Fake Job Detection Using Advanced Machine Learning Techniques

Lahari Challa
Arizona State University
Tempe, AZ
lchalla2@asu.edu

Shashank Navad
Arizona State University
Tempe, AZ
snavad@asu.edu

Rahul Babu
Arizona State University
Tempe, AZ
rbabu3@asu.edu

Kushal Keshav Ravipati
Arizona State University
Tempe, AZ
kravipa1@asu.edu

ABSTRACT

The surge in digital job portals has led to a parallel rise in deceptive job advertisements that endanger applicants by exposing them to identity theft and financial scams. This research introduces a machine learning-driven framework to automatically identify such fraudulent listings. Treating it as a binary classification problem, we develop a robust pipeline incorporating advanced natural language processing techniques such as TF-IDF weighting, GloVe-based word embeddings, sentiment scoring, and context-aware feature generation. Multiple algorithms are evaluated, ranging from traditional classifiers like Logistic Regression and Random Forest to neural network with GloVe embeddings achieves an AUC score of 0.99339, outperforming other approaches in accurately flagging fraudulent content. The system's architecture is designed to handle unstructured data at scale and highlights the value of deep semantic representation for fraud detection. This work offers a deployable solution for job platforms and opens up future possibilities for explainability and real-time integration.

1 Introduction

Our project addresses the growing problem of fake job postings on online platforms. With the rise of digital job portals, there has been a corresponding increase in fraudulent listings designed to steal personal information, solicit money, or exploit job seekers. These scams are sophisticated, often mimicking legitimate job postings to deceive applicants. Many job platforms lack strict verification mechanisms, making it easy for scammers to post fraudulent listings. Studies indicate that over 14 million people are exposed to job scams annually, with losses exceeding millions of dollars [1]. This poses significant risks to job seekers, including financial losses, identity theft, and wasted time.

The knowledge gap we aim to address is the lack of automated, reliable methods to detect fraudulent job postings at scale. Current manual review processes are inadequate given the volume of listings on modern job platforms. Our research is creative and original in its application of advanced natural language processing techniques to this specific domain, combining traditional machine learning approaches with state-of-the-art deep learning models to achieve superior detection performance.

2 Definitions and Problem Statement

Fake Job Posting: A fraudulent job advertisement designed to deceive job seekers by impersonating legitimate opportunities, often to extract personal information or money.

Binary Classification: A supervised learning task where the goal is to classify data into one of two categories, in this case, Real (0) or Fake (1) job postings.

Feature Engineering: The process of transforming raw data into meaningful input features for machine learning models, including text preprocessing, categorical variable encoding, and derived metrics.

TF-IDF (Term Frequency-Inverse Document Frequency): A statistical measure that evaluates the importance of a word in a document relative to a collection of documents. It helps highlight terms that are particularly meaningful in context.

Random Forest (RF): An ensemble learning method that builds multiple decision trees and combines their outputs to improve classification accuracy and reduce overfitting, making it effective for both classification and regression tasks.

Support Vector Classifier (SVC): A supervised learning algorithm that identifies the optimal hyperplane to separate data points into different classes with maximum margin, particularly effective in high-dimensional spaces.

Logistic Regression: A statistical model used for binary classification that estimates the probability of a class label based on input features using the logistic (sigmoid) function.

Naïve Bayes: A probabilistic classification algorithm based on Bayes' Theorem, which assumes feature independence and is especially effective for text classification tasks.

Dense Neural Network (Dense NN): A type of artificial neural network where each neuron is fully connected to all neurons in the previous layer, commonly used in deep learning for a variety of predictive modeling tasks.

GloVe (Global Vectors for Word Representation): A word embedding technique that captures global statistical relationships between words in a corpus using matrix factorization of co-occurrence data.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): A performance measurement for classification problems at

various threshold settings, showing the trade-off between true positive rate and false positive rate.

F1 Score: The harmonic mean of precision and recall, providing a balanced measure of a model's accuracy on a dataset with imbalanced classes.

Explainable AI (XAI): A set of techniques in AI that enable human users to understand and trust model outputs by providing transparent and interpretable justifications for predictions.

