FraudSpotter: Job Posting Detection Using NLP & ML Models

Maureen Ekwebelem and YaeJin (Sally) Kang

Fordham University,

Masters of Data Science

***Abstract*—Fraudulent job postings are a growing problem on online recruitment platforms, exposing job seekers to scams and reducing trust in hiring systems. This project investigates how natural language processing (NLP) and machine learning classification models can be used to automatically detect fake postings. We preprocess and analyze textual data from job listings, extract relevant features, and train multiple classifiers to evaluate performance. Our approach includes DistilBERT, a lightweight version of BERT, alongside traditional machine learning models to distinguish between legitimate and fraudulent entries. Results show that simpler baseline models can outperform deeper language models on this dataset, highlighting both the potential and the limitations of automated fraud detection systems.**

***Index Terms*—Machine learning, DistilBERT, SHAP, job scams, classification models**

I. Introduction

Employment scams have escalated sharply in recent years, aided by Artificial Intelligence. Reports of such scams rose by 118% in 2023 compared to the prior year, according to the Identity Theft Resource Center. The Federal Trade Commission reported that consumers lost $367 million to job and business opportunity scams in 2022—a 76% increase from the year before—with the typical victim losing around $2,000. Beyond financial losses, applicants face long-term risks such as identity theft when personal data, including Social Security numbers or bank account details, is exposed [1].

Fraudulent job postings on online recruitment sites create significant reputational and operational risks. They erode public trust in the companies being impersonated, while also damaging the credibility of the platforms that host them. Smaller or lesser-known businesses are especially vulnerable, as even limited exposure to scams can discourage applicants and hinder their growth. Over time, the persistence of these scams reduces confidence in third-party hiring platforms such as Indeed, LinkedIn, and Glassdoor, making users less willing to rely on them for legitimate opportunities [2].

This project applies BERT-based embeddings, DistilBert, to generate contextualized text representations that capture nuanced meanings in job postings. These embeddings will then be combined with machine learning classifiers to improve classification efficiency and enhance the accuracy of distinguishing legitimate opportunities from fraudulent ones. We also integrate model explainability techniques such as SHAP (Shapley Additive Explanations) to highlight the words that contribute most to the model's predictions, providing insights into which textual patterns are most strongly associated with fraudulent postings.

A. Significance

The current methods for detecting fake job postings often fall short. Keyword-based filtering is easily circumvented by scammers who can simply rephrase their listings. While human moderation offers higher accuracy, it is both costly and slow to scale. Traditional models that rely on word frequency or bag-of-words representations also miss the deeper semantic cues that distinguish legitimate from fraudulent postings. Scammers frequently mimic authentic job descriptions but embed subtle red flags such as vague company information, unrealistic requirements, or urgent calls to action that rule-based systems struggle to detect. For example, words like "remote" or "flexible" are not inherently suspicious; their meaning and risk level depend heavily on the surrounding context, which simple models cannot adequately capture.

Our approach improves upon traditional machine-learning baselines by integrating modern NLP techniques with interpretable modeling. While the baseline Logistic Regression model provides a strong foundation using structured features and TF-IDF text signals, we extend this by leveraging DistilBERT embeddings, which transform job descriptions and company profiles into dense contextual representations that capture deeper semantic patterns beyond simple word frequencies. To ensure transparency in model behavior, we apply SHAP to the baseline Logistic Regression model where feature meanings remain interpretable to identify which textual and structured indicators most strongly influence fraud predictions. By combining the interpretability of the baseline model with the improved semantic understanding of DistilBERT, our system achieves stronger robustness to subtle or context-dependent scam language while still offering clear, human-readable explanations. This hybrid design makes the approach suitable for real-world deployment.

This project demonstrates how statistical NLP methods and machine learning models can be combined to detect fraudulent job postings more effectively than keyword filtering alone. It highlights both the strengths and limitations of these approaches and contributes to understanding how LLMs may be applied responsibly in online recruitment security. Crucially, the system addresses the harms of job scams. By enabling large-scale, automated classification of postings, this approach supports safer, more trustworthy hiring pipelines and has potential extensions to other domains where online fraud detection is essential.

II. Related works

Each of the following studies utilized the Real or Fake Job Posting Prediction dataset described in the Dataset section, applying various modeling and preprocessing techniques to detect fraudulent job postings.

