International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comparitive Study on Fake Job Post Prediction Using Different Data Mining Techniques

Mrs.K.Surekha[1], Jammu Yoshitha Devi[2], Chintapalli Durga Bhavani[3], Koppisetti Sri Durga[4], Madiki Durga Ganesh[5]

[1]Assistant Professor, CSE Department, VSM College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh – 533255
[2,3,4,5,6]Student, CSE Department, VSM College Of Engineering, Ramachandrapuram, Dr.B.R.Ambhedkar Konaseema Dt, Andhra Pradesh – 533255

ABSTRACT
In today's digital job market, the rise of online recruitment has also led to a surge in fraudulent job postings, putting job seekers at risk. This project, titled *"**A Comparative Study on Fake Job Post Prediction Using Different Data Mining Techniques**,"* focuses on identifying fake job listings through a machine learning approach. Using a **Fake job Postings dataset** from Kaggle that includes both genuine and fraudulent job posts, the study applies a series of preprocessing steps, including feature selection (job title, location, description, requirements, and company profile) and text transformation using TF-IDF vectorization.
The dataset is split into training and testing sets, and five classification algorithms are evaluated: **Logistic Regression, Multinomial Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine (SVM)**. Each model's performance is measured using accuracy, precision, recall, F1-score, and confusion matrix. A final bar chart visualizes the comparison, highlighting which model performs best for this task. This comparative study offers valuable insights into the effectiveness of different data mining techniques in detecting fake job postings and supports the development of safer, more trustworthy online hiring platforms.

Keywords: Machine learning, Data mining techniques, Classification algorithms, TF-IDF vectorization, Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), Feature selection, Job post dataset, Kaggle dataset, Model evaluation metrics, Accuracy, Precision, Recall, F1-score, Confusion matrix, Comparative analysis, Fraud detection,

INTRODUCTION

In recent years, the rapid expansion of online job recruitment platforms has introduced not only convenience for job seekers but also a significant increase in fraudulent job postings. These fake job advertisements are often created by scammers to steal sensitive information, extract money, or exploit victims under the guise of employment opportunities. The aim of this project is to develop a robust and scalable system capable of identifying such deceptive job posts using various data mining and machine

learning techniques.

This project performs a comparative study on the effectiveness of multiple machine learning classifiers in detecting fake job advertisements. The study uses the Fake Job Postings Dataset sourced from Kaggle, which contains thousands of real and fraudulent job listings. Key features selected from the dataset include job title, location, description, requirements, and company profile. These features are combined into a single text field to serve as the input for the machine learning models.The text data undergoes preprocessing and is then vectorized using TF-IDF (Term Frequency–Inverse Document Frequency) to convert textual information into numerical features suitable for classification tasks. After vectorization, the dataset is split into training and testing subsets. Five different machine learning classifiers are implemented and compared:

- Logistic Regression
- NaïveBayes (MultinomialNB)
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

Each model is trained on the training set and evaluated on the test set using metrics such as accuracy, classification report (precision, recall, F1-score), and confusion matrix. A performance comparison chart is created to visually represent and analyze the accuracy of each model.

This study demonstrates the potential of classical machine learning models in detecting fraudulent job posts. By identifying patterns in text descriptions, the system is able to identify the fake job postings with considerable accuracy. The project serves as a foundation for implementing fraud detection tools in job portals, HR platforms, and social recruitment systems.

EXISTING SYSTEM

Current fraud detection mechanisms primarily address general-purpose online threats such as spam detection in emails or social media platforms. Techniques developed for platforms like Twitter and Facebook focus on identifying spam users or bot accounts based on behavioral patterns and user interaction metrics. While effective in their respective domains, these systems lack the specialization required to detect fraud in job recruitment platforms.

Job scams are often more nuanced and context-driven. They may contain legitimate-sounding titles, descriptions, and company names, making them difficult to identify using standard spam filters or generalized models. Furthermore, these systems often rely on limited features or outdated data, which reduces their ability to adapt to the constantly evolving tactics used by scammers.

