
Volume 1 Issue 1

Article Number: JASAT20260006

**Data Quality-Aware Detection of Fraudulent Job Advertisements Using
Text and Missing-Metadata Patterns**

Karthik Juluri¹ and Krishna Kant Pandey^{* 2}

¹Independent Researcher, United States

²Department of Mechatronics, Manipal University Jaipur, Jaipur, Rajasthan 303007, India

Abstract

Fraudulent job advertisements can expose applicants to financial loss, identity theft, and wasted effort. Most automated detectors emphasize wording, although an advertisement's missing or incomplete metadata may provide complementary evidence about how it was created. This paper proposes a data quality-aware decision-support framework that joins term frequency-inverse document frequency (TF-IDF) text features with explicit missingness indicators, aggregate completeness measures, and simple text-quality variables. A reproducible CPU-only simulation uses the open Employment Scam Aegean Dataset and a fixed stratified holdout. After removing 330 repeated predictor-identical records, the text-only logistic model obtained precision 0.673, recall 0.830, F1-score 0.743, balanced accuracy 0.905, and precision-recall area 0.877. Fusion increased recall to 0.889, balanced accuracy to 0.925, and precision-recall area to 0.886, but reduced precision to 0.535 and F1-score to 0.668. Metadata quality alone was considerably weaker. Thus, missingness provided a modest recall-oriented signal rather than a universal improvement. The practical contribution is a transparent ablation protocol for deciding when completeness patterns warrant inclusion in human-reviewed recruitment-fraud triage.

Keywords: Fraudulent Job Advertisements; Data Quality; Missing Metadata; Natural Language Processing; TF-IDF; Imbalanced Classification; Logistic Regression

1. Introduction

Online recruitment allows legitimate employers to reach large candidate pools, but the same low-friction publishing process can be abused to solicit money or personal information. Vidros *et al.* characterized online recruitment fraud and released the Employment Scam Aegean Dataset (EMSCAD), a public corpus containing both free text and structured advertisement attributes [1]. Subsequent studies explored feature selection and random forests [2], ensemble classification [3], imbalance-aware learning [4], localized contextual predictors [5], and crowd-informed enrichment [6]. Their collective evidence supports automated triage, but also shows that results depend on representation, class balance, platform, and evaluation design.

Text is indispensable: misleading offers must still present a role, organization, requirements, or inducement. Yet content-only models omit a second source of evidence—whether fields expected by the platform are present. A missing company profile, absent logo, omitted employment type, or sparse requirements may arise from concealment or low-effort

***Corresponding Author:** Krishna Kant Pandey (krishna.pandey@jaipur.manipal.edu)

Received: 25 June 2026; **Revised:** 30 July 2026; **Accepted:** 09 September 2026; **Published:** 30 September 2026

© 2026 The Author(s).

This is an open access article licensed under a [Creative Commons Attribution-NonCommercial 4.0 International Licence \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/).

construction. The same pattern can also arise innocently from optional fields, privacy choices, sector conventions, or a small firm’s limited recruiting capacity. Missingness should therefore be represented as a testable pattern, not converted into an automatic fraud rule.

This paper makes three contributions. First, it specifies a two-branch framework that preserves interpretable text and data-quality signals before fusion. Second, it supplies an executable, leakage-aware experiment comparing text-only, metadata-only, and fused logistic models on EMSCAD. Third, it reports the trade-off honestly: fusion improves fraud recall, balanced accuracy, and precision–recall area, but not precision or F1-score. The framework is conceptual and experimentally illustrated; it has not been deployed in a recruitment platform.

2. Related Work

Table I summarizes six open-access studies inspected for this work. EMSCAD-based research has combined text, binary flags, and categorical variables, often with tree ensembles or neural representations. Alghamdi and Alharby used selected advertisement characteristics with an ensemble design [2]; ORFDetector combined multiple engineered feature groups [3]; and Li *et al.* focused explicitly on severe class imbalance [4]. Mahbub *et al.* demonstrated that contextual predictors can be platform- and country-dependent [5]. Nanath and Olney studied crowd-derived information as enrichment rather than relying only on the original fields [6].