Pillai (2023) explored the use of deep learning techniques to detect fraudulent job postings [2]. The study first applied CountVectorizer for exploratory text analysis, followed by a TextVectorization layer with an Embedding layer to prepare the text data for modeling. Categorical variables were one-hot encoded, and preprocessing included cleaning text by removing symbols, punctuation, HTML tags, and extra spaces, followed by tokenization. A Bidirectional LSTM (Bi-LSTM) was used as the primary model, with ensemble models like Random Forest, LightGBM, and GBM tested for comparison. The project's main contribution was integrating text and numeric features within one Bi-LSTM framework to capture context in job descriptions, achieving 98.71% accuracy and a 0.91 ROC-AUC score.

Boka et al. (2024) referred to the dataset as the EMSCAD , which is another name for the Kaggle dataset. The study focused on improving model efficiency and interpretability by converting text-based features into their categorical forms instead of using complex NLP methods [3]. Preprocessing involved removing stop words, extra spaces, and attributes such as department and salary_range, which contained large amounts of missing data. The authors compared several supervised models, including Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) from scikit-learn's library. Logistic Regression achieved the highest accuracy (98.37%), supporting the conclusion that simple and interpretable models can perform well for fake job detection.

Vu et al. (2024) developed an enhanced framework called NLP2FJD to improve fake job description (FJD) detection using deep learning-based NLP techniques [4]. Preprocessing involved text cleaning, stopword removal, tokenization, and one-hot encoding for categorical variables such as employment type, required education, and industry. Text was represented using Word2Vec embeddings, and Convolutional Neural Networks (CNNs) were applied to capture local phrase patterns. The extracted text representations were then fused with structured features for final classification. Their model achieved 96% accuracy and an AUC of 0.984, outperforming Logistic Regression and other baselines. The study's main contribution was demonstrating that integrating text and job metadata in a unified deep learning model, and tuning decision thresholds to reduce false positives, significantly enhances detection performance.

Hanif et al. (2024) developed a machine learning framework to predict fraudulent job advertisements [5]. Their study evaluated multiple supervised learning algorithms on both imbalanced and balanced datasets to determine the most effective classification approach. Text preprocessing involved removing duplicate records, cleaning HTML tags and URLs, converting text to lowercase, removing punctuation, stop words, and numerical values, followed by stemming. The authors experimented with three feature extraction techniques: Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and Hash vectorizer. To address severe class imbalance, they applied Random Undersampling (RUS) to create a balanced dataset. Each model was then trained and tested using both dataset versions and evaluated using accuracy, precision, recall, F1-score, and AUC. The Decision Tree with BoW vectorization on the balanced dataset achieved the best performance (accuracy = 0.705, F1 = 0.71, AUC = 0.68), demonstrating that simpler models can effectively classify fake job postings when data balance and appropriate feature representation are ensured. Their methodology highlights the importance of comparing multiple vectorization and resampling strategies for robust fraud detection. We will incorporate similar methodologies by comparing multiple text vectorization and class balancing techniques to evaluate how they perform alongside our BERT-based model for robust and fair fraud detection.

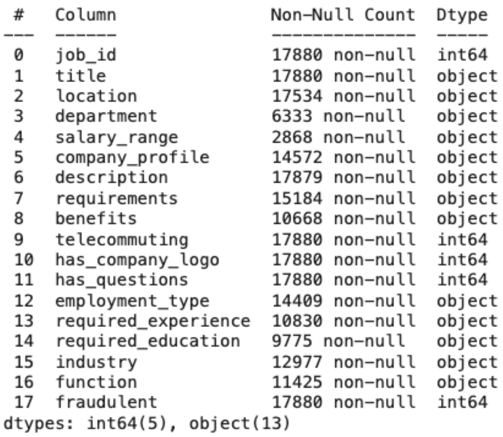
Taneja et al. (2025) introduced Fraud-BERT, a transformer-based model for detecting fraudulent job postings [6]. Their study provided useful methodological insights that inspired parts of our own analysis. In their work, the dataset was referred to as EMSCAD. The authors combined all relevant textual fields (such as job title, description, requirements, and benefits) into a single text column for BERT input, while removing low-correlation structured attributes such as job_id, salary_range, telecommuting, has_company_logo, and has_questions. Preprocessing included text cleaning and tokenization using BERT's WordPiece tokenizer, with sequences standardized for BERT input. Fraud-BERT was fine-tuned using the pre-trained BERTbase (uncased) model and compared against Logistic Regression, Support Vector Machine (SVM), Naive Bayes, Random Forest, and Bi-LSTM. The "uncased" version of BERT ignores capitalization, treating words like "Job" and "job" as the same token, which simplifies training on noisy real-world text data. By leveraging BERT's bidirectional attention and contextual embeddings, the model was able to interpret subtle language patterns and contextual cues in job descriptions more effectively than traditional models. It achieved 98.45% accuracy, along with high precision, recall, and F1-scores, all without the use of oversampling techniques. This study was the first to apply BERT to fake job detection and demonstrated that transformer-based contextual representations significantly outperform conventional feature-based or deep learning approaches. Building on this work, the authors proposed exploring lighter transformer variants such as DistilBERT, which preserves approximately 97% of BERT's language understanding capacity while being 60% faster and smaller. Accordingly, our study implements DistilBERT to balance performance with computational efficiency.