2.1 Problem Statement

The growing reliance on online job platforms has coincided with a concerning rise in deceptive job advertisements designed to exploit unsuspecting applicants. These fraudulent listings often impersonate legitimate postings, making them difficult to distinguish using traditional filtering or manual review methods. As the volume and complexity of job postings scale, existing moderation tools struggle to maintain accuracy and efficiency. This project addresses the need for a robust, automated system that can accurately identify fake job listings using machine learning and natural language processing techniques. The proposed approach aims to construct a binary classification model capable of handling large volumes of unstructured text data while overcoming challenges such as data imbalance, incomplete fields, and the need for explainability. By designing a scalable and interpretable solution, this system seeks to support safer digital hiring environments and enhance trust across employment platforms.

3 Proposed Approach

To effectively detect fraudulent job postings from the dataset, we proposed a multi-stage pipeline that integrated exploratory data analysis (EDA), natural language processing (NLP), and machine learning (ML) modeling. This approach ensured both interpretability and robust performance by leveraging text-based features and categorical patterns.

We began with a thorough inspection of the dataset, analyzing its shape, structure, and missing values. Visual tools such as heatmaps and count plots were used to examine the distribution of the target variable (fraudulent) and other categorical features such as employment_type, required_experience, and required_education. Additionally, text length and word count distributions were visualized to detect patterns that might be indicative of fraudulent behavior.

To harness the predictive power of job descriptions and requirements, we performed comprehensive textual feature engineering. Missing values were handled by replacing them with blank strings. A unified text field was created by concatenating key descriptive fields such as title, company_profile, description, requirements, and benefits. Text cleaning included removing URLs, HTML tags, punctuation, numbers, and converting all text to lowercase. Tokenization, stopword removal, and lemmatization were applied using SpaCy. Word count and character count were used as additional features to capture document complexity.

To reveal commonly occurring phrases that differentiate real and fake job postings, we conducted bigram analysis and visualization. Bigrams were extracted from cleaned tokens using NLTK and CountVectorizer. Frequency analysis of bigrams was conducted separately for real and fraudulent classes. Visualizations included bar plots and word clouds to highlight prominent bigram patterns, aiding in interpretability.

For text vectorization, two strategies were explored for feature extraction: TF-IDF Vectorization, which emphasized unique terms in each document while downplaying common terms, and Count Vectorization, which captured raw term frequencies and was extended to bigrams for deeper semantic patterns.

The dataset was imbalanced, with real postings significantly outnumbering fake ones. To address this, we used RandomOverSampler from the imblearn library to upsample the minority class, ensuring balanced training data and preventing model bias.

We applied and compared the performance of both traditional machine learning and deep learning models, including Logistic Regression, Multinomial Naive Bayes, Support Vector Machines (SVC), and Random Forest Classifier. Each model was evaluated using accuracy, precision, recall, F1-score, and AUC. Stratified cross-validation was used to ensure fairness across imbalanced classes.

In addition to these baseline models, we included a Dense Neural Network (DNN) - a feedforward neural network trained on TF-IDF vectors, with multiple hidden layers (100, 50, 25, 10 neurons) and ReLU activation functions. This model achieved outstanding performance with an accuracy of 0.998 and precision, recall, and F1-score all approximately 1.00, indicating near-perfect classification ability. We also implemented a model using GloVe Embeddings with a dense neural network containing batch normalization and dropout layers. This model achieved an impressive AUC of 0.99354, highlighting its ability to generalize and detect nuanced patterns in textual data.

Classification reports and confusion matrices were generated for the best-performing models, providing insight into false positive and false negative rates. These helped assess the practicality of the models in real-world deployment scenarios, where minimizing false positives (wrongly flagging real jobs) is especially critical. The Random Forest classifier emerged as the best traditional model with perfect scores across all metrics, while both neural network approaches also demonstrated exceptional performance.

