Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Explainable Job Category Detection Using DistilBERT: A Three-Class Transformer Model for Classifying Real, Fake, and Invalid Online Job Offers

Nnaemeka Ugwumba

nnaemeka_ugwumba@uniport.edu.ng

University of Port Harcourt https://orcid.org/0009-0000-2493-9846

Ugochi A. Okengwu

ugochi.okengwu@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Clement Barisua Ndeekor

clement.ndeekor@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Peter Sunday Jaja

peter_jaja@uniport.edu.ng

University of Port Harcourt, Port Harcourt, Nigeria

Research Article

Keywords: fake job posting detection, online recruitment fraud, DistilBERT, transformer models, multi-class text classification, explainable AI, SHAP, natural language processing

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11679641/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Explainable Job Category Detection Using DistilBERT: A Three-Class Transformer Model for Classifying Real, Fake, and Invalid Online Job Offers

Nnaemeka Kingsley Ugwumba[1] · Ugochi A. Okengwu[1] · Clement Barisua Ndeekor[1] · Peter Sunday Jaja[1]

[1]Department of Computer Science, University of Port Harcourt, Port Harcourt, Nigeria

Corresponding author: Nnaemeka Kingsley Ugwumba, nnaemeka_ugwumba@uniport.edu.ng

ugochi.okengwu@uniport.edu.ng · clement.ndeekor@uniport.edu.ng · peter_jaja@uniport.edu.ng

Abstract

Online job portals have widened access to employment, but they have also become a primary channel for Online Recruitment Fraud (ORF), in which fake job postings are used to extract money, harvest personal data, or recruit victims into illicit schemes. Most fake job detection systems treat the task as binary Real/Fake classification, which forces incomplete, nonsensical, or spam-like postings into the wrong class and makes them less reliable in practice. This study develops and evaluates an explainable, lightweight DistilBERT model that classifies online job offers as Real, Fake, or Invalid, and deploys it as a real-time web application. The distilbert-base-uncased classifier was fine-tuned on a corpus that combines the Employment Scam Aegean Dataset (EMSCAD), the job-scam domain of the DIFrauD benchmark, and a researcher-curated Invalid Token dataset. The combined corpus of 34,325 records was highly imbalanced (21.8:1), so it was rebalanced to 3,000 records per class through controlled undersampling and textual augmentation instead of SMOTE, and balanced class weighting was kept during training. The model was trained for four epochs with AdamW (learning rate 2×10^{-5}) and a 256-token sequence length on a stratified 70:15:15 train–validation–test split, with the best checkpoint chosen by validation loss, and was then served through a FastAPI backend with SHAP-based token-level explanations. On a held-out test set of 1,350 samples, the model achieved 98.74% accuracy and 98.74% balanced accuracy and misclassified only 17 samples. It reached F1-scores of 0.98, 1.00, and 0.98 and one-versus-rest ROC-AUC values of 0.998, 1.000, and 0.998 for the Fake, Invalid, and Real classes, respectively, and the deployed system returned predictions in about 1.0 to 2.9 seconds. These results show that a distilled transformer can deliver accurate, efficient, and interpretable three-class fake job detection, and that the explicit Invalid category and SHAP explanations take automated job offer screening beyond conventional binary approaches.

Keywords: fake job posting detection; online recruitment fraud; DistilBERT; transformer models; multi-class text classification; explainable AI; SHAP; natural language processing

1 Introduction

The global shift toward digital recruitment has fundamentally reshaped how employers and job seekers connect. Online job portals such as LinkedIn, indeed, and Glassdoor offer efficiency and unprecedented reach; however, the same openness that makes them valuable has allowed malicious actors to perpetrate Online Recruitment Fraud (ORF), a form of cybercrime in which deceptive job advertisements are used to target vulnerable job seekers (Vidros et al., 2017). Fake job postings are engineered to extract advance fees, harvest personal data for identity theft, or lure applicants into money-laundering and reshipping schemes, and they therefore cause substantial financial and psychological harm. Consumer-protection bodies continue to report that employment scams are among the most common and costly frauds affecting job seekers (Better Business Bureau, 2026; Federal Trade Commission, n.d.).

Early computational approaches to fake job detection relied on traditional machine learning classifiers such as Support Vector Machines, Random Forest, and Naïve Bayes, typically combined with Term Frequency–Inverse Document Frequency (TF-IDF) features extracted from benchmark corpora such as EMSCAD. Although these approaches achieved encouraging accuracy, they struggled with sophisticated postings that closely mimic genuine advertisements, because manual feature engineering

limits generalisation to evolving fraud tactics. Subsequently, transformer-based language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) brought a marked improvement in contextual language understanding, enabling models to capture the subtle linguistic cues that separate fraudulent from legitimate recruitment content. Nevertheless, large transformer architectures carry computational and memory costs that constrain their use in real-time, high-volume screening environments (Sanh et al., 2019).

A further limitation runs across the existing literature: the classification problem is almost always restricted to two categories, Real and Fake. This binary framing does not reflect the actual composition of online job content, a meaningful share of which is incomplete, poorly structured, corrupted, or nonsensical rather than deliberately fraudulent. Forcing such content into the Real or Fake class degrades reliability and prevents platform moderators from distinguishing malicious postings from low-quality submissions. For example, the recent transformer-based system of Akram et al. (2024) combined BERT and RoBERTa with ten variants of the Synthetic Minority Oversampling Technique (SMOTE) and reported a balanced accuracy of approximately 90%, yet it neither introduced an Invalid category nor extended the trained model into a deployed, user-facing application.

To address these gaps, this study develops an explainable DistilBERT-based three-class classification system that labels job offers as Real, Fake, or Invalid. DistilBERT is a knowledge-distilled variant of BERT that retains approximately 97% of BERT's language-understanding performance while being about 40% smaller and 60% faster (Sanh et al., 2019), which makes it well suited to the accuracy–efficiency trade-off that governs practical fraud screening. The study integrates EMSCAD, the job-scam domain of the DIFrauD benchmark (Boumber et al., 2024), and a researcher-curated Invalid Token dataset; it resolves severe class imbalance through controlled sampling and textual augmentation instead of synthetic interpolation; and it deploys the resulting model as a FastAPI web application with SHAP-based explanations.

Accordingly, the paper makes four contributions: (i) a three-class Real/Fake/Invalid formulation of fake job posting detection; (ii) an augmentation- and weighting-based strategy for class imbalance suited to transformer text pipelines; (iii) empirical evidence that a distilled transformer achieves high accuracy on this harder task, supported by validation-based checkpoint selection and a fully held-out test set; and (iv) an explainable, deployable web application for real-time job offer screening. The remainder of the paper is organised as follows. Section 2 reviews related work, Section 3 describes the materials and methods, Section 4 presents the results, Section 5 discusses the findings and limitations, and Section 6 concludes the paper.

2 Related Work

2.1 Machine Learning Approaches to Fake Job Detection

Vidros et al. (2017) provided one of the earliest systematic treatments of ORF, introducing the EMSCAD corpus of 17,880 job postings and reporting that about 4.8% of postings on the studied platform exhibited characteristics of fraud. Their study, and much of the work that followed, applied classical supervised classifiers such as Random Forest, Naïve Bayes, Logistic Regression, and decision trees over TF-IDF or Bag-of-Words representations. While these models established an important empirical foundation, they depend on manually engineered lexical features and therefore degrade when fraudulent postings deliberately imitate professional recruitment language.

Building on this foundation, later studies turned to ensemble learning and deep learning to handle imbalanced recruitment corpora. Ensemble techniques such as Random Forest, which rests on the variance-reduction principle formalised by Breiman (2001), were frequently combined with oversampling to improve minority-class recall, while recurrent architectures such as Bidirectional Long Short-Term Memory networks modelled sequential dependencies in job-description text more

effectively than feature-based classifiers. The class-imbalance problem that underlies this work is comprehensively treated by He and Garcia (2009), and SMOTE (Chawla et al., 2002) remains the most widely used oversampling technique, including in the baseline evaluated here. Because skewed label distributions can inflate aggregate accuracy while the minority class is poorly detected, and because improper validation can introduce optimistic bias (Kohavi, 1995), the present study reports balanced accuracy and per-class metrics on a fully held-out test set.

2.2 Transformer Language Models and DistilBERT

The transformer architecture, and specifically BERT (Devlin et al., 2019), marked a substantial advance in contextual language modelling. Unlike sequential models, BERT processes an entire input sequence simultaneously through self-attention, which allows it to capture long-range dependencies regardless of word position. RoBERTa (Liu et al., 2019) refined BERT's pretraining with larger corpora and improved objectives, typically raising accuracy at a higher computational cost, while ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020) pursued parameter efficiency and alternative pretraining tasks. Moreover, transfer learning allows such models to reuse representations learned from large corpora, improving performance on smaller task-specific datasets while reducing training time and resources (Okengwu et al., 2024).

