring them with enticing but fraudulent job offers, aiming to extract sensitive personal information such as addresses, bank account details, and social security numbers, or to solicit money under the guise of application fees or investments. As a university student, I have personally received numerous scam emails offering lucrative job opportunities that turned out to be 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 sophisticated tool employing advanced machine learning algorithms. Our project utilizes data from Kaggle, which includes a mix of genuine and fake job postings. Despite the small proportion of fraudulent postings, their identification is essential to protect job seekers. Our solution employs a variety of machine learning models, focusing on Random Forest classifiers and Support Vector Machines (SVM). The Random Forest model achieved an impressive accuracy of 97.22%, while the SVM model attained a 95% accuracy. The system is designed to save the trained models and deploy them using Streamlit, enabling users to easily determine whether a job posting is real or fake.

The project is structured into five stages: defining the problem, collecting data, cleaning and preprocessing data, modeling, and evaluation. The primary objective is to develop a classifier that can accurately differentiate between real and fake job postings. This classifier integrates both numeric and textual features to assess job postings comprehensively. 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.

**MOTIVATION**

The rise in employment scams has paralleled the increase in global economic uncertainty, particularly exacerbated by events like the coronavirus pandemic. With job losses mounting and unemployment rates soaring, individuals are increasingly vulnerable to fraudulent job offers that promise lucrative opportunities but ultimately aim to extract personal information or financial investments. As a university student, I have personally encountered deceptive job postings that initially appeared legitimate but later turned out to be scams. Addressing this issue requires innovative approaches, leveraging advanced Machine Learning techniques such as Natural Language Processing (NLP) to distinguish between authentic and fraudulent job postings. This project's motivation stems from the need to protect job seekers from falling victim to such scams by providing a reliable, technology-driven solution that ensures a safer job search experience.

# LITERATURE SURVEY

1. Employment Scams Are on the Rise This segment highlights the growing prevalence of employment scams exacerbated by economic uncertainties and the pandemic, emphasizing machine learning and NLP's role in protecting job seekers. [1]
2. Prediction of Fake Job Postings Using Machine Learning This project uses machine learning algorithms like Naive Bayes and Decision Trees, along with topic modelling (LDA), to predict fake job postings, aiming for real-time detection. [2]
3. Job Posting Fraud Detection Using NLP This research employs text analysis to detect fake job postings with an imbalanced dataset. Results from a Small-BERT model improved AUC scores significantly, demonstrating advanced NLP's effectiveness in fraud detection. [3]
4. Impact of Employment Scams: Defined by the FBI, employment scams extract personal information and money from victims. In 2020, over 16,000 reported losses exceeding $59 million. This project develops a classifier to identify fake listings and safeguard job seekers. [4]

**PROPOSED SYSTEM**

The proposed system leverages machine learning and natural language processing (NLP) techniques to automatically detect fake job postings. This system aims to address the limitations of the existing system by providing a scalable, accurate, and adaptive solution.:

**Machine Learning Model**:

1. **Data Collection:** Gather job posting data from various sources.
2. **Data Preprocessing**: Clean and preprocess the data, including handling missing values and text preprocessing (tokenization, stop word removal, lemmatization).
3. **Feature Engineering:** Create new features from the existing data, such as combining text fields and generating character counts.
4. **Model Training:** Train machine learning models (Naive Bayes, SGD Classifier and LSTM) on the preprocessed data.
5. **Model Evaluation:** Evaluate the models using accuracy and F1-score metrics.
6. **Final Classification:** Combine the outputs of the models to determine the authenticity of
7. job postings.

The proposed system integrates advanced machine learning and natural language processing (NLP) techniques to automate the detection of fraudulent job postings, addressing inherent limitations in current methods. Key enhancements and components of the system include:

Data Collection: Job posting data is gathered from diverse sources, providing a comprehensive dataset that encompasses various attributes such as job title, location, and company details.

Data Preprocessing: The collected data undergoes rigorous preprocessing steps to enhance quality and usability. This includes handling missing values, text preprocessing (tokenization, stopword removal, lemmatization), and ensuring data integrity.