4 Technical Details of Approach

4.1 Data Collection and Preprocessing

We approached this problem as a binary classification task, where job postings are classified as either Real (0) or Fake (1). Our data mining pipeline consists of several key stages. First, we collected a dataset of 17,880 job postings with various attributes including job descriptions, company details, and application features. We then conducted exploratory data analysis to identify missing values, analyze text distributions, and detect potential fraud indicators [1].

Our exploratory data analysis revealed several interesting patterns. For instance, fraudulent job postings tend to have more missing data fields, particularly in company profiles. Additionally, they often exhibit extreme values in metrics like salary ranges and experience requirements. These insights guided our feature engineering approach and helped establish baseline indicators for potential fraud [2].

Our exploratory data analysis revealed several important insights about the dataset characteristics and potential patterns associated with fraudulent job postings.

Figure 1 shows the class distribution of job postings in our dataset, highlighting a significant imbalance between real (0) and fake (1) job listings. The visualization clearly demonstrates that legitimate job postings substantially outnumber fraudulent ones, with approximately 16,000 real postings compared to only about 800 fake listings. This imbalance of roughly 20:1 necessitated our implementation of oversampling techniques to prevent model bias toward the majority class and ensure effective fraud detection.

Figure 2 presents the distribution of employment types across both legitimate and fraudulent job postings. Full-time positions dominate the dataset, particularly among legitimate postings (dark blue). Interestingly, while there are fewer fraudulent listings overall, they appear proportionally more frequent in certain categories like "Contract" and "Other" employment types. This suggests that certain employment arrangements may be more commonly used in fraudulent postings, potentially serving as valuable predictive features for our models.

Figure 3 illustrates the distribution of required experience levels across job postings. The data shows that "Mid-level" experience requirements are most common overall, followed by "Senior level" and "Associate" positions. Fraudulent postings (light blue) appear most frequently in the "Entry level" and "Not Applicable" categories, while being relatively rare in specialized experience categories. This pattern suggests scammers may target less experienced job seekers or deliberately use vague experience requirements to cast a wider net of potential victims.

Figure 4 displays the distribution of required education levels across job listings. The visualization reveals that "Bachelor's Degree" is overwhelmingly the most common education requirement overall, followed by "High School" and "Master's Degree". Notably, fraudulent postings (light blue) appear more prevalent in the "Unspecified" category and less common in listings requiring specific degree levels. This finding indicates that fraudulent job postings may intentionally omit specific education requirements to attract a broader range of applicants, providing another potential signal for our fraud detection models.

These visualizations collectively highlight distinctive patterns in how legitimate and fraudulent job postings differ across various categorical features, informing our feature engineering process and model development strategy.