DistilBERT (Sanh et al., 2019) addresses the accuracy–efficiency trade-off directly through knowledge distillation, in which a compact student network learns to reproduce the output distribution of a larger teacher model. The resulting architecture halves the number of encoder layers relative to BERT-base (six instead of twelve) while retaining most of its language-understanding capability. Notably, a comparable trade-off appears in computer vision: Trisnawarman et al. (2025) found that a Vision Transformer slightly outperformed a CNN-based EfficientNet-B0 on fine-grained image classes (96.25% versus 95.62%), whereas the CNN offered superior computational efficiency. This cross-domain evidence reinforces the view that model choice should be guided by deployment context rather than benchmark accuracy alone, which is the rationale for selecting DistilBERT in this study.

2.3 Transformer-Based Online Recruitment Fraud Detection

Applied specifically to ORF, Taneja et al. (2025) proposed Fraud-BERT, a BERT model fine-tuned on EMSCAD that achieved high accuracy through contextual embeddings, while acknowledging that its computational cost limits deployment in resource-constrained settings. Similarly, Vu et al. (2025) proposed NLP2FJD, which combines Word2Vec embeddings with convolutional layers over textual and metadata attributes, but noted a continued dependence on dataset-specific characteristics. Most directly relevant to this paper, Akram et al. (2024) merged EMSCAD with newly collected United States and Pakistani job postings, faced a majority-to-minority ratio of about 36,162:735, and evaluated ten SMOTE variants with BERT and RoBERTa. Their best configuration, BERT with SMOBD-SMOTE, reached a balanced accuracy and recall of approximately 90%; however, the study remained binary and reported no deployed application.

2.4 Explainability and Deceptive-Language Cues

Interpretability is increasingly required for automated decisions that affect consumers. Ribeiro et al. (2016) introduced LIME to approximate model behaviour locally around a prediction, whereas Lundberg and Lee (2017) proposed SHapley Additive exPlanations (SHAP), which attributes a prediction to individual input features with consistency and local-accuracy guarantees drawn from cooperative game theory. Doshi-Velez and Kim (2017) argue that such transparency is essential for user trust and accountability in high-stakes classification. In addition, research on deceptive language, including linguistic markers of deception in computer-mediated communication (Hancock et al., 2008), lexical cues for lie detection (Mihalcea & Strapparava, 2009), and deceptive opinion spam (Ott et al., 2011), supports the premise that fraudulent and legitimate text differ measurably, a premise that contextual transformer embeddings are well placed to exploit.

2.5 Research Gaps

Three gaps therefore motivate this study. First, published transformer-based ORF systems, including the Akram et al. (2024) baseline, remain confined to binary classification and cannot represent content that is invalid rather than fraudulent. Second, class imbalance in this domain has been treated mainly through SMOTE-style interpolation, which operates on fixed-dimensional feature vectors rather than on the raw text that transformers consume. Third, much of the literature stops at offline evaluation, leaving a gap between reported accuracy and practical, explainable use. The system described in the following section addresses all three gaps.

3 Materials and Methods

3.1 Research Design

This study adopts a Design Science Research (DSR) orientation, in which the principal outcome is a functional artefact, namely a three-class job offer classification system, whose effectiveness is empirically evaluated. Within this orientation, the machine learning work follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), covering business understanding, data understanding, data preparation, modelling, evaluation, and deployment, while a quantitative experimental design underpins the performance evaluation. The procedure thus moves from problem identification through dataset integration and preprocessing to model fine-tuning, evaluation, and deployment.

3.2 Data Sources and Integration

Three complementary data sources were combined to build the modelling corpus. EMSCAD (Vidros et al., 2017) and the job-scam domain of the DIFrauD benchmark (Boumber et al., 2024) supplied labelled genuine and fraudulent postings, from which the Real and Fake classes were derived using each dataset's original fraud indicator (for example, EMSCAD values *f* and *t* were mapped to Real and Fake). In addition, a researcher-curated Invalid Token dataset supplied examples of job-related text that is nonsensical, malformed, incomplete, or otherwise unsuitable for classification as a valid advertisement, thereby providing the Invalid category that distinguishes this formulation from binary ORF detection.

Because the sources differ in structure, each dataset was first identified by its column layout, and the available textual fields among the job title, company profile, description, requirements, and benefits were concatenated in a fixed order into a single text field per record. Missing fields were replaced with empty strings before concatenation, so that the absence of one attribute did not cause an entire record to be discarded. After cleaning and integration, the corpus contained 34,325 records, as summarised in Table 1. As the table shows, Real postings dominated, while Fake postings formed the smallest class, yielding an imbalance ratio of 21.8:1 between the largest and smallest classes. Figure 1 contrasts this distribution with the balanced corpus used for training.

Table 1. Class distribution of the integrated corpus before balancing.

Class	Number of Records	Percentage
Real	30,269	88.18%
Fake	1,390	4.05%
Invalid	2,666	7.77%
Total	34,325	100.00%

Note. Real and Fake labels were derived from the fraud indicators of EMSCAD and DIFrauD; Invalid records originate from the researcher-curated Invalid Token dataset.

3.3 Text Preprocessing

Once integrated, the text was cleaned using a common function. HTML markup was removed, whitespace was normalised, and extraneous special characters were stripped, while full stops, commas,

exclamation marks, question marks, hyphens, and colons were retained because they preserve natural sentence structure. Records whose consolidated text contained 20 characters or fewer were excluded, since such content cannot constitute a meaningful advertisement. Labels were then standardised into the common Real/Fake/Invalid scheme and encoded numerically with a label encoder, which was saved so that the identical mapping could be reused during evaluation and inference.

Alongside the primary text input, an auxiliary scam-keyword lexicon was constructed from consumer-protection guidance published by the U.S. Federal Trade Commission and the Better Business Bureau (Better Business Bureau, 2026; Federal Trade Commission, n.d.). This lexicon counted known scam indicators, such as upfront fees, urgency, and requests for banking details, and was used to verify that Fake postings exhibited a measurably higher density of scam language than Real postings, thereby providing a construct-validity check on the labels. Importantly, the keyword count was used for analysis and augmentation design only and was not supplied to the model, which learned exclusively from the raw text.

3.4 Class Balancing Through Augmentation and Class Weighting

Given the 21.8:1 imbalance, a single equal target of 3,000 records per class was adopted. The Real class was reduced to 3,000 records by random undersampling (random state = 42), whereas the Fake and Invalid classes were raised to 3,000 records through textual augmentation. For the Invalid class, augmented examples were generated from the curated seed records using distinct construction patterns, including alphanumeric noise, symbol spam, disorganised text containing stray HTML fragments, unfilled template placeholders, contact-only messages, and scrambled word sequences. Unlike SMOTE, which interpolates between points in a feature space, this strategy increases the volume and diversity of genuine textual examples that the transformer can learn from directly.

The resulting balanced corpus of 9,000 records was then partitioned into training, validation, and test subsets using stratified 70:15:15 sampling (random state = 42), as shown in Table 2. Consequently, every subset preserved exact class parity: each class contributed 2,100 training, 450 validation, and 450 test samples. The validation subset was used only for monitoring and checkpoint selection, while the test subset was held out entirely until final evaluation. Finally, balanced class weights were retained in the loss function as an additional safeguard against any residual imbalance.

Table 2. Final balanced dataset and stratified 70:15:15 partition.

Class	Total	Training (70%)	Validation (15%)	Test (15%)
Real	3,000	2,100	450	450
Fake	3,000	2,100	450	450
Invalid	3,000	2,100	450	450
Total	9,000	6,300	1,350	1,350

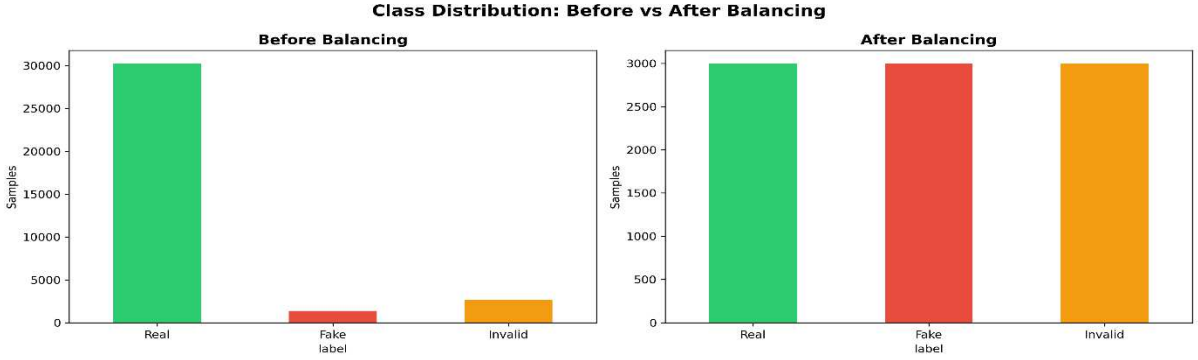


Fig. 1. Class distribution of the integrated corpus (a) before balancing and (b) after balancing to 3,000 records per class.

3.5 Proposed System Architecture

Figure 2 presents the architecture of the proposed system. The pipeline begins with data collection, in which EMSCAD, DIFrauD, and the Invalid Token dataset are merged into a single job-description repository. Next, the preprocessing stage removes symbols, URLs, numbers, HTML tags, email addresses, duplicates, and null records. In the feature-engineering stage, the relevant textual fields are merged into one input and the standardised labels are assigned (Real = 0, Fake = 1, Invalid = 2 before label encoding) before the datasets are concatenated. The integrated corpus then passes to class balancing, where undersampling, augmentation, and class weighting are applied ahead of the stratified 70:15:15 split.

