JOB SCAM ALERT

**Dr. Y J Nagendra Kumar,** Head of the Department of Information Technology, Gokaraju RangarajuInstitute of Engineering and Technology,Nizampet Road, Bachupally, Hyderabad-50090, Telangana, India. [jeevannagendra@griet.ac.in](mailto:jeevannagendra@griet.ac.in)

J. Krishna Chaitanya

C. Abhinav Reddy

Vinod Pawar

**Abstract**

In response to the rise in online scams, we created Project Scam Detection, an innovative solution that uses advanced machine learning. Our software uses random forest and support vector machine (SVM) models to detect fake job postings. The technology helps job seekers stay safe in the digital job market, protecting them from fraud and protecting their personal information. By improving the integrity of online job listings, job fraud detection creates a safer, more trustworthy environment for job seekers and employers. Our solutions not only help people avoid financial scams like false claims and fake employment promises, but also help them make informed decisions when applying for jobs online. The model is recorded and sent using Streamlit, allowing users to easily check if the job posting is fake or real. We are committed to continuously developing and improving our software to increase its effectiveness, change the way job seekers engage in the digital marketplace, and provide confidence and peace of mind during the job search. Through continuous research and development, we strive to stay ahead of emerging threats and provide a harmonious, secure environment for all stakeholders.

*Keywords:fraud, machine learning, random forest, support vector machine (SVM), fraud detection, Streamlit distribution*

**Introduction**

According to CNBC, employment scams have been on a steady rise, with this type of scam doubling from 2017 to 2018. creates fertile ground. These scammers lure interested job seekers with attractive but deceptive job offers designed to remove personal information such as addresses, bank account information, and Social Security numbers, or to trick them into asking for a fee or making a deposit. As a college student, I personally received many scam emails offering lucrative jobs that later turned out to be scams. Addressing this issue is important, and advanced machine learning techniques combined with natural language processing (NLP) are showing promise. It is a sophisticated tool that uses advanced machine learning algorithms. Our project leverages Kaggle data, which includes real and fake job search data. Although the percentage of lies is small, identifying them is important to protect job seekers. Our solutions use different types of machine learning, focusing on random forest classifiers and support vector machines (SVM). The random forest model achieved 97.22% accuracy, while the SVM model achieved 95% accuracy. The system is designed to record training examples and send them using Streamlit, allowing users to easily determine whether the posted job is real or fake. Data, models, and evaluation. The main aim is to create a classification system that can distinguish true and false job postings. This classification includes numbers and text for better understanding of the job posting. With this approach, the Fraud Detection Project aims to provide job seekers with reliable tools to help them stay safe in the market and make informed decisions.

**1.1.*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. 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.

1. **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 help candidates stay alert. Their model uses nlp to analyze sentiments & patterns postings, training a sequential neural network. 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 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 the use of various data mining techniques and classification algorithms such as KNN, decision trees, support vector machines, naive Bayes, random forests, multilayer perceptrons and deep neural network methods to predict unselected data. They conducted an experiment on the Ege Employment Fraud Dataset (EMSCAD) using 18,000 samples. Their deep neural network classifier with three thick layers achieves about 98% accuracy in predicting invalid task names [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. discuss the increase in online job postings during the pandemic and the need to accurately detect fake jobs. Their study uses various data mining and classification algorithms, including KNN, decision trees, support vector machines, naive Bayes, random forests, multilayer perceptrons, and deep neural networks, using the EMSCAD dataset containing 18,000 samples. Deep neural network classifiers with three thick layers achieved approximately 98% accuracy in predicting fake job titles.

This section presents the performance of a proposed method to predict fake job postings using machine learning techniques and natural language processing (NLP). This study includes several stages: problem definition, data collection, data cleaning and prioritization, modeling and evaluation. Figure 1 shows the operation of the proposed model.