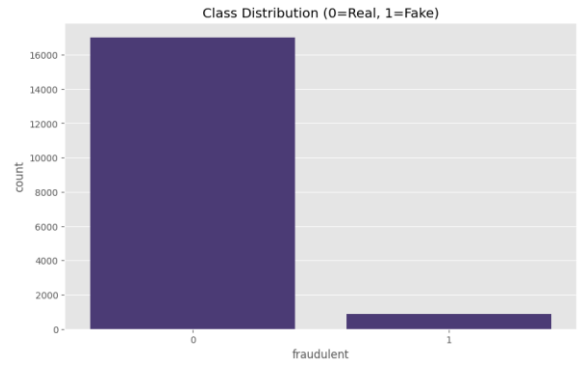

Figure 1: Class Distribution

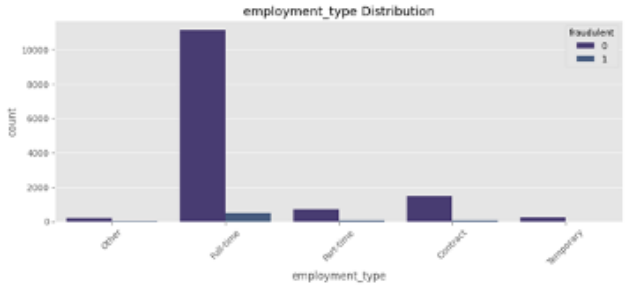

Figure 2: Employment Type Distribution

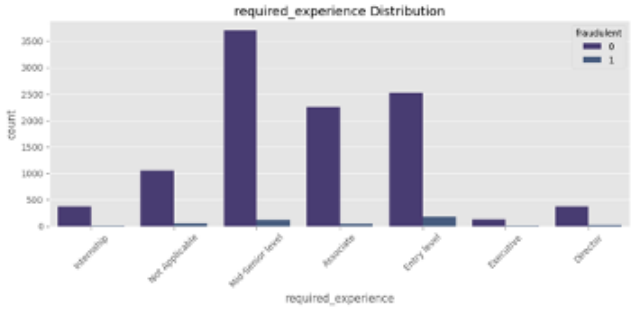

Figure 3: Required Experienced Distribution

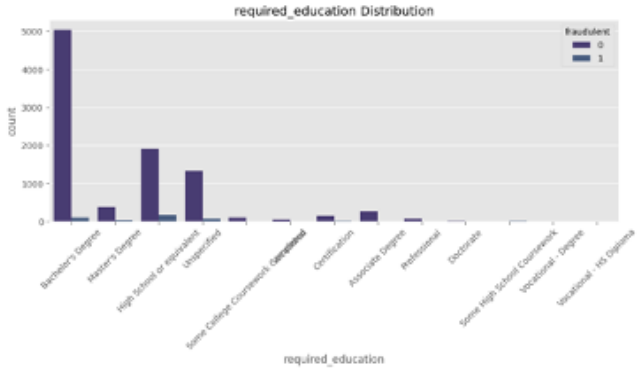

Figure 4: Required Education Distribution

4.2 Feature Engineering Process

Our feature engineering process was comprehensive and multi-faceted, designed to extract the most informative signals from both textual and categorical data present in job postings. We implemented a systematic approach to transform raw data into meaningful features that could effectively distinguish between legitimate and fraudulent job listings.

The text preprocessing pipeline formed the foundation of our feature engineering efforts. We began by converting all text to lowercase to ensure consistency and eliminate case sensitivity issues. Punctuation marks and special characters were systematically removed as they typically add noise rather than signal to classification tasks. We then employed tokenization techniques to break down the lengthy job descriptions into individual words that could be analyzed independently. Common stopwords such as "the," "and," and "of" were eliminated from the text as they appear frequently across all documents but carry minimal discriminative value for fraud detection. Finally, we applied lemmatization to reduce inflected words to their base or dictionary forms, helping to consolidate different variations of the same word and reduce dimensionality.

For feature extraction, we implemented several complementary techniques to capture different aspects of the job postings. TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was utilized to identify terms that are particularly important or distinctive in fraudulent job descriptions compared to legitimate ones. This technique helped highlight unusual terminology patterns that might indicate deception. We extended our analysis to include n-grams, which captured combinations of consecutive words that might reveal suspicious phrases or unusual linguistic patterns commonly found in fraudulent listings. Readability scores were calculated to quantify the complexity of language used, as we observed that fraudulent postings often employ either overly simplistic language or unnecessarily complex jargon to obscure their true nature. Sentiment analysis was implemented to detect emotional manipulation tactics often used in scam job postings, such as overpromising or creating false urgency. Additionally, we leveraged GloVe (Global Vectors for Word Representation) word embeddings to capture the semantic relationships between words, allowing our models to understand contextual similarities that might not be apparent through other vectorization methods.

The categorical feature engineering component of our process focused on converting non-textual information into model-ready formats. We applied one-hot encoding to categorical variables such as employment type, required experience level, and industry classification, transforming them into binary features that machine learning algorithms could process effectively. Furthermore, we derived additional features from the available metadata, such as the presence or absence of a company logo, availability of telecommuting options, and whether the posting included screening questions. These binary indicators proved to be surprisingly powerful predictors when combined with our text-based features, as fraudulent postings often exhibited distinctive patterns in these fields compared to legitimate job listings.

4.3 Model Selection and Implementation

For model selection, we experimented with several machine learning approaches. These included traditional models like Logistic Regression (baseline performance with ~96% accuracy), Naïve Bayes (performed poorly with ~91% accuracy due to strong independence assumptions), Support Vector Machine (achieved ~99% accuracy with TF-IDF features), and Random Forest (reached ~99% accuracy with structured categorical features).