Subsequently, in the classifier stage, tokenised input is mapped to token and position embeddings and passed through six transformer encoder layers with multi-head self-attention and feed-forward sub-layers. The contextualised representation of the classification token is then fed to a three-class softmax output layer. Predictions are evaluated with accuracy, balanced accuracy, precision, recall, F1-score, a confusion matrix, and ROC-AUC, after which the fine-tuned model is exposed through a deployment layer consisting of a web frontend, a FastAPI backend, and a SHAP explainability module.

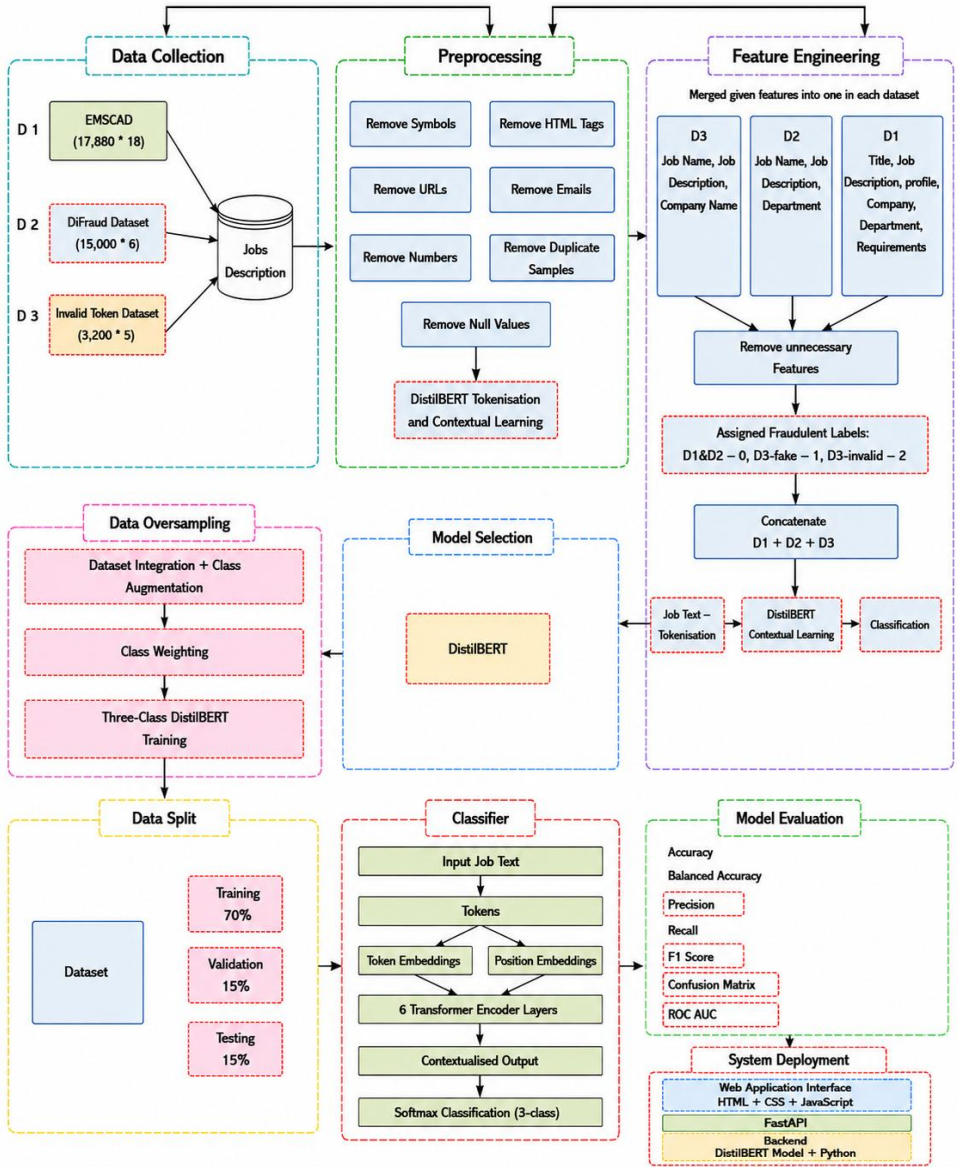


Fig. 2. Architecture of the proposed explainable DistilBERT-based three-class system for Real, Fake, and Invalid job offer detection.

3.6 Model Training Configuration

The model was initialised from the pretrained *distilbert-base-uncased* checkpoint and fine-tuned on the training subset. Each input was tokenised to a maximum length of 256 tokens, with longer sequences truncated and shorter ones padded, and the resulting token IDs and attention masks formed the model input. Optimisation used AdamW with a learning rate of 2×10^{-5} and a linear learning-rate scheduler with warm-up over four epochs, with a training batch size of 16 and an evaluation batch size of 32. After each epoch, validation loss and accuracy were computed, and the checkpoint was saved whenever validation loss improved; the best checkpoint was then reloaded for the single final evaluation on the test subset. Experiments were run in a Kaggle notebook environment using Python 3.13.7, PyTorch, Hugging Face Transformers 4.52.4, Datasets 3.6.0, and scikit-learn. Table 3 summarises the configuration.

Table 3. DistilBERT experimental and training configuration.

Parameter	Configuration
Pretrained model	distilbert-base-uncased (6 encoder layers)
Classification type	Three-class (Real, Fake, Invalid)
Final dataset size	9,000 samples (3,000 per class)
Train / validation / test split	70% (6,300) / 15% (1,350) / 15% (1,350), stratified
Random state	42
Maximum sequence length	256 tokens
Batch size	16 (training); 32 (validation and test)
Optimiser and learning rate	AdamW; 2×10^{-5}
Learning-rate scheduler	Linear, with warm-up
Number of epochs	4
Checkpoint selection	Lowest validation loss
Class-imbalance treatment	Undersampling + textual augmentation + balanced class weights

3.7 Baseline System

The ORF detection system of Akram et al. (2024) was adopted as the comparative baseline. That system fine-tunes BERT and RoBERTa on EMSCAD combined with United States and Pakistani postings, applies byte-pair encoding and ten SMOTE variants to a 70:30 train–test split, and formulates the task as binary fraudulent/non-fraudulent classification, reporting a balanced accuracy and recall of approximately 90% for its best configuration (BERT with SMOBD-SMOTE). Because the baseline solves a different problem, with two classes rather than three and different data, its figures are used as contextual benchmarks for transformer-based ORF detection rather than as a strictly like-for-like comparison.

3.8 Evaluation Metrics

Performance was assessed using overall accuracy, balanced accuracy, and class-level precision, recall, and F1-score, together with a confusion matrix and one-versus-rest ROC analysis. For each class, with TP, TN, FP, and FN denoting true positives, true negatives, false positives, and false negatives, the metrics are defined as follows:

$$Accuracy = \Sigma \text{ correct predictions} / N \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall) \quad (4)$$

$$\textit{Balanced accuracy} = (1/K)\,\Sigma\,\textit{Recall},\ K = 3 \qquad (5)$$

Balanced accuracy averages per-class recall and is therefore less sensitive to class imbalance than accuracy. Particular attention is given to Fake-class recall, because an undetected fraudulent posting exposes a job seeker to harm. Furthermore, the confusion matrix reveals the specific pattern of errors between classes, while ROC-AUC evaluates discrimination across all decision thresholds rather than at a single fixed threshold.

3.9 Deployment and Explainability

Finally, the fine-tuned model was integrated into a client–server web application to demonstrate practical, real-time use. The frontend, built with HTML, CSS, and JavaScript, provides a text area through which a user submits job advertisement content. The FastAPI backend receives the text through a RESTful endpoint, applies the same cleaning and tokenisation pipeline used during training, performs inference, and returns the predicted class and per-class confidence scores as JSON. On request, an explainability module built with the SHAP Partition explainer and a tokenizer-aware text masker computes token-level Shapley values, allowing the interface to highlight the words that pushed a prediction towards or away from each class (Lundberg & Lee, 2017). The application was containerised with Docker to support consistent, reproducible deployment.

4 Results

4.1 Exploratory Data Analysis

Exploratory analysis confirmed the class imbalance reported in Table 1 and, in addition, revealed marked differences in text length across the three categories. Real job offers had a mean length of approximately 2,880 characters (median 2,661; range 80–77,060), and Fake offers had a mean of approximately 2,275 characters (median 1,850; range 23–20,528). By contrast, Invalid entries were far shorter, with a mean of approximately 110 characters (median 74; range 21–383). Figure 3 shows these distributions. These differences indicate that, beyond semantic content, structural cues such as length provide an additional signal that a transformer classifier can implicitly exploit. Moreover, the very long Real postings confirm the need for the 256-token truncation limit adopted during tokenisation.