Building on insights from prior research, our project implements Logistic Regression, K-Nearest Neighbors (KNN), and XGBoost, evaluating their performance based on accuracy and precision, as done in earlier studies. In addition, we incorporate the DistilBERT-based transformer model discussed above to capture contextual semantics in job descriptions and company profiles. Our approach draws methodological inspiration from prior multi-feature fusion frameworks that integrate both textual and structured attributes, while extending them through the inclusion of context-aware embeddings and comparative evaluation across traditional machine learning and transformer-based models.

III. Methodology

A. Dataset

Our project uses the Real or Fake Job Posting Prediction from Kaggle (originated from the Employment Scam Aegean Dataset (EMSCAD), compiled by the University of the Aegean). The dataset contains approximately 17,880 job postings, of which around 800 are labeled as fake, making it imbalanced but realistic for fraud detection . The dataset includes 18 features combining structured and unstructured data, for example, job title, location, company profile, description, requirements, benefits, and other meta fields (Fig. 1). Of these, five are numerical columns and thirteen are text-based columns. The target, fraudulent, uses a binary label (0=real, 1=fake). Because it includes both textual and categorical/metadata attributes, it supports a variety of modeling approaches. We will use an 80/20 train-test split for evaluation.



```
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   job_id               17880 non-null  int64
 1   title                17880 non-null  object
 2   location             17534 non-null  object
 3   department           6333 non-null   object
 4   salary_range         2868 non-null   object
 5   company_profile      14572 non-null  object
 6   description          17879 non-null  object
 7   requirements         15184 non-null  object
 8   benefits             10668 non-null  object
 9   telecommuting        17880 non-null  int64
 10  has_company_logo     17880 non-null  int64
 11  has_questions        17880 non-null  int64
 12  employment_type      14409 non-null  object
 13  required_experience  10830 non-null  object
 14  required_education   9775 non-null   object
 15  industry             12977 non-null  object
 16  function             11425 non-null  object
 17  fraudulent           17880 non-null  int64
dtypes: int64(5), object(13)
```

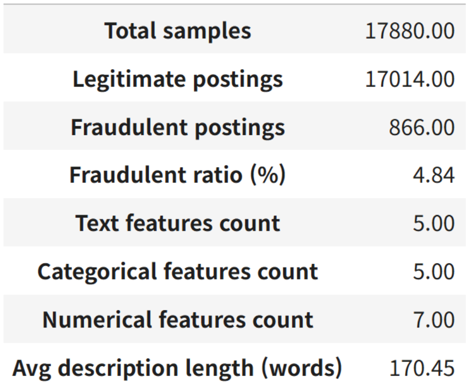
Figure 1. Columns and their data types

B. Data cleaning & Preprocessing

Missing data were handled by dropping columns with more than 80% missing values (salary_range column removed). For other columns with less than 80% missing data, missing entries were filled with empty strings before analysis to ensure that text-based vectorizers such as BERT would not fail. We combined all text fields into one column and applied standard preprocessing removing HTML tags, URLs, punctuation-like artifacts, and excessive whitespace to produce a clean, normalized text feature. We then applied one-hot encoding to convert categorical variables (such as employment type, required experience, and industry) into numerical format so they could be integrated into the machine learning models.

The summary of dataset features can be found in Fig 2. Given the substantial class imbalance of approximately 18,000 real postings versus 800 fraudulent ones, we applied stratified random resampling and class-sensitive learning through adjusted class weights. This technique preserves the original class distribution across folds and prevents classifiers from defaulting to majority-class predictions, enabling more reliable fraud detection.