Table 1: Summary of related open-access studies.

Ref.	Dataset/domain	Features and model	Limitation relevant to this paper
[1]	EMSCAD; English job advertisements	Lexical, linguistic, and structured fields; classical supervised classifiers	Dataset-era and platform specificity; missingness not isolated as an ablation branch
[2]	EMSCAD	Profile/content and selected structured characteristics; SVM selection and random forest	Emphasizes final classifier performance rather than explicit missing-pattern utility
[3]	EMSCAD	Employer/job language plus contextual metadata; ensemble learning	Multiple feature groups are combined, so the marginal value of missingness is unclear
[4]	EMSCAD	Advertisement representation and original structured fields; imbalance-oriented models	Focuses on balancing and model choice, not data-quality indicators
[5]	Australian advertisements	Advertisement text and local contextual features; supervised comparison	Context is geographically and platform dependent
[6]	Online recruitment fraud; crowd enrichment	Semantic/context and crowd-derived information; multiple ML models	Enrichment requires information beyond the original advertisement

The literature-based gap is narrower than claiming that metadata has been ignored. Existing fraudulent-job-advertisement studies commonly use categorical or binary predictors, but metadata completeness is often handled during preprocessing or embedded in a broad feature set. Its independent and incremental predictive value is consequently difficult to identify. This paper treats each absent field, the overall pattern, and simple content completeness as an explicit, auditable view.

Missing-data research supplies the necessary caution. “Missing completely at random” (MCAR) means the probability of absence is unrelated to observed or unobserved values; “missing at random” (MAR) permits dependence on observed values; and “missing not at random” (MNAR) allows dependence on the unobserved value itself or another unobserved process [7]. Reviews show that the mechanism and pattern influence appropriate treatment [8]; classification studies also demonstrate that missing-value handling affects predictive behavior [9]. Recent work calls attention to structured missingness, where co-occurring absences carry organization beyond isolated nulls [10]. Job-advertisement missingness could reflect any mixture of these mechanisms, so correlation with the fraud label cannot establish malicious intent.

3. Problem Definition and Research Motivation

Let advertisement i have binary target y_i , where $y_i = 1$ denotes a posting labeled fraudulent by the dataset provider and $y_i = 0$ denotes a legitimate posting. The prediction task estimates risk from information available in the advertisement itself. It does not infer the employer’s legal guilt, nor does it cover later messages, payment requests, identity verification, or external corporate registries.

Five motivations guide the design. First, a scammer may deliberately omit traceable company information. Second, low-effort advertisements may leave optional fields blank. Third, platform form design can induce widespread absence independently of fraud. Fourth, industries differ in salary disclosure and credential conventions. Fifth, small firms may publish incomplete but genuine postings. Thus, a missing salary is not semantically equivalent to an absent logo, and a count of missing fields may obscure which combination occurred. The experiment retains both field-level indicators and summary measures.

The research question is whether quality features add measurable holdout value beyond TF-IDF. The ablation answers three subsidiary questions: whether quality-only classification is useful; which fields differ most by class; and whether any recall gain is accompanied by an unacceptable precision or F1 loss. Because only one historical corpus is used, the result tests this representation on EMSCAD rather than asserting general behavior across recruitment systems.

4. Proposed Data Quality-Aware Conceptual Framework

Figure 1 presents the original framework. Collection and quality assessment precede two parallel branches. The textual branch normalizes and combines the title, company profile, description, requirements, and benefits, then computes TF-IDF. The quality branch encodes field-level absence, aggregate missingness, and text completeness. Fusion occurs only after branch-specific transformations; an interpretable probabilistic classifier then supports evaluation and human review.