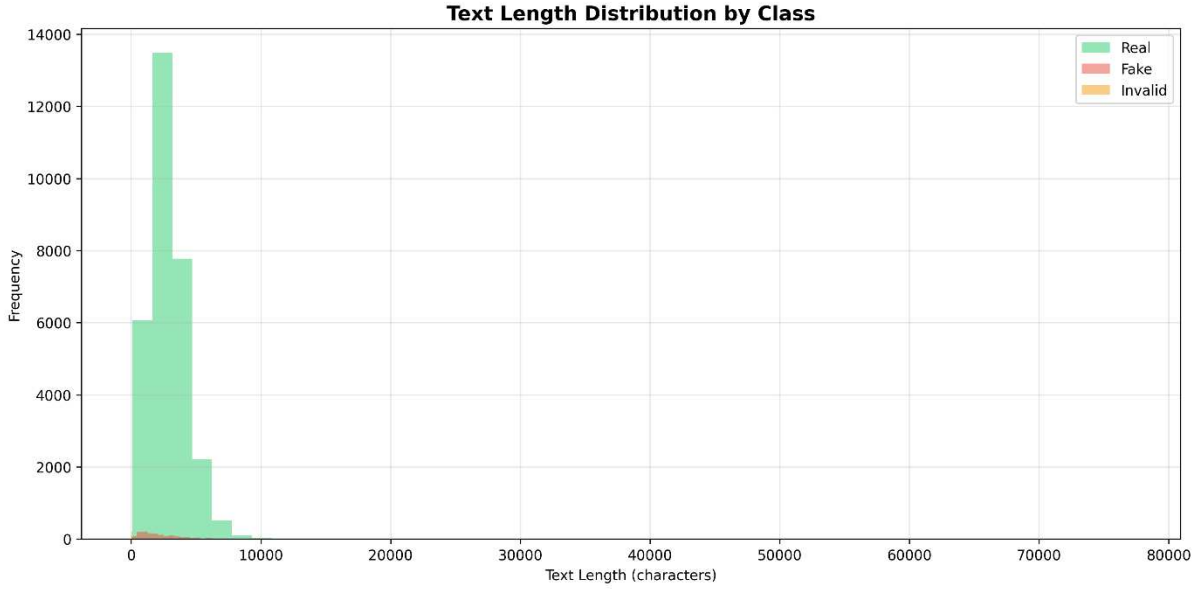


Fig. 3. Distribution of text length (characters) across the Real, Fake, and Invalid job offer categories.

4.2 Training Convergence

Training and validation behaviour was monitored across the four fine-tuning epochs, as shown in Figure 4. After the first epoch, training accuracy was 84.00% and validation accuracy was already 96.37%, indicating that the pretrained representations adapted rapidly to the task. Thereafter, training loss fell steadily from 0.3932 in the first epoch to 0.0056 in the fourth, while validation accuracy rose from 96.37% to 98.74%. Crucially, validation loss decreased at every epoch, so the best checkpoint coincided with the final epoch and no evidence of overfitting was observed within the training regime. The complete four-epoch fine-tuning run took approximately 622 seconds, which underlines the efficiency of the distilled architecture.

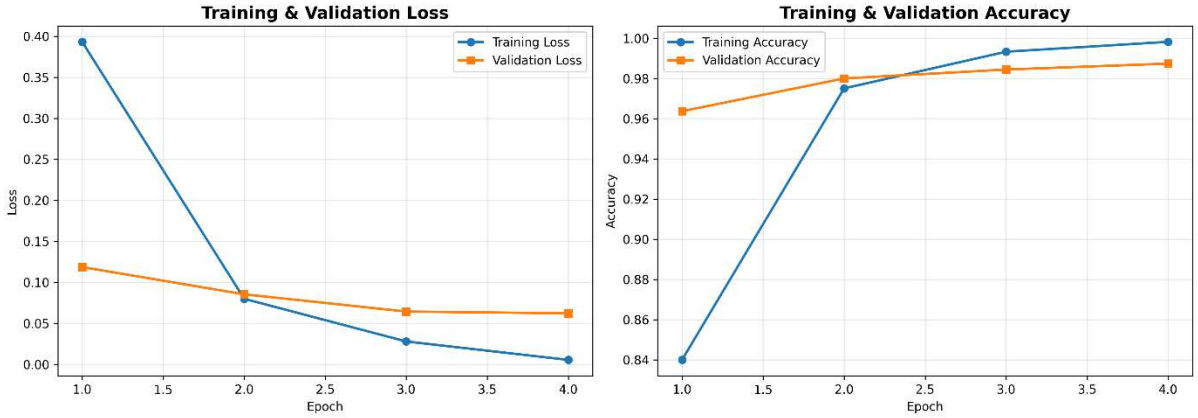


Fig. 4. Training and validation loss (left) and accuracy (right) across four fine-tuning epochs.

4.3 Overall and Class-Level Performance

When the best checkpoint was evaluated on the 1,350-sample held-out test set, the model achieved an overall accuracy of 98.74% and a balanced accuracy of 98.74% (Table 4). Because the test set contains exactly 450 samples per class, the two measures coincide, and the result confirms that predictive strength is distributed evenly across the three categories rather than concentrated in one. In total, only 17 of the 1,350 test samples (1.26%) were misclassified.

Table 4. Overall DistilBERT performance on the held-out test set.

Evaluation Metric	Result
Test accuracy	98.74%
Balanced accuracy	98.74%
Macro-average precision / recall / F1	0.99 / 0.99 / 0.99
Weighted-average precision / recall / F1	0.99 / 0.99 / 0.99
Test samples	1,350 (450 per class)
Misclassified samples	17 (1.26%)
Total training time (4 epochs)	≈ 622 s

Table 5 and Figure 5 break these results down by class. The Fake class achieved a precision, recall, and F1-score of 0.98, meaning that 98% of fraudulent postings were detected while few legitimate postings were wrongly flagged. The Invalid class achieved perfect scores of 1.00 on all three measures, confirming that the structural irregularities of invalid content are readily separable from both genuine and fraudulent text. Likewise, the Real class achieved 0.98 on all three measures. As a result, the macro- and weighted-average scores were both 0.99, showing that no single class dominated the evaluation.

Table 5. Class-level precision, recall, F1-score, and ROC-AUC on the held-out test set.

Class	Precision	Recall	F1-Score	ROC-AUC	Support
Fake	0.98	0.98	0.98	0.998	450
Invalid	1.00	1.00	1.00	1.000	450
Real	0.98	0.98	0.98	0.998	450
Macro average	0.99	0.99	0.99	–	1,350
Weighted average	0.99	0.99	0.99	–	1,350

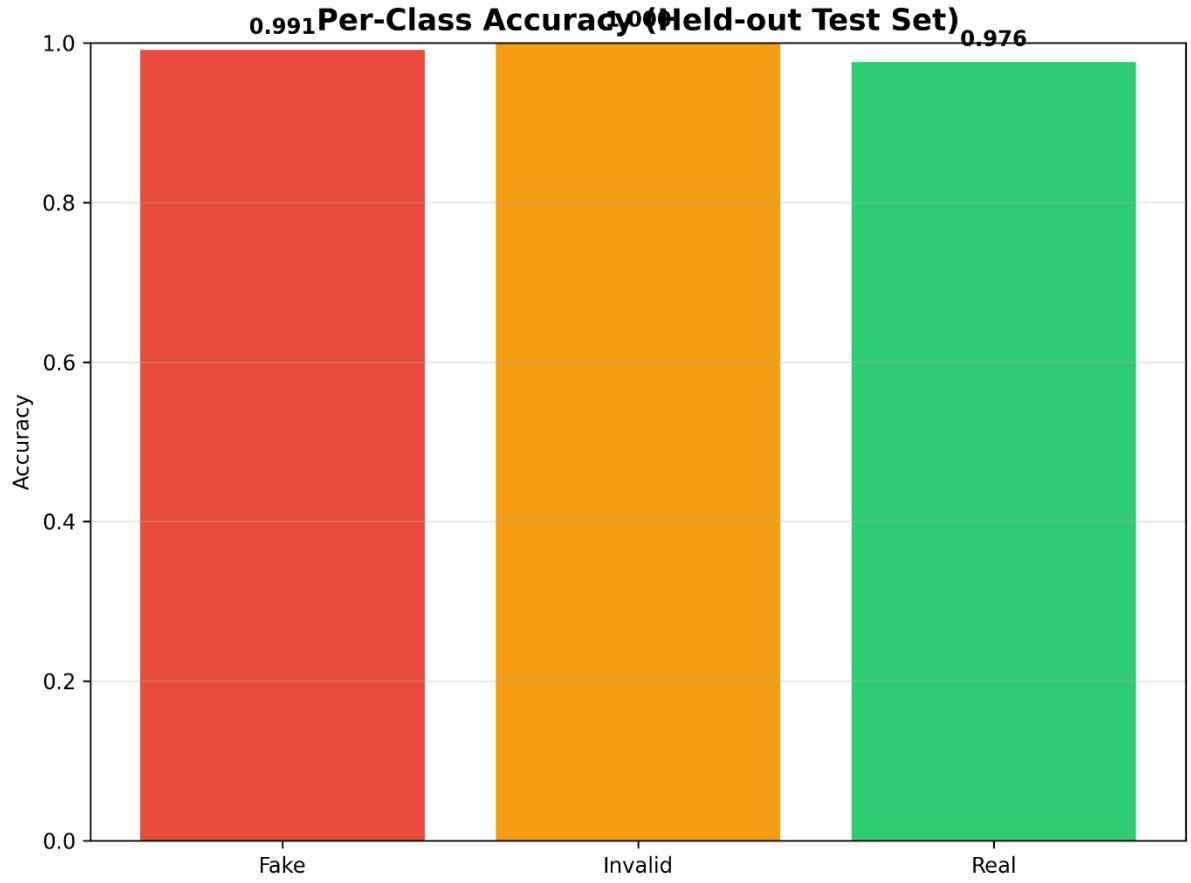


Fig. 5. Per-class precision, recall, and F1-score for the Real, Fake, and Invalid classes.