Many existing models also suffer from practical limitations:

- Low predictive accuracy due to generic feature selection.
- Inadequate datasets because of restrictions on accessing real-world job data.
- Overly complex models that are difficult to deploy and maintain in real-world environments.

These constraints underscore the need for a domain-specific, lightweight, and high-performing model that can detect fake job postings with improved accuracy and minimal computational resources.

PROBLEM STATEMENT

In modern technology and social communication, advertising new job posts has become very common issue in the present world. So, fake job posting prediction task is going to be a great concern for all. Like

many other classification tasks, fake job posing prediction leaves a lot of challenges to face.

PROPOSED SYSTEM

The proposed system introduces a comparative machine learning framework to predict fraudulent job postings using a clean and structured dataset. It takes advantage of classic classification algorithms known for their efficiency, interpretability, and ease of deployment. The process begins with feature selection from the raw dataset, where key textual elements such as job title, job location, description, requirements, and company profile are selected and combined into a single textual field.

This textual data is processed using TF-IDF vectorization, a widely used technique in text mining, which transforms the text into a matrix of numerical features by measuring the importance of each word relative to the corpus. The transformed data is then split into training and testing sets.

The system implements the following models for classification:

- Logistic Regression
- Multinomial Naive Bayes
- Decision Tree Classifier
- Random Forest Classifier
- Support Vector Machine (SVM)

Each model is evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Finally, a bar chart is generated to compare the accuracy of all models visually, enabling straightforward interpretation of which algorithm performs best for this particular dataset and use case.

Unlike some advanced models, the proposed system does not rely on deep learning or active learning techniques in this phase but is modular enough to be extended with such features in the future. The use of multiple models provides a broad overview of what works best in the domain of fake job detection.

IMPLEMENTED ALGORITHMS

LOGISTIC REGRESSION:

Purpose: Logistic Regression is a classification algorithm used for binary outcomes (e.g., real vs. fake job posts).

Working: It calculates the probability of an input belonging to a class using the sigmoid (logistic) function.

Decision Rule: If the predicted probability > 0.5, it is classified as one class; otherwise, it belongs to the other.

In Context of Fake Job Post Prediction:

Prediction Use: Logistic Regression predicts whether a job post is real or fake based on features like title, description, and company profile.

Confidence & Accuracy: It provides a probability score to flag suspicious listings and achieved **99.6% accuracy** in our study.

NAIVE BAYES:

Purpose: Naive Bayes is a classification algorithm based on Bayes' Theorem, commonly used for text-based classification tasks.

Working: It calculates the probability of each class based on the input features, assuming all features are independent (naive assumption).

Decision Rule: The model predicts the class with the highest posterior probability given the input data.

In Context of Fake Job Post Prediction:

Prediction Use: Naive Bayes classifies job posts as real or fake by analyzing textual features like title, description, requirements, and company profile after TF-IDF vectorization.

Confidence & Accuracy: It identifies word usage patterns in genuine vs. fake listings and achieved **98.0% accuracy** in our study.

DECISION TREE:

Purpose: Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks.

Working: It functions like a flowchart—internal nodes represent decisions based on features, branches show outcomes, and leaves indicate final class labels.

Decision Rule: The model follows if-else conditions to split the data and make predictions based on feature values.

In Context of Fake Job Post Prediction:

Prediction Use: The Decision Tree Classifier analyzes TF-IDF-transformed text from job posts, learning rules from fields like title, description, and company profile to identify fraud.

Confidence & Accuracy: It detects patterns linked to fake listings and achieved **99.64% accuracy** in our study, making it highly effective.

RANDOM FOREST:

Purpose: Random Forest is an ensemble classification algorithm that combines the outputs of multiple decision trees to enhance accuracy and reduce overfitting.

Working: Each tree is trained on a random subset of the data; predictions are made by majority voting across all trees.

Decision Rule: The final prediction is determined by the class with the most votes from the individual trees.

In Context of Fake Job Post Prediction:

Prediction Use: Random Forest classifies job posts as real or fake by learning patterns in TF-IDF-transformed features like job title, description, and company profile.