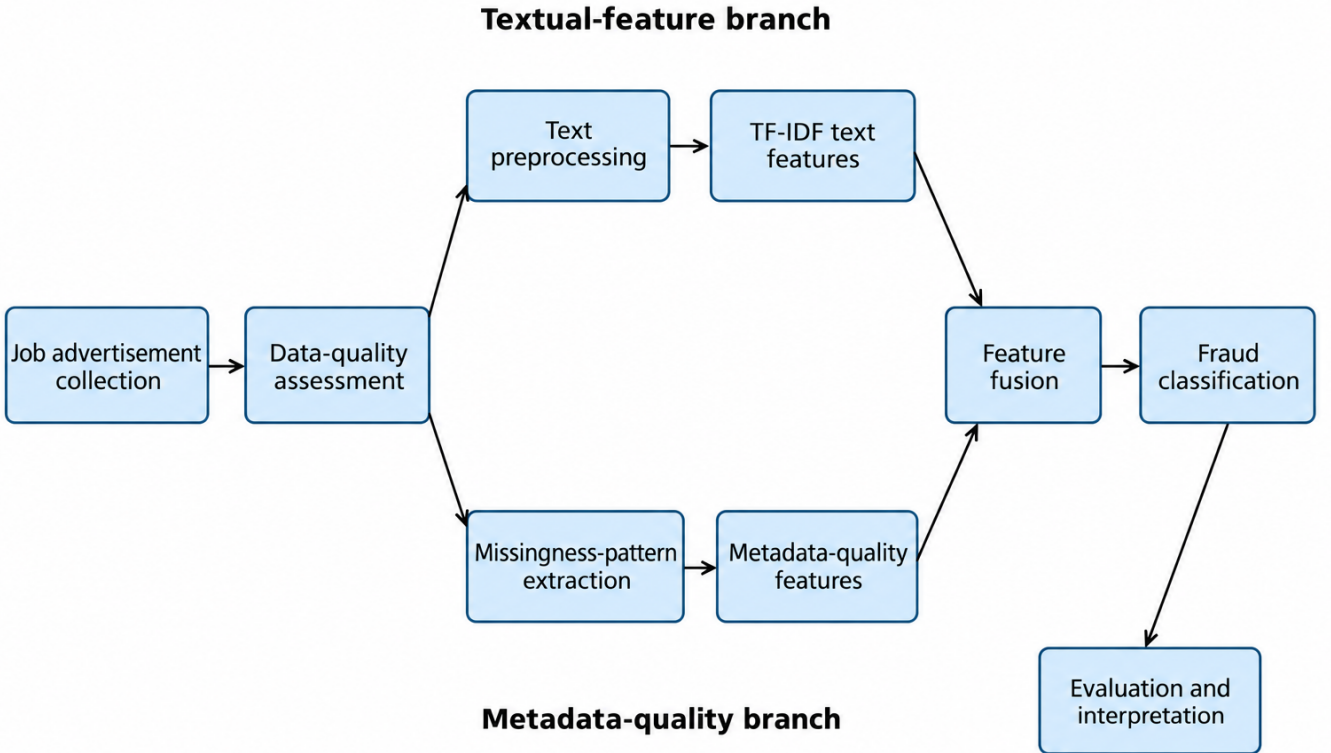


Figure 1: Proposed data quality-aware job-advertisement fraud-detection framework. Textual and metadata-quality branches are fused before classification; outputs are intended for evaluation and human-reviewed decision support.

For advertisement i , let d_i be the concatenated textual content. The textual representation is

$$\mathbf{t}_i = \text{TFIDF}(d_i), \quad (1)$$

where term-frequency and inverse-document-frequency weighting emphasize discriminative vocabulary [11]. For metadata field j ,

$$m_{ij} = \begin{cases} 1, & \text{if field } j \text{ is missing for advertisement } i, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The missingness vector and ratio are

$$\mathbf{m}_i = [m_{i1}, m_{i2}, \dots, m_{ip}], \quad r_i = \frac{1}{p} \sum_{j=1}^p m_{ij}. \quad (3)$$

In this study, $p = 10$: eight nullable metadata fields plus indicators that the posting has no company logo and no screening questions. These last two are platform-provided absence flags rather than CSV nulls. Let $\mathbf{q}_i$ contain the missing count, available-field count, character lengths, and a short-description flag. Fusion forms

$$\mathbf{x}_i = [\mathbf{t}_i \| \mathbf{m}_i \| r_i \| \mathbf{q}_i]. \quad (4)$$

The logistic model computes

$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp \left[- \left(\beta_0 + \beta^T \mathbf{x}_i \right) \right]}. \quad (5)$$

The linear classifier is deliberately modest: its coefficients support inspection and avoid adding a complex black box to a high-impact screening context, consistent with arguments for interpretable models where errors affect people [12].