4.4 Confusion Matrix and ROC Analysis

To examine the pattern of errors in more detail, Figure 6 presents the confusion matrix, which shows strong diagonal dominance across all three classes. Since the Invalid class achieved perfect precision and recall, none of the 17 errors involved Invalid content; instead, every misclassification occurred between the Real and Fake classes. This outcome is consistent with the inherent difficulty of the task, because fraudulent postings are deliberately written to resemble legitimate advertisements. From a user-protection standpoint, the most consequential errors are Fake postings predicted as Real, since these could expose a job seeker to an undetected scam; the Fake-class recall of 0.98 indicates that such errors were rare.

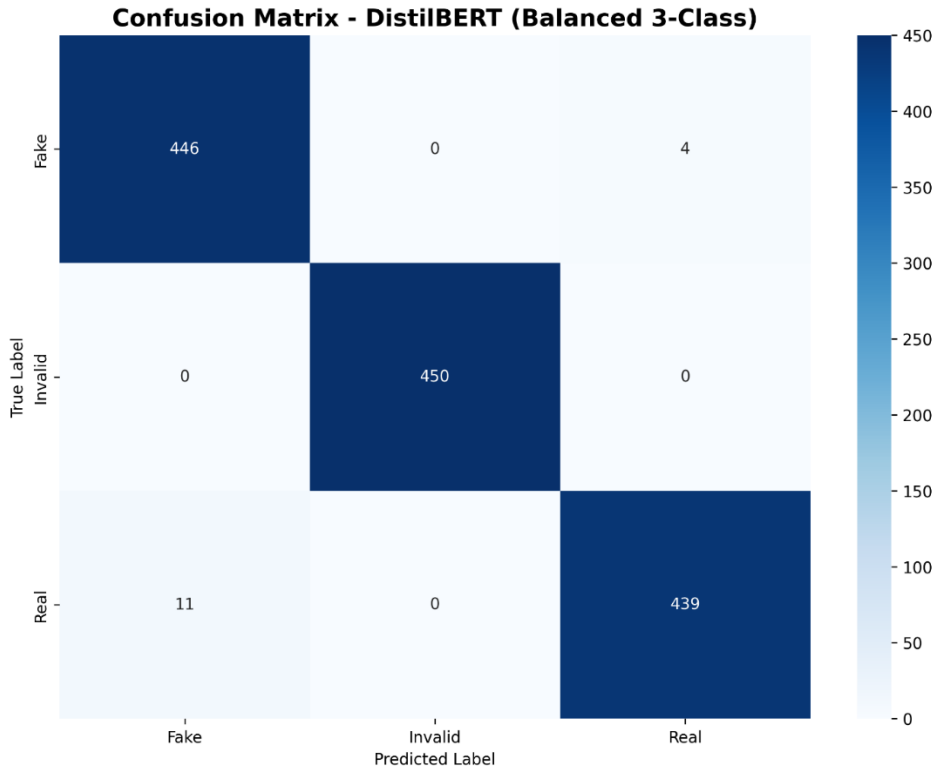


Fig. 6. Confusion matrix for Real, Fake, and Invalid job offer classification on the held-out test set.

Complementing the confusion matrix, one-versus-rest ROC curves were computed for each class to evaluate discrimination independently of a fixed decision threshold (Figure 7). The resulting AUC values were 0.998 for Fake, 1.000 for Invalid, and 0.998 for Real, indicating that the model separates each class from the other two with very high reliability across the full threshold range. In practice, this property allows the operating threshold to be tuned to deployment priorities; for instance, a platform that prioritises user safety could lower the Fake-class threshold to raise recall further at a modest cost to precision.

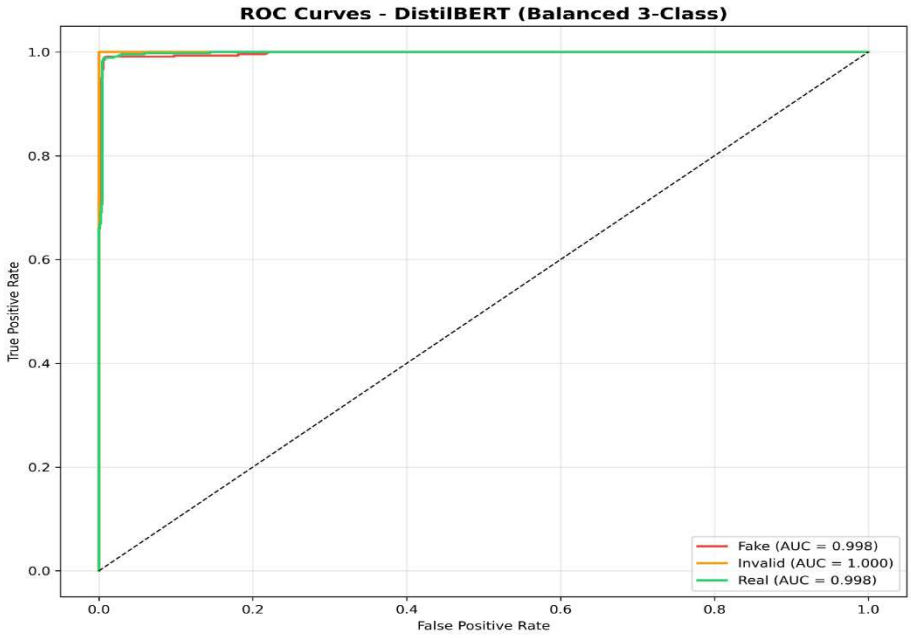


Fig. 7. One-versus-rest receiver operating characteristic (ROC) curves for the Real, Fake, and Invalid classes.

4.5 Comparison with the Baseline System

Table 6 summarises the structural and performance differences between the proposed system and the Akram et al. (2024) baseline. Whereas the baseline addresses binary classification with BERT, RoBERTa, and ten SMOTE variants and reports approximately 90% balanced accuracy, the proposed system performs three-class classification with DistilBERT and reaches 98.74% accuracy and balanced accuracy. Although the two systems use different data and problem formulations, and the comparison is therefore contextual rather than like-for-like, the proposed model attains higher reported performance on a more granular task while also providing Invalid-content detection and a deployed, explainable interface.

Table 6. Comparison between the Akram et al. (2024) baseline and the proposed system.

Feature	Akram et al. (2024) Baseline	Proposed System
Classification formulation	Binary (fraudulent / legitimate)	Three-class (Real / Fake / Invalid)
Transformer model(s)	BERT and RoBERTa	DistilBERT (6 layers)
Dataset sources	EMSCAD + US + Pakistan postings	EMSCAD + DIFrauD + Invalid Token dataset
Class-imbalance approach	Ten SMOTE variants	Undersampling + textual augmentation + class weighting
Data partition	70% train / 30% test	70% train / 15% validation / 15% test
Reported best performance	≈ 90% balanced accuracy / recall	98.74% accuracy / 98.74% balanced accuracy
Invalid-content detection	Not included	Included (F1 = 1.00)
Deployed application	Not reported	FastAPI web application with SHAP explanations

4.6 Explainability and Feature Attribution

To examine the linguistic basis of the model's decisions, SHAP token attributions were computed for representative predictions. For the Fake class, tokens associated with urgency, unrealistic monetary incentives, and requests for sensitive financial information, for example references to immediate start dates, unsolicited payment platforms, bank details, or unusually high salaries, consistently received high positive Shapley values. For the Invalid class, the most influential tokens were random character sequences, repeated or truncated words, and fragments lacking coherent sentence structure, which agrees with the short and irregular character of Invalid examples described in Section 4.1. In contrast, for the Real class, tokens describing specific organisational detail, structured qualification requirements (such as "bachelor's degree" and "years of experience"), and professional terminology carried the strongest positive attribution. Taken together, these patterns indicate that predictions rest on interpretable, domain-plausible cues rather than on spurious correlations, and they give end users a transparent rationale for each classification.

4.7 Deployed System Performance

Finally, the deployed web application was evaluated with representative submissions covering all three categories (Figure 8). Table 7 summarises the predicted classes, confidence scores, and end-to-end response times measured from submission to displayed result. Response times ranged from about 1.0 to 2.9 seconds, and the interface remained responsive throughout inference. These results are consistent with the intended real-time use case and reflect the efficiency of DistilBERT relative to full-scale BERT or RoBERTa deployment.