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 use natural language processing (NLP) to analyze the text in job postings and provide useful information. NLP is an important form of artificial intelligence that allows computers to understand and interpret human speech or writing. By transforming raw data into structured patterns, NLP enables machine learning models to be more efficient at reporting tasks as correct or incorrect. It combines the concepts of computational speech (the study of how things work) with statistical models, machine learning, and deep learning. These technologies allow computers to analyze written information, understand context, and understand the full context of a text, including the author’s intent and intention.

In this project, we created an NLP pipeline that includes several important steps:

Sentence Segmentation: Split the text into sentences to facilitate further analysis. Words or symbols. Single words like “and”, “the” and “is”.

Dependency syntax analysis: Analyze the grammatical structure of sentences to understand the relationship between words.

Part of Speech (POS) tag: Assign a part of speech such as noun, verb, adjective, etc. to each word. Semantic parts such as noun phrases or verb phrases. This preliminary step is important to remove the video and ensure that the document is in a format suitable for machine learning algorithms. The model was trained on a dataset containing real and fake job postings. Using NLP technology allows us to identify the truth and classify fake job postings; this is a significant improvement in guidelines and legal procedures.

Random Forest

Random Forest is a powerful machine learning algorithm widely used for classification and regression. It works by creating multiple decision trees throughout the learning process. Each tree in the forest is created using a random subset of the dataset, and a random subset of features is analyzed at each node. This diversity of trees helps reduce overfitting and improves the overall prediction of the model. The output data is preprocessed and transformed using NLP techniques to create a set of codes. Each tree is trained on a different random subset of the dataset and evaluated on a different random subset of features. This ensures that each tree provides a unique view of the material. The final distribution is determined by the majority vote of the trees. For example, if most of the trees classify the job posting as fake, then random forest matching will flag it as fake. This is a common problem in decision tree algorithms. The random forest algorithm provides better and more accurate results by averaging the predictions of each tree. It processes many features from the job posting to ensure the distribution.

Advantages of Random Forest in Our Studies

High Accuracy: The combination of methods combining multiple decision trees contributes to the high accuracy of classification of published studies. Adaptability ensures that the overall model is good for new, unseen information. Real world application.Support Vector Machine (SVM)

***SVM in Job Scam Detection***

SVM in Finding the Optimal Hyperplane

Optimal Hyperplane: In our project, SVM aims to find the hyperplane that maximizes the separation of the closest points from multiple clusters. This margin is important because it determines the boundary of the distinction between real and fake job advertisements. The decisions are extremely unfair. This change allows SVM to capture complex patterns and relationships in the text extracted from published work. Converting the text into usable numbers can be done by SVM. its performance helps provide strong and reliable predictions in detecting fake job advertisements. Preventing job fraud is important to protect job seekers from financial fraud and scams.

Existing System

The existing system for detecting fake job postings primarily relies on manual review processes or basic rule-based filtering methods. These systems typically involve human reviewers who examine job postings for signs of fraud based on their experience and predefined rules. For instance, they might check for discrepancies in job details, unusual salary ranges, or suspicious email addresses. However, this approach has several limitations:

Scalability Issues: As the volume of job postings increases, manual review becomes impractical and inefficient.

Subjectivity and Human Error: The accuracy of detecting fake job postings depends heavily on the expertise and vigilance of the reviewers, making it prone to human error.

Slow Response Time: Manual review processes are time-consuming, resulting in delayed identification and removal of fake job postings.

Limited Adaptability: Rule-based systems are often static and cannot adapt quickly to new types of scams or evolving fraudulent techniques.

**Proposed System**

The proposed system uses machine learning and natural language processing (NLP) techniques to detect fake jobs. The system is designed to solve the limitations of the existing system by providing broad, accurate and flexible solutions.

Key Components:

1. Data Collection
2. Feature Engineering: Create new features from the existing data, such as combining text fields and generating character counts.
3. Model Training: Train machine learning models (Naive Bayes, SGD Classifier and LSTM) on the preprocessed data.
4. Model Evaluation: Evaluate the models using accuracy and F1-score metrics.
5. Final Classification: Combine the outputs of the models to determine the authenticity of job postings.