Confidence & Accuracy: It generates stable, high-accuracy predictions through ensemble learning and achieved **99.6% accuracy** in our study.

SUPPORTVECTOR MACHINE:

Purpose: SVM is a supervised machine learning algorithm designed for binary classification tasks, particularly effective in high-dimensional feature spaces.

Working: It constructs an optimal hyperplane that maximizes the margin between two classes. Only critical data points (support vectors) influence this boundary.

Decision Rule: The model assigns a class label based on which side of the hyperplane the input data lies.

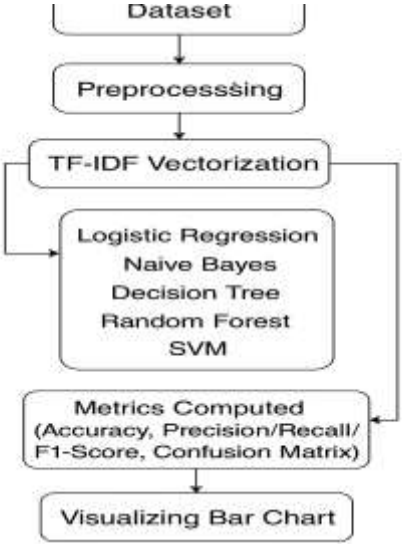
In Context of Fake Job Post Prediction:

Prediction Use: SVM identifies whether a job post is real or fake by analyzing TF-IDF vectorized text from fields like title, description, requirements, and company profile.

Confidence & Accuracy: Leveraging its strength in handling complex, high-dimensional data, SVM

achieved **99.6% accuracy**, making it one of the most precise models in the study for detecting fraudulent listings.

SYSTEM ARCHITECTURE



The proposed system features a modular architecture for detecting fake job postings using classical machine learning and efficient natural language processing. It begins with acquiring data from sources like Kaggle, followed by preprocessing steps such as handling missing values, removing duplicates, and consolidating features. Text fields like job title, description, and requirements are vectorized using TF-IDF to create numerical input features. Multiple supervised models—Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and SVM—are trained and evaluated using a 70-30 train-test split, with performance visualized through bar graphs and confusion matrices. Prioritizing explainability, models like Decision Trees and Logistic Regression are emphasized for transparency, making the system suitable for deployment in HR screening tools and job post verification platforms. Though currently limited to traditional models, the design supports future integration of real-time learning and advanced AI, delivering a reliable, scalable, and interpretable solution for fraud detection in digital recruitment.

Data Ingestion Module: The Data Ingestion Module is responsible for loading and preparing the job postings dataset, typically sourced from platforms like Kaggle. This module reads the CSV file containing structured job listing information, including fields such as title, location, description, requirements, and company_profile. The module then extracts the relevant columns and performs initial preprocessing tasks, including the removal of missing or null values to ensure data consistency. All selected text-based columns are concatenated to form a single unified textual representation of each job posting, which becomes the input for further analysis and vectorization.

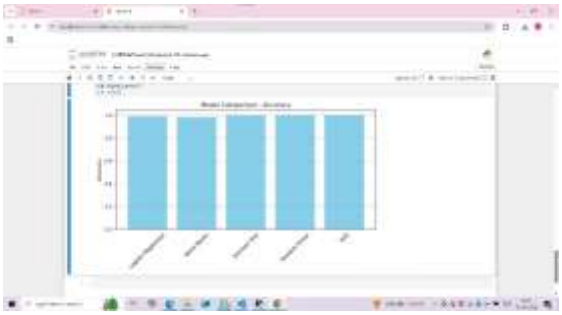
TF-IDF Vectorization Module: This module performs **text vectorization** using the TF-IDF (Term Frequency–Inverse Document Frequency) technique, a common method to convert textual data into numerical features that can be understood by machine learning algorithms. The unified job posting text is transformed into a matrix of TF-IDF features, where each word is weighted according to its importance in the document relative to the entire corpus. This module eliminates stop words and limits the influence of extremely frequent terms using parameters like max_df. The result is a sparse matrix that numerically represents each job posting, optimized for training classification models.