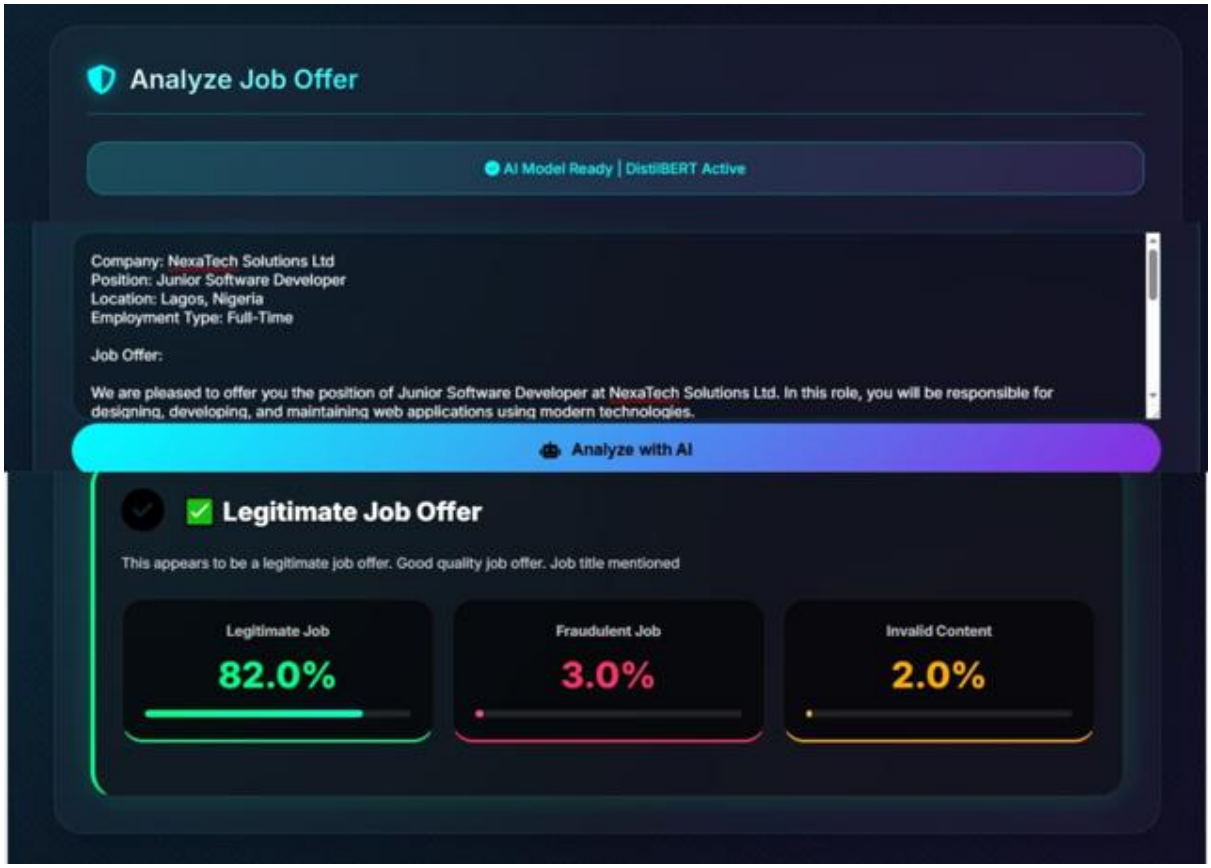


Fig. 8. Deployed web interface showing a sample classification result with per-class confidence scores.

Table 7. Sample real-time prediction results from the deployed system.

Test Case	Predicted Class	Confidence Score	Response Time (s)
Input 1	Real	82.0%	1.1
Input 2	Fake	96.0%	1.3
Input 3	Invalid	85.0%	1.0
Input 4	Fake	98.0%	2.9

5 Discussion

The results show that a knowledge-distilled transformer can achieve high accuracy on a fake job detection task that is formally more demanding than the binary formulation addressed in prior work. The accuracy and balanced accuracy of 98.74% exceed the approximately 90% balanced accuracy reported for the strongest Akram et al. (2024) configuration, even though the present model must additionally separate Invalid content from both Real and Fake postings. This finding is consistent with the broader transformer literature, in which contextual self-attention captures subtle cues such as exaggerated claims, inconsistent detail, and manipulative phrasing (Devlin et al., 2019; Liu et al., 2019), and it extends that evidence to a three-class recruitment-fraud setting.

Compared with an earlier iteration of this pipeline, which left the classes unequal after sampling and evaluated on a single train–test split, the revised design brings every class to exact parity, introduces a separate validation set for checkpoint selection, and keeps the test set untouched until the final evaluation. As a result, the gain in Fake-class F1-score from 0.95 to 0.98 is obtained under a stricter protocol, which strengthens confidence that the reported performance is not an artefact of bias towards

a dominant class. Similarly, the smooth, monotonic decline in validation loss across all four epochs indicates stable convergence rather than memorisation of the training data.

The perfect scores for the Invalid class nonetheless merit careful interpretation. Invalid content is structurally distinctive: its mean length of about 110 characters is more than an order of magnitude shorter than that of Real or Fake postings, and its augmentation patterns deliberately introduce corrupted or templated text that rarely overlaps with recruitment language. Therefore, while the result validates the feasibility of an explicit Invalid category, performance on a broader and more heterogeneous population of naturally occurring invalid submissions may be lower. By contrast, the residual errors between the Real and Fake classes reflect the core difficulty of ORF detection, namely that fraudsters imitate the structure, tone, and vocabulary of legitimate advertisements; some overlap is thus expected even for a strong contextual model.

From a systems perspective, the results support the central premise behind selecting DistilBERT: high accuracy was achieved with a four-epoch training time of roughly ten minutes and end-to-end response times under three seconds. This efficiency–accuracy balance mirrors the pattern reported by Trisnawarman et al. (2025) for Vision Transformers and CNNs in image classification, which suggests that the trade-off is a general property of attention-based versus compact architectures under real-world constraints rather than a quirk of the recruitment domain. Furthermore, the SHAP explanations distinguish the system from prior ORF research that reports only a label; by grounding each prediction in visible cues, the system supports user trust and accountability (Doshi-Velez & Kim, 2017; Lundberg & Lee, 2017; Ribeiro et al., 2016) and gives job seekers actionable evidence for why a posting was flagged.

Several limitations qualify these findings. First, EMSCAD and DIFrauD may not represent the full diversity of contemporary recruitment fraud across platforms, industries, and regions, and the curated Invalid Token dataset may not capture every type of invalid input that real users submit. Second, the study is confined to English-language postings. Third, because augmentation was applied before partitioning, augmented variants derived from the same seed record may appear in more than one subset, which could make estimates for the augmented Fake and Invalid classes optimistic; applying augmentation only within the training partition would remove this risk. Fourth, the balanced test set does not reflect the natural prevalence of fraud (about 4% in the integrated corpus), so precision for the Fake class would be expected to fall under real-world class ratios. Finally, the three-class scheme does not separate fraud sub-types such as advance-fee and multi-level-marketing schemes.

To contextualise the performance of the fine-tuned DistilBERT model, two additional architectures, LSTM and RNN, were trained and evaluated under identical conditions: the same balanced dataset of 9,000 records, the same stratified 70:15:15 train, validation, and test split, the same class weighted loss, and the same four training epochs. This allowed the transformer-based approach to be benchmarked directly against sequential neural network baselines without any difference in data exposure influencing the comparison.

The bidirectional LSTM model achieved a test accuracy of 98.44% and a balanced accuracy of 98.44%, slightly below the 98.89% achieved by DistilBERT but considerably above conventional expectations for a recurrent architecture trained from scratch rather than pretrained. This result indicates that the balanced and augmented dataset provided sufficient signal for the LSTM to learn meaningful distinctions between the Real, Fake, and Invalid classes, although its performance remained marginally weaker than the transformer-based model, consistent with DistilBERT's advantage of pretrained contextual language representations. Figure 9 shows the confusion matrix for the LSTM model, Figure 10 shows its ROC curves across the three classes, and Figure 11 shows its training and validation loss and accuracy curves across the four epochs.