4.1. Data-Quality Assessment and Control Logic

The first branch is more than a null counter. Collection should preserve the original schema, platform field definitions, and raw values before any cleaning. Quality assessment then distinguishes four conditions: a true CSV null; a whitespace-only string; a platform flag indicating that an item is absent; and a valid but terse value. These conditions may have different causes. In the simulation, null and blank strings are combined for the eight nullable metadata fields, while the logo and screening-question indicators retain their platform semantics. Text length is measured after null replacement but before TF-IDF tokenization. No missing field is imputed with a plausible category because doing so would erase the very signal under evaluation.

At inference time, the two branches must be generated from the same immutable advertisement snapshot. Schema validation should reject unexpected column changes rather than silently reordering features. A production implementation would also record form version, collection time, language, and whether each field was mandatory when the advertisement was submitted. These records allow analysts to determine whether an apparent missingness shift reflects scam behavior or a platform redesign. The classifier output is paired with branch-level evidence: influential text terms, present/absent quality indicators, and the aggregate completeness ratio. A review policy can then require corroboration across branches, route uncertain cases to a reviewer, and prohibit any rule in which one blank field alone causes enforcement.

5. Dataset and Data Preparation

5.1. Source and Audit

The experiment uses EMSCAD, formally described and released by Vidros *et al.* [1]. The local file is named `fake_job_postings.csv`; the executed copy has SHA-256 `cec194ca...85cae3e2`. Before cleaning it contained 17,880 rows, 18 columns, 17,014 legitimate labels, and 866 fraud labels. These values are computed by the supplied script rather than copied into its source.

Predictor-level exact duplicates were identified after lowercasing, whitespace normalization, and null normalization across every feature except `job_id` and the target. There were 281 duplicate groups and 330 later copies; no group contained conflicting labels. Removing later copies before splitting reduced direct duplicate leakage. The analysis set therefore contains 17,550 records (16,694 legitimate; 856 fraudulent), a 4.88% fraud rate. Two descriptions were empty, but every record retained some text across the five combined fields; missing text cells were replaced only with empty strings for concatenation. Near-duplicate detection was omitted because a dependable similarity threshold would introduce an additional modeling choice.

5.2. Observed Data Quality

Table II reports selected rates, while Fig. 2 shows all ten indicators by class. Company-profile absence differed by 51.61 percentage points (fraudulent 67.76% versus legitimate 16.14%), logo absence by 48.83 points, and screening-question absence by 21.20 points. The direction was not universal: salary was missing less often among fraudulent postings (74.18%) than legitimate postings (84.39%). This reversal is evidence against treating “more missing” as a categorical fraud rule.

Table 2: Dataset characteristics and selected missingness rates.

Item	Overall	Legit.	Fraud
Raw records	17,880	17,014	866
Analysis records	17,550	16,694	856
Company profile missing	18.66%	16.14%	67.76%
Salary range missing	83.89%	84.39%	74.18%
Employment type missing	19.52%	19.10%	27.57%
Required experience missing	39.63%	39.10%	49.88%
Required education missing	45.57%	45.25%	51.87%
Company logo absent	20.60%	18.22%	67.06%
Screening questions absent	50.75%	49.71%	70.91%

Percentages in the lower panel are computed after exact-duplicate removal. “Absent” flags are derived from platform binary columns; other entries are null/blank rates.