Model Development & Metrics Comparitive Module: The Model Development Module defines, trains,

and organizes multiple supervised machine learning algorithms—Logistic Regression, Multinomial Naive Bayes, Decision Tree, Random Forest, and SVM—using TF-IDF vectorized text data to classify job posts as fake or legitimate. Models are stored in a dictionary and trained using. fit () for streamlined evaluation. The Model Metrics Comparative Module assesses each model's performance through accuracy, confusion matrices, and classification reports with precision, recall, and F1-score, offering a comprehensive view of effectiveness, especially for imbalanced datasets.

Visualization & Comparision Module: This module visualizes the comparative performance of all trained models through graphical representation. Using libraries like **matplotlib**, it generates a bar chart that displays the accuracy scores of each classifier side by side. This visualization aids in quickly identifying which model performs best for the fraud detection task. The chart is styled for clarity with color coding, axis labels, grid lines, and rotated tick labels to enhance readability. This module plays a key role in supporting data-driven decision-making for model selection.

OUTPUT



CONCLUSION

This project focused on detecting fake job postings using machine learning models applied to textual job data. After preprocessing the dataset and combining relevant text features such as job title, description, requirements, and company profile, the data was vectorized using the TF-IDF technique to convert the text into numerical features. Several machine learning models, including **Logistic Regression, Naive Bayes, Decision Tree, Random Forest, and SVM**, were trained and evaluated using accuracy, classification reports, and confusion matrices.

The results showed that all models performed well, with particularly high accuracy achieved by the Decision Tree and Random Forest classifiers. The Decision Tree model stood out with an accuracy of approximately **99.6%** and very few misclassifications, making it highly effective in distinguishing between real and fake job postings. Overall, this project demonstrates that machine learning can be a powerful tool in identifying fraudulent job ads and can contribute to building safer and more trustworthy online job platforms.

REFERENCES

1. C. Chen, S. Wen, J. Zhang, Y. Xiang, J. Oliver, A. Alelaiwi, and M. M. Hassan, ‘‘Investigating the deceptive information in Twitter spam,’’ Future Gener. Comput. Syst., vol. 72, pp. 319–326, Jul. 2017.
2. M. Babcock, R. A. V. Cox, and S. Kumar, ‘‘Diffusion of pro- and anti-false information tweets: The black panther movie case,’’ Comput. Math. Org. Theory, vol. 25, no. 1, pp. 72–84, Mar. 2019.
3. S. Keretna, A. Hossny, and D. Creighton, ‘‘Recognising user identity in Twitter social networks via text mining,’’ in Proc. IEEE Int. Conf. Syst., Man, Cybern., Oct. 2013, pp. 3079 3082.
4. C. Meda, F. Bisio, P. Gastaldo, and R. Zunino, ‘‘A machine learning approach for Twitter spammers detection,’’ in Proc. Int. Carnahan Conf. Secur. Technol. (ICCST), Oct. 2014, pp. 1–6.
5. W. Chen, C. K. Yeo, C. T. Lau, and B. S. Lee, ‘‘Real-time Twitter content polluter detection based on direct features,’’ in Proc. 2nd Int. Conf. Inf. Sci. Secur. (ICISS), Dec. 2015, pp. 1–4.
6. H. Shen and X. Liu, ‘‘Detecting spammers on Twitter based on content and social interaction,’’ in Proc. Int. Conf. Netw. Inf. Syst. Comput., pp. 413–417, Jan. 2015.
7. G. Jain, M. Sharma, and B. Agarwal, ‘‘Spam detection in social media using convolutional and long short-term memory neural network,’’ Ann. Math. Artif. Intell., vol. 85, no. 1, pp. 21 44, Jan. 2019.
8. M. Washha, A. Qaroush, M. Mezghani, and F. Sedes, ‘‘A topic-based hidden Markov model for real-time spam tweets filtering,’’ Procedia Comput. Sci., vol. 112, pp. 833–843, Jan. 2017.