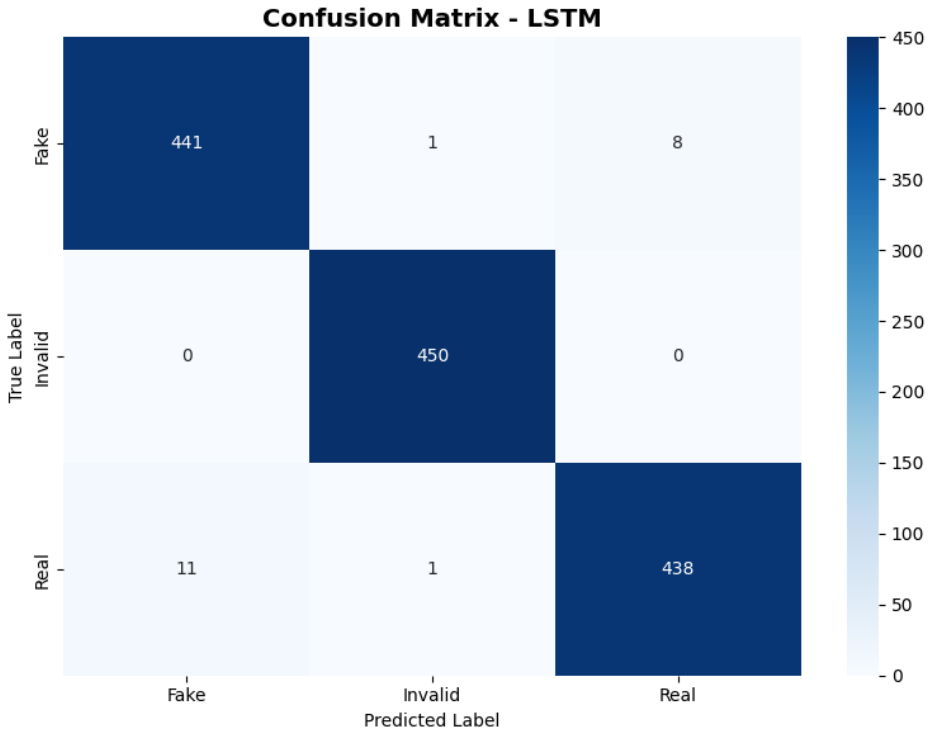


Fig. 9. LSTM Confusion Matrix.

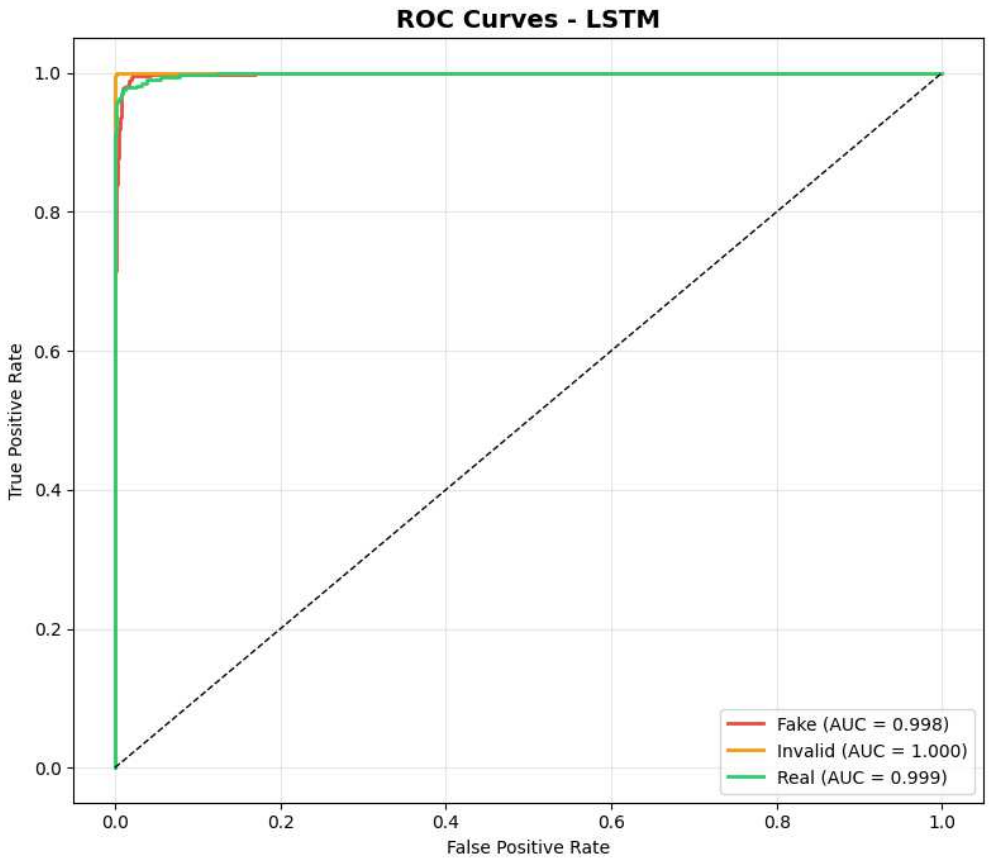


Fig. 10. LSTM ROC Curves.

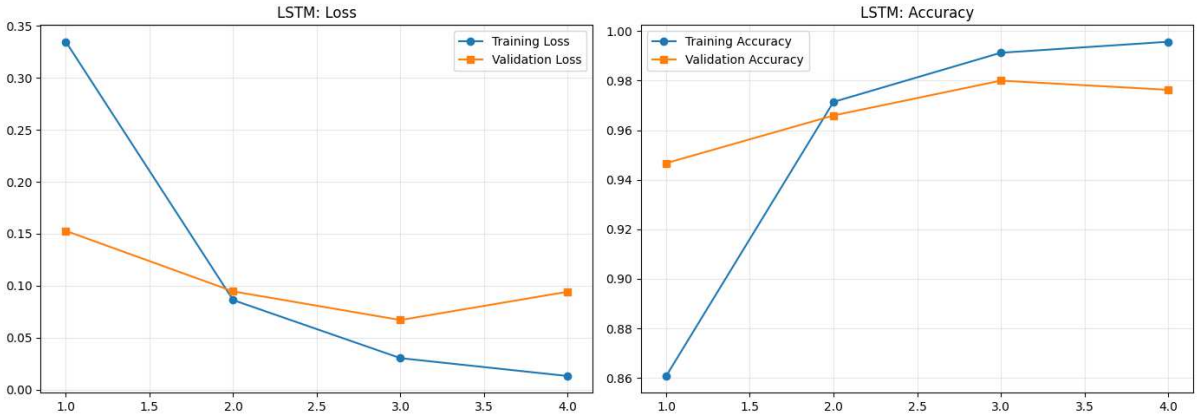


Fig. 11 LSTM Training and Validation Loss and Accuracy Curves

The bidirectional RNN model achieved a test accuracy of 97.04% and a balanced accuracy of 97.04%, the lowest among the three models evaluated. This outcome is consistent with the known limitations of simple RNN architectures, which are more susceptible to vanishing gradient effects over longer input sequences than LSTM or transformer-based architectures, and therefore less able to capture long range dependencies within longer job postings. Despite this, the RNN still achieved a strong overall result, confirming that the balancing procedure applied to the dataset benefited all three architectures rather than being advantageous to DistilBERT alone. Figure 12 shows the confusion matrix for the RNN model, Figure 13 shows its ROC curves across the three classes, and Figure 14 shows its training and validation loss and accuracy curves across the four epochs.

