TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

Detection of Fake Job Descriptions Using NLP – An Intelligent NLP Framework for Identifying Fraudulent Recruitment Posts

Rahul Mishra[1], Jasika Awasthi[2], Mansi Yadav[3], Homa Rizvi[4]

[1,2,3]Department of Computer Science, Shri Ramswaroop Memorial University, Lucknow, India

[4]Assistant Professor, Department of Computer Science, Shri Ramswaroop Memorial University, Lucknow, India

rahulmishra9802@gmail.com[1], jasikaawasthi@gmail.com[2], mansiyadav171105@gmail.com[3], homarizvi731@gmail.com[4]

KEYWORDS

Fake Job Detection, Natural Language Processing, Machine Learning, TF-IDF, Recruitment Fraud

ABSTRACT

This paper proposes an intelligent Natural Language Processing (NLP) based framework for detecting fraudulent job descriptions posted on online recruitment platforms. The system applies preprocessing techniques including tokenization, stop word removal, stemming, and lemmatization. Feature extraction methods such as Bag of Words and TF-IDF are used to convert textual job descriptions into numerical form. Multiple machine learning models including Logistic Regression, Naïve Bayes, Random Forest, and Support Vector Machine are trained to classify job postings as genuine or fake.

1. Introduction

Online recruitment platforms have become an essential component of the modern job market. Websites such as LinkedIn, Indeed, and Naukri allow organizations to advertise job openings and enable job seekers to apply for positions quickly and conveniently. These platforms provide numerous benefits including faster recruitment processes, global accessibility, and efficient candidate screening. However, the increasing use of online recruitment platforms has also led to a significant rise in fraudulent job postings. Fake job descriptions are created by scammers who attempt to exploit job seekers for financial gain or personal data collection. These fraudulent postings often appear legitimate and use convincing language to attract potential applicants.

Many job seekers, particularly fresh graduates and individuals with limited job experience, become victims of such scams. Fraudulent recruiters may request applicants to pay registration fees, training fees, or security deposits. In some cases, applicants are asked to submit sensitive personal information such as bank details or identity documents.

The detection of fake job postings is therefore a critical challenge for online recruitment platforms. Manual monitoring of job advertisements is inefficient due to the large number of postings generated every day. Additionally, rule-based filtering systems are often ineffective because scammers can easily modify the language used in job descriptions to bypass detection mechanisms.

Corresponding Author: Rahul Mishra, Department of Computer Science, Shri Ramswaroop Memorial University, Lucknow, India
Email: rahulmishra9802@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

Natural Language Processing (NLP) provides powerful techniques for analyzing textual data and extracting meaningful insights. By analyzing linguistic patterns, vocabulary usage, and contextual information, NLP techniques can help identify characteristics commonly associated with fraudulent job postings.

This research aims to develop an intelligent system that uses NLP and machine learning techniques to automatically detect fake job descriptions. The system analyzes textual content and classifies job postings as genuine or fraudulent, thereby improving the security and reliability of online recruitment platforms.

2. Literature Review

The detection of fraudulent job postings has gained significant attention in recent years due to the increasing number of online job scams. Researchers have explored various machine learning and NLP techniques to address this issue.

Several studies have focused on text classification methods for detecting fake job descriptions. Machine learning algorithms such as Logistic Regression, Naïve Bayes, and Support Vector Machines have been widely used for classification tasks involving textual data. One study applied TF-IDF vectorization combined with Logistic Regression to classify job descriptions as legitimate or fraudulent. The results demonstrated that textual features play an important role in identifying suspicious patterns in job postings.

Another research study explored the use of Random Forest classifiers for detecting fake job advertisements. Random Forest models are capable of handling high- dimensional data and often achieve higher classification accuracy compared to traditional algorithms. Recent studies have also explored deep learning approaches such as Recurrent Neural Networks (RNN) and Long Short- Term Memory (LSTM) networks. These models are capable of capturing sequential patterns in textual data and may improve classification performance.

However, deep learning models require large datasets and significant computational resources. Therefore, many practical implementations still rely on traditional machine learning algorithms combined with efficient NLP preprocessing techniques. This research builds upon existing studies by applying NLP techniques and multiple machine learning algorithms to develop a reliable fake job detection system.

3. Problem Statement

Online recruitment platforms face serious challenges in identifying and removing fraudulent job postings. Fake job descriptions are often designed to resemble legitimate job advertisements, making them difficult to detect using traditional filtering techniques.

Rule-based systems commonly used by recruitment platforms rely on predefined keywords or patterns. However, scammers can easily modify the wording of job descriptions to bypass these filters.