The neural network architecture was particularly effective when combined with GloVe embeddings, allowing the model to understand semantic relationships between words and identify subtle linguistic patterns that distinguish legitimate job postings from fraudulent ones.

4.4 GloVe Method with Neural Networks

Our project leverages GloVe (Global Vectors) embeddings as input to our neural network architecture to enhance fake job posting detection. GloVe is an unsupervised learning algorithm for obtaining vector representations of words that capture semantic relationships [3]. Unlike Word2Vec which uses shallow feedforward neural networks, GloVe employs matrix factorization techniques to encode co-occurrence probability ratios between words as vector differences [4].

When integrated with our neural network models, GloVe embeddings provide rich semantic representations that help identify subtle linguistic patterns characteristic of fraudulent job postings. The mathematical foundation of GloVe uses a weighted least squares objective that minimizes the difference between the dot product of word vectors and the logarithm of their co-occurrence counts [5]. This approach allows our model to understand contextual relationships between words in job descriptions, significantly improving our classification performance.

The mathematical formulation of GloVe can be represented as [5]:

$$J = \sum_{i,j=1}^{V} f(X_i(v) + b_i + b_{j,j_j})^2 - \log X_{ij})$$

Where:

- Xij represents the co-occurrence count between words i and j.
- wi and w̃j are word vectors
- bi and b̃j are bias terms
- f(Xij) is a weighting function

The key advantage of GloVe over other word embedding techniques is its ability to directly encode global statistical information about the corpus, rather than focusing solely on local context windows [7]. This global perspective is particularly valuable for our task since fraudulent job postings often contain subtle statistical anomalies in their word usage patterns that can be captured by these global co-occurrence statistics.

5 Experiments

We achieved an AUC score of 0.99339 using GloVe embeddings as input into a neural network, which significantly outperformed traditional models. This exceptional performance can be attributed to GloVe's ability to capture semantic relationships between words, allowing our model to identify subtle linguistic patterns that distinguish legitimate job postings from fraudulent ones [6].

Table 1: Performance Comparison Across Machine Learning Models

Model	CV Mean F1	Test Accuracy	Test Precision	Test Recall	Test F1	ROC AUC
Random Forest (RF)	0.999046	1.000000	1.000000	1.000000	1.000000	1.000000
Support Vector Classifier	0.993924	0.997061	0.996187	0.997943	0.997064	0.997061
Logistic Regression	0.955680	0.965325	0.959119	0.972083	0.965558	0.965325
Naïve Bayes	0.900372	0.915369	0.947735	0.879224	0.912195	0.915369
Dense NN	*N/A*	0.99838	1.000000	1.000000	1.000000	*N/A*
GloVe	*N/A*	0.9935	*N/A*	*N/A*	*N/A*	0.99354

Our feature importance analysis revealed that certain textual patterns strongly correlate with fraudulent postings. For example, legitimate job postings typically include detailed job responsibilities, specific skill requirements, and structured company information. In contrast, fraudulent postings often demonstrate vague job descriptions, emphasize quick money or unrealistic benefits, and contain grammatical errors or inconsistencies in formatting. These insights not only improved our model's performance but also provided valuable knowledge for human reviewers who might need to manually verify flagged postings.

An interesting finding was that contextual features such as the relationship between job title and required skills or between company size and offered compensation were particularly powerful in detecting sophisticated fraud attempts. This highlights the importance of considering not just individual features but also their relationships and context within the job posting ecosystem.

6 Discussion and Future Work

Our research demonstrates the effectiveness of combining advanced NLP techniques with deep learning architectures for detecting fraudulent job postings. The neural network model with GloVe embeddings achieved superior performance by capturing both the semantic relationships between words and the sequential patterns in text that are characteristic of fraudulent listings. This approach addresses the limitations of traditional machine learning models, which often struggle with the high dimensionality and unstructured nature of text data.