Feature Engineering: New features are engineered from the processed data to enrich the dataset. Techniques such as combining text fields, generating character counts, and extracting relevant keywords are employed to enhance model performance.

Model Selection and Training: The project focuses on utilizing Random Forest and Support Vector Machine (SVM) algorithms due to their demonstrated efficacy in classification tasks. These models are trained on the preprocessed dataset to learn patterns and distinguish between genuine and fake job postings.

Model Evaluation and Validation: The trained models are evaluated using appropriate metrics to assess their performance accurately. The evaluation criteria include accuracy, precision, recall, and F1-score, ensuring robustness and reliability in detecting fraudulent job postings.

Integration with Streamlit: Post model training and validation, the best-performing model, particularly the Random Forest and SVM, is saved and integrated into a Streamlit-based application. This application provides a user-friendly interface where users can input job postings and receive real-time predictions on their authenticity. System Objectives:

● Identification of Genuine vs. Fake Job Postings: The primary goal is to accurately classify job postings, thereby enabling job seekers to focus exclusively on legitimate opportunities.

● Enhanced Data Handling: Leveraging a Kaggle dataset ensures comprehensive coverage of job attributes, facilitating thorough analysis and detection.

● Streamlined Data Processing: Rigorous preprocessing ensures that the input data is cleaned and optimized for predictive modeling, enhancing the accuracy and reliability of the system. By leveraging these advancements, the proposed system aims to revolutionize the detection of fake job postings, providing a scalable, accurate, and adaptive solution that addresses the evolving challenges in the digital job market.

# METHODOLOGY

This section outlines the workflow of the proposed methodology for predicting fake job postings using machine learning techniques and natural language processing (NLP). The workflow involves several stages: problem definition, data collection, data cleaning and preprocessing, modelling, and evaluation. Figure 1 illustrates the workflow of the proposed model.

A diagram of a process

Description automatically generated

By following this methodology, the project aims to develop a robust classifier that effectively identifies fraudulent job postings, leveraging the strengths of both Random Forest and SVM classifiers. The project employs a comprehensive approach to process the dataset from Kaggle, which includes genuine and fake job postings. The classifier integrates both numeric and textual features to provide a holistic analysis of job postings. The final model is then deployed using Streamlit, enabling real-time predictions of job posting authenticity. This structured approach ensures that the developed model is both accurate and practical for real-world applications.

**Natural Language Processing (NLP)**

In this project, we utilized Natural Language Processing (NLP) to analyze and extract meaningful information from the text within job postings. NLP is a crucial component of artificial intelligence that enables computer programs to understand and interpret human language as it is spoken or written. By transforming raw text data into a structured format, NLP allows machine learning models to effectively classify job postings as either real or fake.

NLP employs various computational techniques to process and analyze large amounts of natural language data. It combines the principles of computational linguistics, which is the study of how language works, with advanced statistical models, machine learning, and deep learning. These technologies enable computers to analyze text data, understand context, and grasp the full meaning of the text, including the intentions and emotions of the writer.

A screenshot of a computer screen

Description automatically generated

In this project, we built an NLP pipeline that includes several key steps:

1. **Sentence Segmentation**: Dividing the text into individual sentences to simplify further analysis.
2. **Word Tokenization**: Breaking down sentences into individual words or tokens.
3. **Stemming and Lemmatization**: Reducing words to their base or root forms to ensure consistency in text analysis.
4. **Identifying Stop Words**: Removing common words that do not contribute significant meaning to the analysis, such as "and," "the," and "is."
5. **Dependency Parsing**: Analyzing the grammatical structure of sentences to understand the relationships between words.
6. **Part-of-Speech (POS) Tagging**: Assigning parts of speech to each word, such as nouns, verbs, adjectives, etc.
7. **Named Entity Recognition (NER)**: Identifying and classifying proper nouns within the text, such as names of people, organizations, and locations.
8. **Chunking**: Grouping words into meaningful chunks, such as noun phrases or verb phrases.

A screenshot of a computer

Description automatically generated

By applying these NLP techniques, we transformed unstructured textual data from job postings into structured numerical data that our machine learning models could use. This preprocessing step is essential for feature extraction and ensures that the text data is in a format suitable for machine learning algorithms.