*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:* This study focuses on the use of random forest and support vector machine (SVM) algorithms as they are effective in classifying tasks. The model is trained using a pre-programmed method to learn patterns and distinguish between real and fake job postings. Tests include precision, accuracy, recall, and F1 score to ensure stability and reliability in detecting job vacancy fraud.

*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 taking advantage of these developments, the system aims to improve the detection of fake job postings, provide solutions, and make corrections and adjustments to solve changing problems in the digital business process.

***ABOUT PYTHON***

Python's environment has evolved significantly, enhancing its capabilities for statistical analysis. It strikes a fine balance between scalability and elegance, placing a premium on efficiency and code readability. Python is renowned for its emphasis on program readability, featuring a straightforward syntax that is beginner-friendly and encourages concise code expression through indentation. Noteworthy aspects of this high-level language include dynamic system functions and automatic memory management.

***APPLICATIONS OF PYTHON***

Python is used in many application areas. It touches all developing areas. It is the fastest programming language and can be used to create any type of application.

***PYTHON IS WIDELY USED IN MACHINE LEARNING***

Python is widely favored in machine learning for its flexibility and open-source nature. It provides extensive functionality for mathematical computations and scientific operations, making it indispensable in developing and deploying machine learning models. Python's simple syntax and vast libraries accelerate the development process, reducing coding time significantly.

This makes it a preferred choice for machine learning practitioners seeking efficiency and robustness in their projects.

The major Python libraries used in machine learning are as follows:

*PANDAS*

Pandas is a Python library for data analysis, data cleaning, searching and management. Generally, the dataset contains important and irrelevant information. Pandas helps to make this information more readable and relevant. It provides powerful information that can be used to calculate large numbers and matrices, making it easier to access and manage information.

*SCIKIT-LEARN*

Scikit-Learn is a Python library for machine learning. It provides tools for machine learning and statistical models, including classification, regression, and clustering.

*SEABORN*

Seaborn is a Python data visualization library based on matplotlib. It provides attractive graphs and an advanced interface for statistical data. You can read the introduction or the text for a brief introduction to the concept behind the library.

*NUMPY*

NumPy is a Python library utilized for numerical data reading, cleaning, exploration, and manipulation. It provides powerful data structures for efficient computation with large arrays and matrices, making the data more accessible and manageable.

*MATPLOTLIB*

Matplotlib is a Python library for plotting graphs. Built on NumPy arrays, it allows for the creation of a wide range of graph types, from basic plots to bar graphs, histograms, scatter plots, and more.

*CSV*

It is a kind of file that stores tabular records, like a spreadsheet or a database. There are one or extra fields in each entry, that are separated by commas. We use the csv built in module to work with CSV files.

*INTERPRETED LANGUAGE*

 Python executes code line by line, without the need for prior compilation. This approach facilitates quicker development cycles and simplifies the debugging process. As a result, developers can iterate and test their code more efficiently.

***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 document is from Kaggle and is useful for answering many important questions. It can create a classification model that uses the features and meta-content of the text to predict which job descriptions are fraudulent. It also helps identify important features or characteristics (words, connotations, phrases) that indicate fraudulent activity. This profile can also be used to run a content analysis model to identify similar job descriptions and perform profile analysis to uncover the best opportunities.

* Total Entries: 17,880
* Fraudulent Jobs: 725 (approximately 7%)
* Real Jobs: 17,155 (approximately 93%)

1. **Proposed Architecture**

A diagram of a software development process

Description automatically generated

Unified Modeling Language (UML) is used as a design language to create models in multiple formats. Its primary purpose is to visually represent structure, similar to the structure in engineering disciplines. In complex applications involving multiple teams, clear communication is important, especially with stakeholders who are not familiar with the policy. UML facilitates this communication by visualizing the expression of formal systems, functions, and processes. UML facilitates collaboration and optimization by defining static models of processes, user interactions, and systems for teams.