Textual features were vectorized using TF-IDF and standardized for models such as Logistic Regression and KNN, which rely on uniform feature scaling.



Total samples	17880.00
Legitimate postings	17014.00
Fraudulent postings	866.00
Fraudulent ratio (%)	4.84
Text features count	5.00
Categorical features count	5.00
Numerical features count	7.00
Avg description length (words)	170.45

Figure 2. Summary of dataset features; 17 of 18 included in analysis (*salary_range* excluded due to its high proportion of missing values)

C. Model Implementation

Following data cleaning and preprocessing, we employed DistilBERT, a pretrained transformer-based language model, to generate contextual embeddings for textual features such as job descriptions, company profiles, and job requirements. Unlike traditional frequency-based representations (e.g., TF-IDF or Bag of Words), DistilBERT captures deeper semantic and contextual relationships between words, enabling the model to recognize subtle linguistic cues in fraudulent job descriptions.

For each job posting, DistilBERT converted the text into a single numerical vector that captured its overall meaning. Text data were first tokenized using the DistilBERT tokenizer, which applied subword segmentation, lowercasing, and truncation to a fixed length of 128 tokens. The tokenized inputs were then processed by the pretrained DistilBERT model in evaluation mode, and the mean of the last hidden layer's outputs was extracted to produce a multi-dimensional embedding for each posting. These text representations were then combined with one-hot-encoded categorical features (such as job type or required education), producing a unified feature set that integrated both the textual and structured aspects of each job listing.

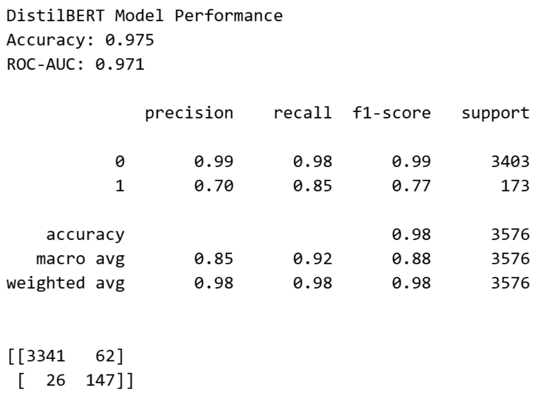
We compared the DistilBERT model's performance to three machine learning classifier models: Logistic Regression, KNN, and XGBoost to establish performance benchmarks. For our models more sensitive to large variations in magnitude (Logistic Regression and KNN), features were scaled prior to model training.

The highest-performing model was then used for SHAP (SHapley Additive exPlanations) analysis. SHAP was applied to identify the features that most strongly influenced the model's predictions. This analysis provided transparency into the decision-making process and highlighted the key factors that distinguished legitimate from fraudulent job postings.

Model performance was evaluated using metrics such as accuracy, precision, recall, F1-score, and by assessing the confusion matrix.

IV. Results

We first evaluated a contextual NLP model using DistilBERT embeddings combined with structured metadata. This model achieved 97.5% accuracy, 0.971 ROC–AUC, and a recall of 0.85 for fraudulent postings, showing strong performance on context-dependent scam patterns (Fig. 3).