We utilized Random Forest and Support Vector Machine (SVM) classifiers to build the predictive models. These models were trained on a dataset that included a mix of genuine and fake job postings. The use of NLP techniques allowed us to accurately identify and classify fake job postings, providing a significant improvement over manual and rule-based systems.

**Dataset Description**

The dataset utilized for this project comprises 17,880 job descriptions, with approximately 800 labeled as fraudulent. It includes a combination of textual and meta-information about each job, making it suitable for creating classification models to detect fake job postings. Key attributes in the dataset include a unique job identifier (**job\_id**), job title (**title**), and geographical location (**location**). The **department** attribute provides information about the corporate department, while the **salary range** attribute indicates the expected salary range for the job. The dataset also contains a brief description of the company (**company profile**), a detailed job description (**description**), job requirements (**requirements**), and offered benefits (**benefits**). Additionally, it specifies whether the position allows telecommuting (**telecommuting**) and includes a target variable (**fraudulent**) indicating if the job is fake or real.

This dataset, sourced from Kaggle, is highly valuable for answering several important questions. It enables the creation of classification models that utilize both text data features and meta-features to predict which job descriptions are fraudulent. It also helps identify key traits or features (words, entities, phrases) that are indicative of fraudulent job descriptions. Additionally, the dataset can be used to run contextual embedding models to identify the most similar job descriptions and perform exploratory data analysis to uncover interesting insights.

A table with text and a description

Description automatically generated

**RESULTS**

**A screenshot of a computer

Description automatically generated**

**CONCLUSION**

In conclusion, this project successfully developed a robust machine learning model to detect fraudulent job postings using a combination of text and numeric data. By employing models like Random Forest and Support Vector Machine (SVM), the project demonstrated the potential of various ML techniques in addressing the issue of employment scams. The results showed that the Random Forest model outperformed the SVM model, achieving an accuracy of 97.22%, thus becoming the chosen model for this task. Despite these achievements, there is still room for improvement. Future enhancements could include integrating more advanced NLP techniques, such as transformers and BERT, to improve the\ detection accuracy further. Additionally, incorporating a larger and more diverse dataset could help the model generalize better across different types of job postings. Implementing real-time detection and continuous model updating can also enhance the system's adaptability to new scam patterns. Overall, this project provides a solid foundation for combating employment scams and protecting job seekers, with opportunities for further refinement and innovation.

# Acknowledgments

We take the immense pleasure in expressing gratitude to our Internal guide**,  Dr. Y J Nagendra Kumar, Head of the Department,IT**, **GRIET**. We express our sincere thanks for his encouragement, suggestions and support, which provided the impetus and paved the way for the successful completion of the project work.

We wish to express our gratitude to **Dr. Y J Nagendra Kumar,** HOD IT,our Project Coordinators **Mrs.A. Pavithra** for their constant support during the project.

We express our sincere thanks to **Dr. Jandhyala N Murthy,** Director, GRIET**,** and **Dr. J. Praveen,** Principal, GRIET**,** for providing us the conductive environment for carrying through our academic schedules and project with ease. We also take this opportunity to convey our sincere thanks to the teaching and non-teaching staff of GRIET College, Hyderabad.

# References

[1]: Fake Job Prediction using Sequential Network ; Devsmit Ranparia; Shaily Kumari; Ashish Sahani

[2]: Fake Job Prediction Using Machine Learning V Anbarasu1\* , Dr. S. Selvakani2 , Mrs. K. Vasumathi3

[3]: Online Job Posting Authenticity Prediction using Machine and Deep Learning Techniques Simiscuka, Anderson Augusto

[4]: FAKE JOB DETECTION USING MACHINE LEARNING APPROACH K. Swetha, M. Tharun Reddy, K. Sravani, B. Subramanyam

[5]<https://www.ieeexpert.com/python-projects/fake-job-post-detection-using-machine-learning/>

[6] https://youtu.be/hTS9MhGDlfU?si=nOKYRAREtHvklXS2

[8]:  Developing A Model to Detect Fraudulent Job Postings: Fake vs. Real T.Nandini1 ,S.Gnana Chandrika2 , P.Mounika3 , V.Sandeeep Kumar4 UG students