Use Case Diagram

A use case diagram is a behavioral diagram that is a description of the functions provided by the system, including actors, their goals, and dependencies on those uses.A diagram of a scam detection system

Description automatically generated

Fig – 5 User Case Diagram

Class Diagram

In Unified Modeling Language (UML), a class diagram is a static structure diagram that describes the groups, objects, and methods of the system to define the structure of the sequence and the relationship between objects. relationship. They are used to view, share, create, and document artifacts of software systems.

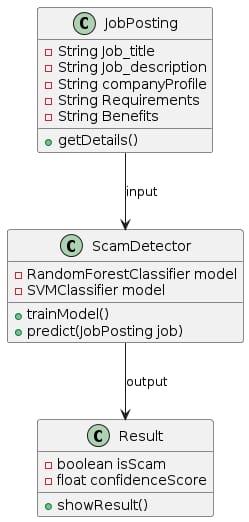
****

Fig – 6 Class Diagram

*Activity Diagram*

This diagram is a more complex version of a flow chart that depicts the flow of information from one activity to the next. It describes the coordination of activities in order to offer a service at various levels of abstraction.

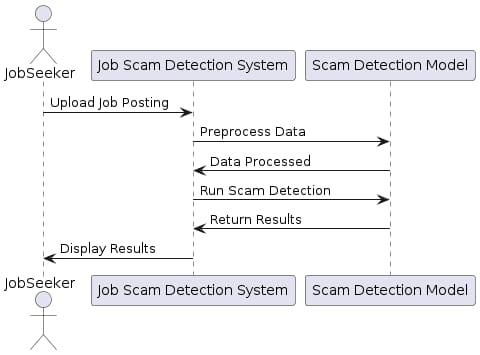


Fig – 7 Activity Diagram

*Sequence Diagram*

One kind of UML (Unified Modeling Language) diagram that shows how items interact in a certain order of time is a sequence diagram. They are especially helpful in describing the steps involved in an activity or the messages that are sent back and forth between objects to complete a task. I'll give a general rundown of the elements and procedures needed to create a sequence diagram below, along with an example that focuses on a CNN-based picture forgery detection system.

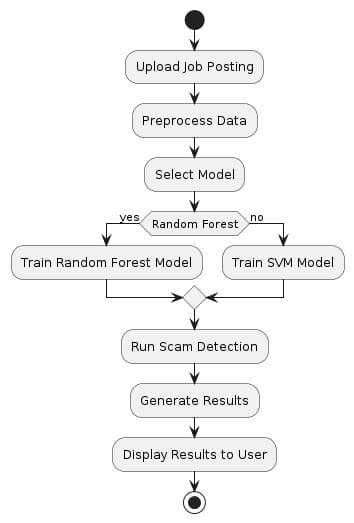
**

Fig – 8 Sequence Diagram

1. **A screenshot of a computer

   Description automatically generatedResults**

Fig – 11 Result

1. **Conclusions**

Overall, the successful project has created a powerful machine learning system that can combine text and digital data to detect fake ads. Using models such as random forests and support vector machines (SVMs), the project demonstrates the potential of various machine learning techniques to solve operational fraud problems. The results show that the random forest model outperforms the SVM model with 97.22% accuracy, making it the preferred model for this task. Despite these achievements, there is still room for improvement. Future improvements could include combining multiple NLP techniques such as Transformer and BERT to increase detection accuracy. Additionally, incorporating larger and more diverse data could help the model better expand to different types of job postings. Tracking ad hoc searches and constantly updated patterns could allow the body to adapt to new fraud patterns. Overall, the program provides a solid foundation for combating employment fraud, protecting job seekers, and providing opportunities for improvement and innovation.

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