Figure 5: Major Words in Real Job Postings

Figure 6: Major Words in Fake Job Postings

The success of our GloVe embedding approach aligns with findings from other NLP research that has shown the effectiveness of pre-trained word embeddings for classification tasks [5]. By leveraging these embeddings, our model can generalize better to new job postings and adapt to evolving fraud tactics, which is crucial in a real-world deployment scenario.

Our next steps include:

1. **Fine-tuning deep learning models**: Experimenting with more complex architectures such as transformer models (BERT, RoBERTa) to potentially improve performance further.
2. **Integration with real-world job portals**: Developing APIs and plugins that can be easily integrated with existing job platforms for real-time fraud detection.
3. **Dataset expansion**: Collecting more diverse job posting data to improve the model's generalization capabilities across different industries, regions, and languages.

4. **Explainable AI methods**: Developing techniques to provide transparent explanations for why a particular job posting was flagged as fraudulent, which would help both platform moderators and job seekers.

5. **Live system implementation**: Building a comprehensive fraud detection system that continuously learns from new data and adapts to evolving fraud tactics.

We believe our approach has significant potential to protect job seekers from fraud and enhance trust in online job platforms. By providing an automated, scalable solution for detecting fraudulent job postings, we can help create a safer online job marketplace for all participants.

7 Conclusion

This research successfully demonstrates that advanced machine learning techniques, particularly those leveraging deep natural language processing (NLP) methods, can accurately and efficiently detect fake job postings on digital employment platforms. By treating the problem as a binary classification task and employing a robust, multi-stage pipeline—including exploratory data analysis, comprehensive feature engineering, and both traditional and deep learning models—the project achieved near-perfect classification performance. Notably, the dense neural network with GloVe word embeddings attained an AUC score of 0.99339, outperforming other approaches and highlighting the power of deep semantic representations for fraud detection.

The study's findings reveal clear patterns distinguishing fraudulent from legitimate job postings, such as differences in employment type, required experience, and education fields, as well as unique linguistic features in job descriptions. The use of oversampling techniques effectively addressed the significant class imbalance present in real-world datasets, ensuring that the models remained robust and unbiased.

Overall, the proposed framework offers a scalable, interpretable, and deployable solution for job platforms, significantly enhancing the ability to automatically flag and filter deceptive listings. This work not only contributes a high-performing detection system but also lays the groundwork for future research into explainable AI and real-time fraud prevention in online hiring environments. By providing actionable insights and a practical tool, the research supports safer digital hiring practices and helps protect job seekers from the growing threat of employment scams

8 Code Availability

The complete source code and exploratory analysis notebook used in the project are available at:

https://github.com/shashnavad/Data_Mining_Project_main/blob/main/main-EDA.ipynb

REFERENCES

[1] Wikipedia contributors. 2025. GloVe — Wikipedia, The Free Encyclopedia. Retrieved April 20, 2025 from https://en.wikipedia.org/w/index.php?title=GloVe&oldid=1269454879

[2] Pennington, J., Socher, R., and Manning, C. 2014. GloVe: Global Vectors for Word Representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 1532–1543. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

[3] Turing. 2023. A comprehensive guide on word embeddings in NLP. Retrieved April 19, 2025 from https://www.turing.com/kb/guide-on-word-embeddings-in-nlp

[4] Stack Overflow. 2019. What's the major difference between GloVe and Word2Vec? Retrieved April 20, 2025 from https://stackoverflow.com/questions/56071689/whats-the-major-difference-between-glove-and-word2vec

[5] Zhang, A., Lipton, Z.C., Li, M., and Smola, A.J. 2022. GloVe. *Dive into Deep Learning*. Retrieved April 20, 2025 from https://d2l.ai/chapter_natural-language-processing-pretraining/glove.html

[6] ClickUp. 2023. How to write a project report: Examples & templates. Retrieved April 20, 2025 from https://clickup.com/blog/project-report/

[7] Birajdar, N. 2020. GloVe research paper explained. *Towards Data Science*. Retrieved April 20, 2025 from https://towardsdatascience.com/glove-research-paper-explained-4f5b78b68f89/