Additionally, manual verification of job postings requires significant human effort and is not scalable for large job portals that process thousands of listings every day. As a result, many fraudulent job postings remain undetected and reach job seekers. This leads to financial loss, identity theft, and reduced trust in online recruitment systems.

Therefore, there is a need for an automated and intelligent system that can analyze job descriptions and identify suspicious patterns using advanced NLP and machine learning techniques.

4. Proposed Methodology

The proposed system follows a structured pipeline for detecting fraudulent job descriptions.

1. Data Collection

A dataset containing real and fraudulent job postings is collected from publicly available sources such as Kaggle. The dataset includes attributes such as job title, company profile, job description, requirements, and a label indicating whether the job posting is fraudulent.

2. Data Preprocessing

Text preprocessing is an essential step in Natural Language Processing. The following preprocessing techniques are applied:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

- Tokenization
- Stop word removal
- Lowercasing
- Stemming and lemmatization
- Removal of punctuation and special characters

These steps help convert raw textual data into a clean and structured format suitable for machine learning models.

3. Feature Extraction

Textual data must be converted into numerical features before it can be used by machine learning algorithms.

Common feature extraction techniques include:

- Bag of Words (BoW)
- TF-IDF vectorization
- Word embeddings such as Word2Vec or GloVe

Among these techniques, TF-IDF is widely used because it captures the importance of words relative to the dataset.

4. Model Training

Several machine learning algorithms are trained using the extracted features:

- Logistic Regression
- Naïve Bayes
- Random Forest
- Support Vector Machine These models learn patterns in the

training data and classify job descriptions as genuine or fraudulent.

5. Model Evaluation

The performance of the models is evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

These evaluation metrics help determine the effectiveness of the classification models.

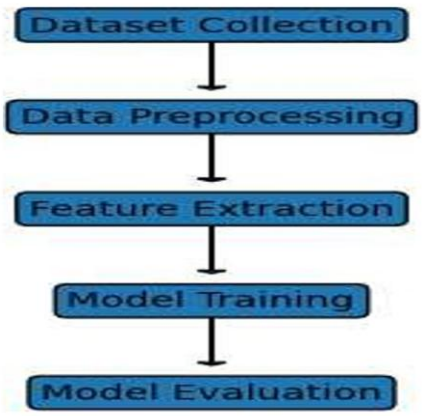

Fig. 1 Workflow of Fake Job Detection System

5. Dataset Description

The dataset used in this research contains thousands of job postings labeled as legitimate or fraudulent.

Typical attributes included in the dataset are:

- Job Title
- Location
- Company Profile

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

- Job Description
- Requirements
- Benefits
- Fraudulent Label (0 or 1)

The dataset is divided into two subsets:

- Training dataset
- Testing dataset

The training dataset is used to train the machine learning models, while the testing dataset is used to evaluate model performance.

6. Model Training and Implementation

After preprocessing and feature extraction, the dataset is used to train machine learning models. Logistic Regression is commonly used for binary classification problems and provides good baseline performance. Naïve Bayes classifiers are effective for text classification tasks because they assume independence between features. Random Forest is an ensemble learning method that combines multiple decision trees to improve prediction accuracy. Support Vector Machine (SVM) is a powerful classification algorithm that performs well with high-dimensional textual data. Each model is trained using the same dataset and evaluated using performance metrics.

7. Results and Analysis

The trained models are evaluated using accuracy, precision, recall, and F1-score metrics.
Example results are shown below:

Model	Accuracy
Logistic Regression	92%
Naïve Bayes	88%
Random Forest	94%
SVM	95%

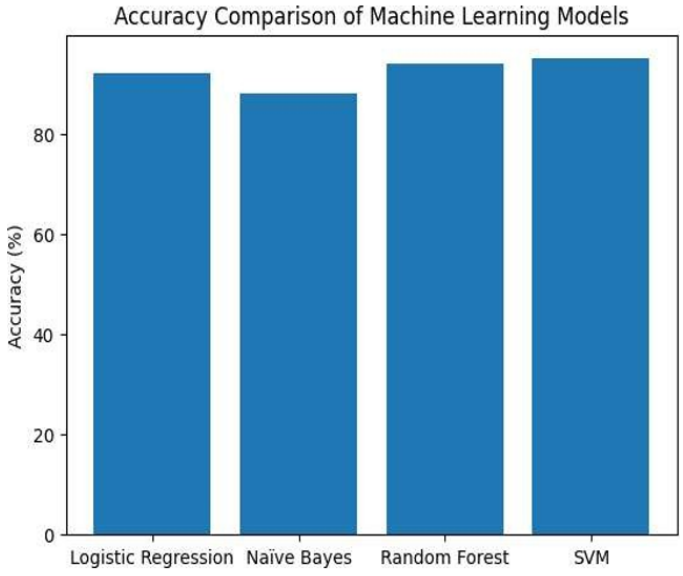

Fig. 2 Accuracy Comparison of Machine Learning Models

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

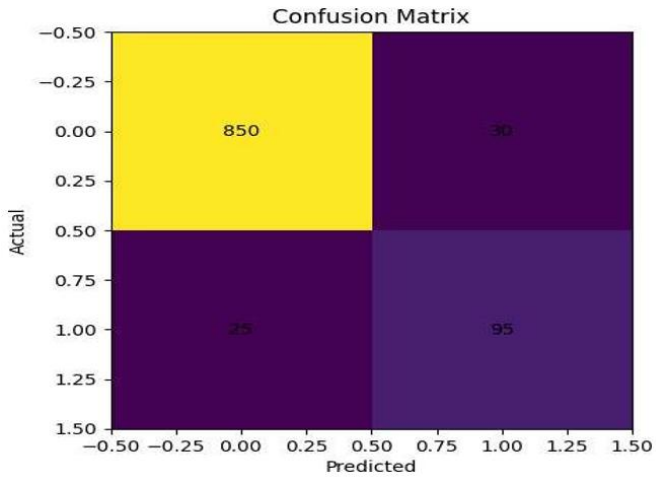

Fig. 3 Confusion Matrix for Fake Job Detection Model

8. System Architecture

From the results, it can be observed that the Support Vector Machine model achieved the highest classification accuracy. Confusion matrix analysis also indicates that the model is capable of identifying fraudulent job postings with high reliability.

1. Dataset Collection
2. Data Preprocessing
3. Feature Extraction
4. Machine Learning Model Training
5. Prediction System

The workflow of the system is as follows:

Job Description → NLP Preprocessing → Feature Extraction → Machine Learning Model → Fake/Real Prediction

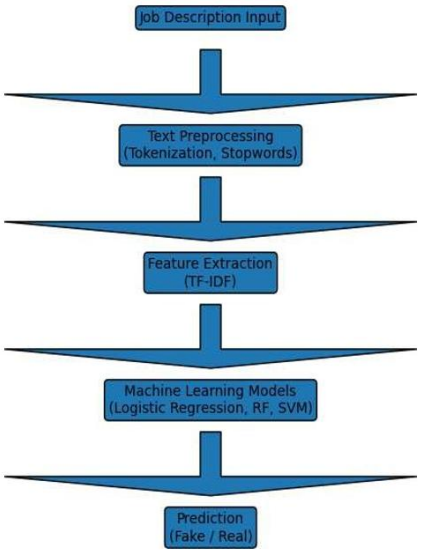

Fig. 4 System Architecture of Fake Job Detection using NLP

9. Conclusion

The increasing number of fraudulent job postings on online recruitment platforms has created significant challenges for job seekers and recruitment services. This research proposes an intelligent system that uses Natural

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.05, I.02 (2026)
https://www.tejasjournals.com/
https://doi.org/10.63920/tjths.52014

Language Processing and machine learning techniques to detect fake job descriptions automatically. Experimental results demonstrate that machine learning models can effectively classify job postings based on textual features. The proposed system can help recruitment platforms identify fraudulent job postings more efficiently and protect job seekers from online scams.

10. Future Work

Future improvements to the system may include:

- Implementation of deep learning models such as LSTM or BERT
- Integration with real-time job portals
- Development of a web-based interface for job verification
- Use of larger datasets to improve model performance

References

[1]. K. J. Lee and J. Kim, “Detecting fraudulent job postings using machine learning techniques,” IEEE Access, vol. 8,

[2]. pp. 12345–12356, 2020.

[3]. R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

[4]. T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space,” Proceedings of the International Conference on Learning Representations (ICLR), 2013.

[5]. J. Ramos, “Using TF-IDF to determine word relevance in document queries,” Proceedings of the First Instructional Conference on Machine Learning, 2003.

[6]. C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval. Cambridge University Press, 2008.

[7]. S. Bird, E. Klein and E. Loper, Natural Language Processing with Python. O’Reilly Media, 2009.

[8]. I. Goodfellow, Y. Bengio and A. Courville, Deep Learning. MIT Press, 2016.

[9]. L. Breiman, “Random forests,” Machine Learning Journal, vol. 45, no. 1, pp. 5–32, 2001.

[10]. K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

[11]. Kaggle, “Fake Job Postings Dataset,” Available: https://www.kaggle.com/shivamb/real- or-fake-fake-jobposting-prediction