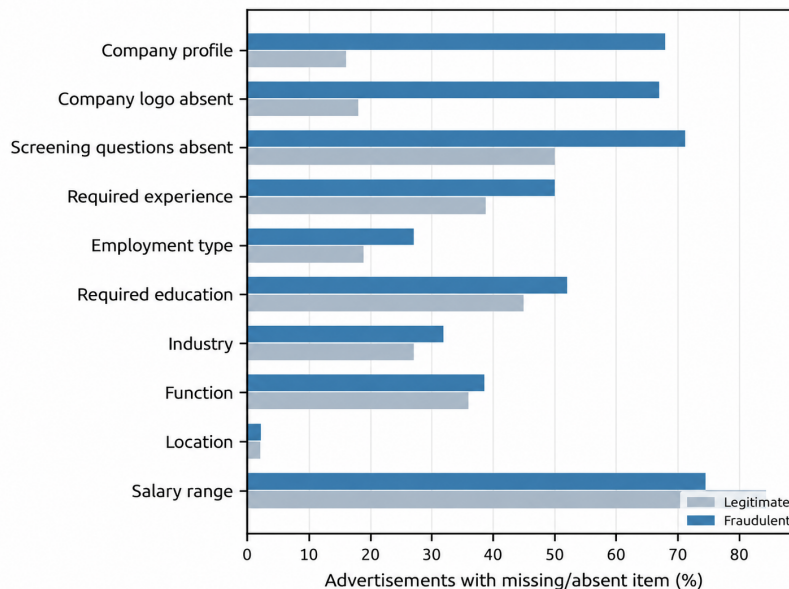


Figure 2: Observed missingness or absence rates by EMSCAD class after exact-duplicate removal. Differences are descriptive associations, not evidence of intent.

5.3. Interpreting the Missingness Mechanism

EMSCAD does not identify why a field was left blank, so MCAR, MAR, and MNAR cannot be established from the observed table alone. MCAR would be plausible for sporadic ingestion failures unrelated to advertisement content. MAR could occur when salary disclosure depends on observed industry, country, or employment type. MNAR could occur if an organization withholds company details because the unreported identity would reveal deception; it could equally occur when a legitimate firm withholds pay because the unreported value is commercially sensitive. Co-occurring absence of profile, logo, and screening questions may also reflect a particular account configuration or bulk-import template. The experiment therefore estimates predictive association under the historical data-generating process. It neither recovers the causal mechanism nor assigns intent to an individual employer.

6. Simulation Methodology

6.1. Models and Leakage Controls

The simulation is implemented with pandas, NumPy, Matplotlib, and scikit-learn [13]. A fixed random seed of 42 creates a stratified 80/20 split: 14,040 training and 3,510 test records; the test set includes 171 fraud and 3,339 legitimate examples. All learned transformations are inside pipelines. Consequently, the TF-IDF vocabulary, inverse-document frequencies, and quality-feature scaling are fitted only on training data.

Model 1 applies TF-IDF to the five-field combined text and logistic regression. Model 2 applies standardized logistic regression to 16 quality features: ten field indicators; missing, available, and ratio summaries; combined-text and description lengths; and a fixed flag for descriptions shorter than 100 characters. Model 3 joins both branches through a column transformer before the same classifier. The TF-IDF configuration uses English stop-word removal, unigrams and bigrams, `min_df=3`, `max_df=0.98`, sublinear term frequency, and at most 20,000 features. All classifiers use L_2 -regularized logistic regression, $C = 1$, a 0.5 threshold, and `class_weight=balanced`. Class weighting is a simple response to minority rarity [14]; no resampling or holdout-driven threshold tuning is performed. Table III records the settings.

Table 3: Experimental settings.

Component	Setting
Split	Stratified 80/20 holdout; seed 42
Text fields	Title, profile, description, requirements, benefits
TF-IDF	1–2 grams; English stops; min df 3; max df 0.98; max 20,000
Quality features	10 absence flags + 6 count/ratio/length indicators
Classifier	Logistic regression; L2; $C = 1$; liblinear; 1,000 iterations
Imbalance	<code>class_weight=balanced</code>
Decision rule	Fraud probability ≥ 0.5
Hardware need	CPU only; no GPU or specialized accelerator

6.2. Reproducibility and Audit Outputs

The supplied script executes data validation, deduplication, feature construction, fitting, evaluation, and plotting in one command. It stops with an explicit error if required columns, binary labels, or the dataset file are absent. A SHA-256 digest records the executed CSV copy. Machine-readable outputs include the raw and analysis class counts, split composition, configuration, metrics, confusion counts, missingness rates, and standardized quality-only coefficients. CSV tables and PDF/PNG figures are written from the same in-memory results used to create the LaTeX macros. Consequently, the manuscript imports its dataset counts, rates, and performance values instead of duplicating them by manual transcription.