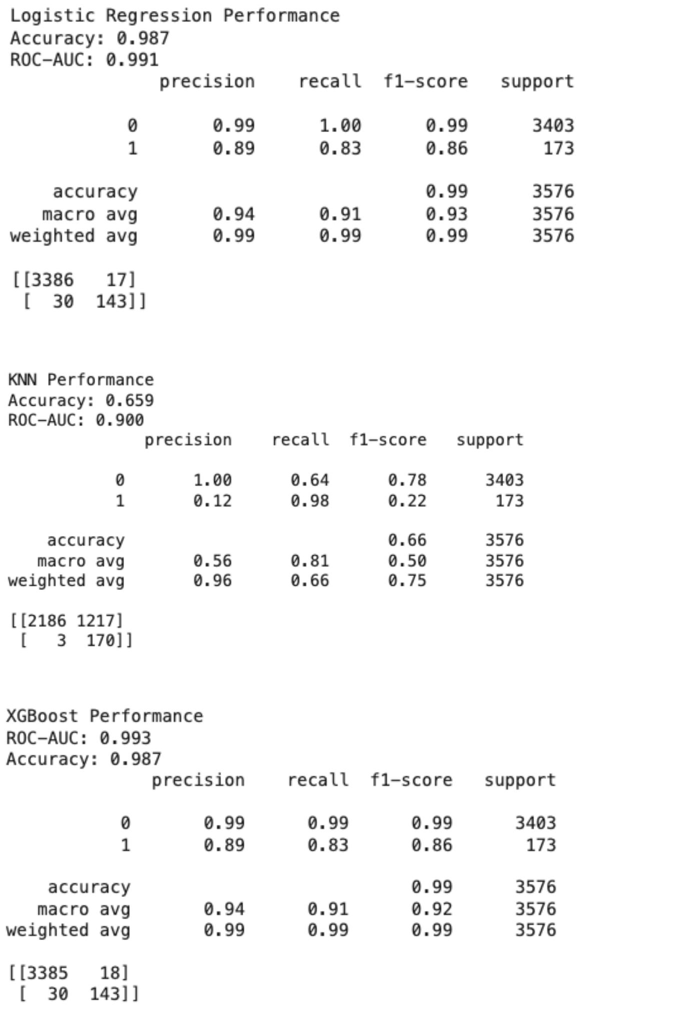


```
DistilBERT Model Performance
Accuracy: 0.975
ROC-AUC: 0.971

              precision    recall  f1-score   support

           0       0.99      0.98      0.99      3403
           1       0.70      0.85      0.77       173

    accuracy                           0.98      3576
   macro avg       0.85      0.92      0.88      3576
weighted avg       0.98      0.98      0.98      3576


[[3341   62]
 [  26  147]]
```

Figure 3. DistilBert Model Performance

We evaluated three baseline machine learning models (K Nearest Neighbors (KNN), XGBoost, and Logistic Regression) using the combined TF-IDF and structured feature representation. Logistic Regression demonstrated the strongest and most reliable performance, achieving 98.7% accuracy, 0.99 precision, 0.83 recall, an F1 score of 0.86, and a ROC-AUC of 0.991 (Fig 4).

XGBoost performed comparably with 98.7% accuracy, 0.99 precision, 0.83 recall, an F1 score of 0.86, and a ROC-AUC of 0.993, although the slightly lower recall still indicates a higher tendency to miss fraudulent postings (Fig 4). The Logistic Regression model produced one fewer false positive than XGBoost, as shown in the confusion matrix. KNN displayed noticeably lower performance across all metrics, with ~ 66% accuracy, 0.12 precision, 0.98 recall, an F1 score of 0.22, and a ROC-AUC of 0.900. Overall, both Logistic Regression and XGBoost were effective baselines, but Logistic Regression offered the optimal combination of predictive performance and interpretability for deeper analysis. Therefore, although XGBoost also performed strongly, we selected Logistic Regression for SHAP explainability because its slightly higher overall performance and linear structure provide clearer, more directly interpretable feature attributions, making it easier to understand which text and metadata signals contribute most to fraud predictions.



```
Logistic Regression Performance
Accuracy: 0.987
ROC-AUC: 0.991
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      3403
           1       0.89      0.83      0.86       173

    accuracy                           0.99      3576
   macro avg       0.94      0.91      0.93      3576
weighted avg       0.99      0.99      0.99      3576

[[3386   17]
 [  30  143]]
```

```
KNN Performance
Accuracy: 0.659
ROC-AUC: 0.900
              precision    recall  f1-score   support

           0       1.00      0.64      0.78      3403
           1       0.12      0.98      0.22       173

    accuracy                           0.66      3576
   macro avg       0.56      0.81      0.50      3576
weighted avg       0.96      0.66      0.75      3576

[[2186 1217]
 [   3  170]]
```

```
XGBoost Performance
ROC-AUC: 0.993
Accuracy: 0.987
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      3403
           1       0.89      0.83      0.86       173

    accuracy                           0.99      3576
   macro avg       0.94      0.91      0.92      3576
weighted avg       0.99      0.99      0.99      3576

[[3385   18]
 [  30  143]]
```

Figure 4. Summary of Baseline ML Model Performance

The SHAP summary plot displays the top twenty text features influencing the Logistic Regression model (Fig 5). Features such as *"companies," "growing," "team,"* and *"clients"* show the highest overall impact, with SHAP values mainly clustered near zero and extending up to roughly +2 . On the negative side, several features reach values near –3, and *"sign"* presents a single extreme outlier around –6, representing the strongest negative contribution observed. Overall, these features reflect the primary drivers of variation in the model's predictions across job postings.