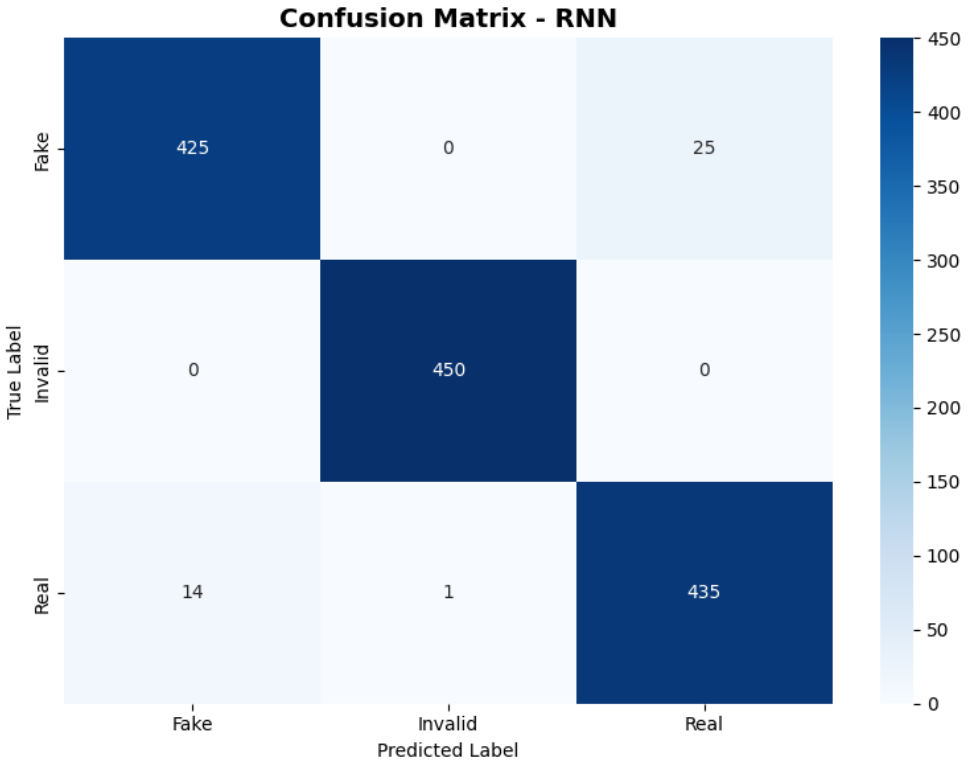


Fig. 12. RNN Confusion Matrix

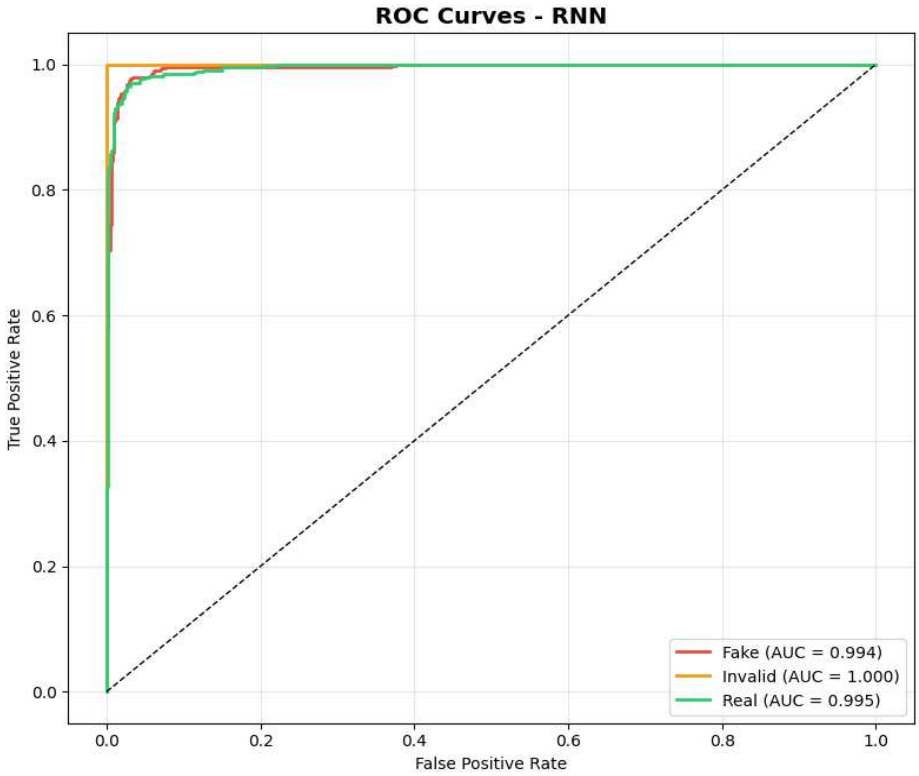


Fig. 13. RNN ROC Curves

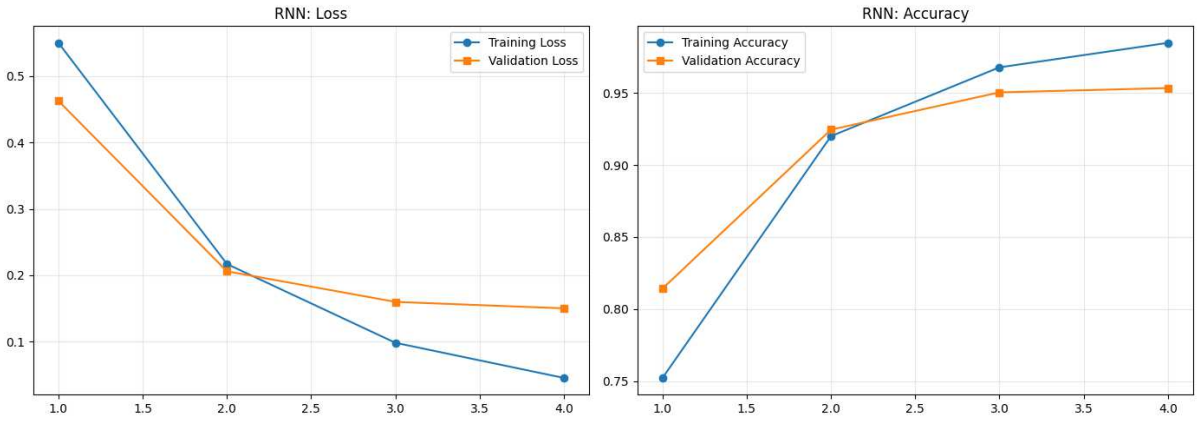


Fig. 11 RNN Training and Validation Loss and Accuracy Curves

6 Conclusion

This study developed and evaluated an explainable DistilBERT-based system that classifies online job offers as Real, Fake, or Invalid, thereby extending conventional binary fake job posting detection to a more comprehensive three-class framework. By integrating EMSCAD, DIFrauD, and a curated Invalid Token dataset, and by resolving a 21.8:1 class imbalance through controlled undersampling, textual augmentation, and class weighting rather than SMOTE, the fine-tuned model achieved 98.74% accuracy and balanced accuracy on a fully held-out test set, with F1-scores of 0.98 for Fake, 1.00 for Invalid, and 0.98 for Real, and ROC-AUC values of 0.998, 1.000, and 0.998. These results exceed the approximately 90% balanced accuracy reported for a comparable binary transformer baseline. In addition, the model was deployed as a FastAPI web application with SHAP explanations, demonstrating that the approach is accurate, efficient, interpretable, and ready for real-time use.

A comparison against LSTM and RNN baselines, trained under identical data and balancing conditions, further supports this conclusion. DistilBERT achieved the highest test and balanced accuracy at 98.89%, followed by the LSTM at 98.44% and the RNN at 97.04%, confirming that the transformer-based approach outperformed both recurrent alternatives while all three benefited from the equal class balancing applied in this study.

In summary, the principal contributions are (i) a validated three-class Real/Fake/Invalid formulation of recruitment-fraud detection; (ii) an augmentation- and weighting-based imbalance strategy suited to transformer text pipelines; (iii) evidence that a distilled transformer can deliver high accuracy on this harder task with low training and inference cost; and (iv) the integration of token-level explainability into a deployed fake job detection tool. Future work should prioritise external validation on independently collected, cross-platform postings; augmentation confined to the training partition; evaluation under natural class prevalence; extension to multilingual content; incorporation of non-textual signals such as recruiter and company verification; and periodic retraining to keep pace with evolving fraud tactics.

Declarations

Funding: The authors declare that no funding was received for the conduct of this research.

Conflict of Interest: The authors declare that they have no conflict of interest.

Data Availability: The Employment Scam Aegean Dataset (EMSCAD) and the DIFrauD benchmark are publicly available research datasets. The researcher-curated Invalid Token dataset and the trained model artefacts are available from the corresponding author upon reasonable request.

Ethical Approval: This study did not involve human participants, personal data collection, or identifiable individuals; it relied exclusively on publicly available and researcher-curated job advertisement text. Formal ethical approval was therefore not applicable.

Author Contributions: N.K.U. conceptualised the study, developed and fine-tuned the model, implemented the deployment system, and drafted the manuscript. U.A.O., C.B.N., and P.S.J. supervised the research, contributed to the methodology, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

Akram, N., Irfan, R., Al-Shamayleh, A. S., & Akhunzada, A. (2024). Online recruitment fraud (ORF) detection using deep learning approaches. *IEEE Access, 12*, 109388–109408. https://doi.org/10.1109/ACCESS.2024.3435670

Better Business Bureau. (2026). *Employment scams 2026 update*. BBB Scam Studies. https://www.bbb.org/all/scamstudies/jobscams/employment-scams-2026-update

Boumber, D. A., Qachfar, F. Z., & Verma, R. (2024). Domain-agnostic adapter architecture for deception detection: Extensive evaluations with the DIFrauD benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 5260–5274). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.468

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://doi.org/10.48550/arXiv.1702.08608

Federal Trade Commission. (n.d.). *Job scams*. Consumer Advice. https://consumer.ftc.gov/articles/job-scams

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45*(1), 1–23. https://doi.org/10.1080/01638530701739181

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143). Morgan Kaufmann.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*. https://doi.org/10.48550/arXiv.1907.11692

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)* (pp. 4765–4774).

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 309–312). Association for Computational Linguistics.

Okengwu, U. A., Akpughe, H. A., Odogu, E., & Ojetunmibi, T. (2024). Transfer learning for tomato leaf disease detection using convolutional neural networks on mobile platforms. In *International Conference on Technological Solutions for Smart Economy (SmartEco'2024)* (pp. 264–273).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 309–319). Association for Computational Linguistics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv*. https://doi.org/10.48550/arXiv.1910.01108

Taneja, K., Vashishtha, J., & Ratnoo, S. (2025). Fraud-BERT: Transformer based context aware online recruitment fraud detection. *Discover Computing, 28*(1), Article 25. https://doi.org/10.1007/s10791-025-09502-8

Trisnawarman, D., Supriyanton, A. A., Mawardi, V. C., & Okengwu, U. A. (2025). Comparative study of CNN and Vision Transformers on Indonesian traditional cakes classification. *International Journal of Advances in Artificial Intelligence and Machine Learning, 2*(2), 86–94. https://doi.org/10.58723/ijaaiml.v2i2.405

Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet, 9*(1), 6. https://doi.org/10.3390/fi9010006

Vu, D. H., Nguyen, K., Tran, K. T., Vo, B., & Le, T. (2025). Improving fake job description detection using deep learning-based NLP techniques. *Journal of Information and Telecommunication, 9*(1), 113–125. https://doi.org/10.1080/24751839.2024.2387380