This design supports two audits. A computational audit can rerun the script and compare the JSON, CSV, and figure outputs. A methodological audit can inspect whether the deduplication key excludes the target, whether transformations are fitted inside pipelines, and whether the test set is used only for final measurement. The code intentionally avoids cross-validated tuning, synthetic oversampling, neural hardware, and external feature services. These exclusions make the simulation accessible on a normal laptop and narrow the number of researcher choices that could otherwise be optimized after viewing test performance.

6.3. Evaluation

For the fraudulent class, with true positives TP , false positives FP , false negatives FN , and true negatives TN ,

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F_1 = \frac{2PR}{P + R} \quad (7)$$

$$BAcc = \frac{1}{2} \left(\text{Recall} + \frac{TN}{TN + FP} \right) \quad (8)$$

Average precision (reported as PR-AUC) summarizes the precision–recall ranking as $AP = \sum_n (R_n - R_{n-1}) P_n$. Precision–recall analysis is preferred to accuracy-centric interpretation when positives are rare [15]; balanced accuracy gives equal importance to sensitivity and specificity [16]. Accuracy is retained only as a secondary value. No confidence intervals or model-selection claims are made from the single holdout.

7. Results and Discussion

7.1. Ablation Results

Table IV and Fig. 3 report the executed holdout results. The text baseline was strongest in precision (0.673) and F1 (0.743). Quality-only logistic regression attained recall 0.725 and balanced accuracy 0.779, showing that completeness patterns were associated with the label, but its precision 0.181, F1 0.290, and PR-AUC 0.179 were weak. It is therefore not suitable as a standalone decision rule.