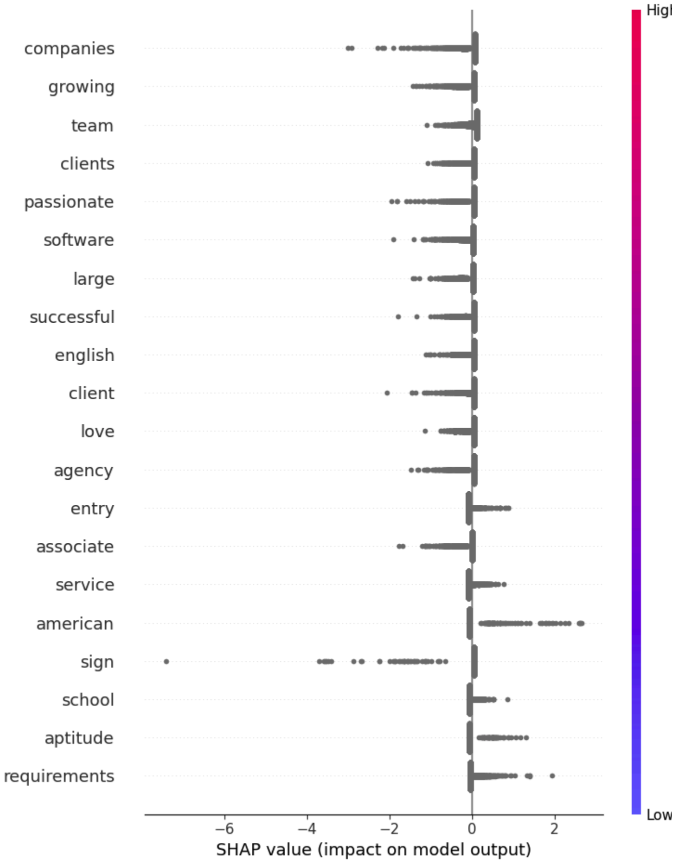


Figure 5. SHAP results from the Logistic Regression Model. Positive SHAP values indicate legitimacy of job postings, while negative values indicate fraud.

V. Discussion

This project compares a contextual NLP approach using DistilBERT to three baseline machine learning models for detecting fraudulent job postings. DistilBERT, paired with Logistic Regression and structured metadata, showed strong performance and demonstrated that transformer-based embeddings can capture deeper semantic meaning and more subtle fraud cues than traditional text representations. Although we initially expected the contextual model to be the top performer due to its ability to model language context, Logistic Regression ultimately outperformed it.

Logistic Regression emerged as the strongest performer, detecting fraudulent postings accurately while minimizing false positives.

A likely reason for this is that many of the fraud signals in our dataset are expressed through clear patterns in wording and job structure, which a linear model can capture effectively. This suggests that, although scammers sometimes disguise intent, a large portion of fraudulent postings still rely on obvious or repetitive language cues that do not require a deeper semantic model to identify.

In comparison, XGBoost did not provide a meaningful advantage over Logistic Regression, despite being a more complex model capable of capturing nonlinear patterns. Its similar performance suggests that additional complexity was not necessary for this task.

Meanwhile, K-Nearest Neighbors struggled considerably with the high-dimensional and sparse TF-IDF feature space, leading to poor recall for the fraudulent class and demonstrating that instance-based methods are not well suited for text fraud detection in this context.

Because Logistic Regression delivered the best combination of performance and interpretability among all models tested, including the embedding based pipeline, it was selected for SHAP analysis to better understand how the system is making predictions. The SHAP summary plot shows that most influential features are clustered close to zero on the SHAP axis, indicating that individual words provide small but consistent contributions when combined with other information. Words such as *“companies”, “team”, “clients”*, and **“***growing***”** tended to show more negative SHAP values, meaning they push predictions toward the legitimate class. These terms are common in postings that clearly reference real work environments with defined responsibilities. The most notable feature is **“***sign***”**, which shows a strong negative outlier near minus six. Although negative values typically suggest legitimacy, this extreme spread indicates that when “sign” appears in certain contexts, such as pressure to sign agreements quickly, it becomes a highly important signal the model uses in combination with other cues.

In contrast, several features extend farther into the positive SHAP range, indicating a stronger influence toward predicting fraud when they appear without supporting context. Words like *“american”, “entry”, “requirements”,* and *“aptitude”* often show higher SHAP values, reflecting language that may appear generic or disconnected from specific job duties, which increases suspicion.

Importantly, no single feature determines whether a posting is labeled as fraudulent. Instead, the model weighs multiple linguistic signals across the description to estimate risk. The alignment between these learned patterns and real scam tactics reinforces that Logistic Regression is identifying meaningful and interpretable indicators rather than relying on random correlations.

Overall, these results demonstrate that Logistic Regression provides an optimal balance of performance, efficiency, and explainability for automated detection of fraudulent job postings. Meanwhile, the DistilBERT-based model shows promising advantages in modeling contextual and semantic signaling of fraud, pointing toward future enhancements that can improve detection of sophisticated or emerging scam patterns. Together, these findings support the development of accurate and transparent fraud detection systems that help protect users on job platforms while maintaining trust in automated screening tools.

This project has several limitations that may influence the interpretation and generalizability of the results. First, the dataset was overly imbalanced and included a relatively small number of confirmed fraudulent postings, which required oversampling and may limit how well the model generalizes to other job platforms. Additionally, although the models performed well, hyperparameters were not selectively optimized, and the DistilBERT model was not fine tuned to job posting language as fully as intended. This was primarily due to time and GPU constraints, which limited our ability to perform a more extensive transformer based training process.

VI. Conclusion

This project compares contextual NLP and traditional baseline machine learning approaches for detecting fraudulent job postings. Logistic Regression achieved the strongest and most interpretable performance, showing that many key fraud indicators in this dataset can be captured through straightforward linguistic and structural patterns. The DistilBERT based model demonstrated potential for capturing more subtle and context-dependent scam language, although it did not outperform the simpler baseline under our non-fine-tuned setup. SHAP analysis confirmed that the logistic regression model relies on meaningful cues that align with real scam tactics, reinforcing confidence in using explainable models for this task. In future work, fully fine tuning DistilBERT or other transformer models on job posting data may better leverage contextual understanding and improve detection of more sophisticated scams. Incorporating more balanced datasets could further improve the system as fraud behaviors continue to evolve. Lastly, exploring explainability tools that examine contextual word phrasing rather than the word by word scoring used in SHAP could provide more meaningful insights for job fraud detection.

References

[1] Identity Theft Resource Center, "2023 trends in identity report: Identity Theft Resource Center sees 118 percent increase in job scams," Jun. 26, 2024. [Online]. Available: https://www.idtheftcenter.org/post/2023-trends-in-identity-report-118-percent-job-scam-increase/

[2] M. Boka, "Predicting fake job posts using machine learning models,"
SSRN, Apr. 4, 2024. [Online]. Available:
https://ssrn.com/abstract=4807145

[3] V. Dinh-Hong, K. Nguyen, K. T. Tran, B. Vo, and T. Le, "Improving fake job description detection using deep learning (NLP2FJD),"
J. Inf. Telecommun., vol. 8, no. 2, pp. 238–260, 2024.
doi: 10.1080/24751839.2024.2387380.

[4] A. Agarwal and R. Sharma, "A novel online job scam detection of imbalanced data using machine learning and NLP models," ResearchGate,Preprint,2024.[Online]. Available: https://www.researchgate.net/publication/377447842

[5] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: Transformer based context aware online recruitment fraud detection," Discover Computing, 2025. doi: 10.1007/s10791-025-09502-8.

[6] More REFERENCES- follow template below

[7] R. Fairbairns. (2014, Jun.) The TEX FAQ. [Online]. Available: http://www.tex.ac.uk/

[8] M. Sharpe. (2015, Jul.) The newtx package. [Online]. Available: http://www.ctan.org/pkg/newtx

[9] (2015, Jul.) Mathtime professional fonts. Personal TEX, Inc. [Online]. Available: http://www.pctex.com/mtpro2.html

[10] D. Carlisle and F. Mittelbach. (2015, Apr.) The bm package. [Online]. Available: http://www.ctan.org/pkg/bm

[11] (2013, Jan.) The amsmath package. The American Mathematical Society. [Online]. Available: http://www.ctan.org/pkg/amsmath

[12] S. Pakin. (2009, Apr.) The IEEEconf.cls package. [Online]. Available: http://www.ctan.org/pkg/ieeec