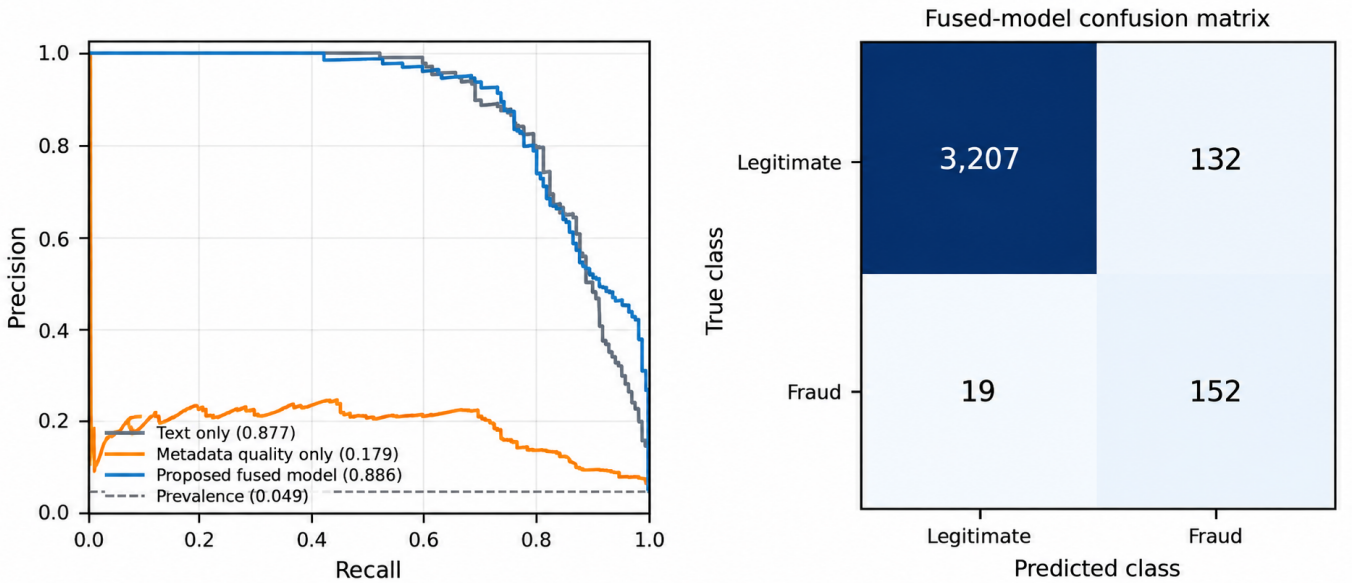


Figure 3: Holdout precision–recall curves (legend gives PR-AUC) and the fused-model confusion matrix.

Table 4: Holdout model performance (fraudulent class).

Model	Prec.	Recall	F1	BAcc	PR-AUC
Text only	0.673	0.830	0.743	0.905	0.877
Quality only	0.181	0.725	0.290	0.779	0.179
Proposed fused	0.535	0.889	0.668	0.925	0.886

Fusion increased recall from 0.830 to 0.889: it found 152 of 171 fraud cases rather than 142. Balanced accuracy increased from 0.905 to 0.925, and PR-AUC from 0.877 to 0.886. However, false positives rose from 69 to 132, reducing

precision from 0.673 to 0.535 and F1 from 0.743 to 0.668. The confusion matrices (TN, FP, FN, TP) were text (3, 270, 69, 29, 142), quality-only (2, 779, 560, 47, 124), and fused (3, 207, 132, 19, 152). Thus, the fused model does not dominate the baseline; it trades more review workload for 10 additional detected fraud cases.

The conclusion also changes with the chosen metric. Overall accuracy favors text only (0.972) over fusion (0.957), but this difference largely reflects the legitimate majority and the fused model’s additional false positives. Fraud recall favors fusion, F1 favors text, and PR-AUC shows only a small ranking gain for fusion. Reporting one of these measures in isolation could therefore support contradictory narratives. The ablation is most defensible when the intended operating cost is stated first: missed scams motivate recall, reviewer workload motivates precision, and ranking quality for a limited queue motivates PR-AUC. The fixed 0.5 threshold is a reproducibility choice, not a universally optimal policy.

7.2. Interpretation and Practical Value

The largest class differences in Fig. 2 concern company identity and interaction structure: missing profiles, absent logos, and absent screening questions. Required experience is also more often missing for fraud, whereas salary absence points in the opposite direction. These patterns are plausible complements to wording, but they may encode historical platform practice or campaign templates rather than a transferable property of fraud. The standardized quality-only coefficients saved by the script provide a second audit view, but coefficients are associative and affected by correlated count and ratio variables.

Is the gain worth the added complexity? Computationally, appending 16 numeric features to a sparse text matrix is minor. Operationally, the answer depends on review capacity and error costs. For recall-oriented queueing, 10 fewer false negatives may justify 63 more false positives in this holdout. For automatic removal, the precision and F1 loss argue against fusion at the fixed threshold. A platform should select a threshold and workflow using validation data, reviewer capacity, and explicit costs; this experiment intentionally does not tune the threshold on the test set.

7.3. Ethical and Operational Considerations

The framework is decision support, not autonomous enforcement. A high score must lead to proportionate human review rather than a public accusation or immediate employer ban. Reviewers should see which text and completeness signals contributed, have access to corroborating evidence, and record outcomes. Employers need an appeal and correction path. Data collection should minimize applicant and recruiter personal information, enforce retention limits, and secure access.

Missingness can systematically burden small, new, or less-resourced organizations. Dataset documentation recommended by Gebru *et al.* [17] and NLP data statements [18] motivate recording collection platform, period, language, geography, annotation process, and missingness mechanisms. Subgroup error audits are necessary where lawful attributes or defensible proxies exist; broader fairness literature warns that apparently neutral variables can reproduce structural disparities [19]. No single absent field should trigger removal, and correlation between absence and fraud does not prove concealment.

8. Limitations and Threats to Validity

Construct validity: EMSCAD labels operationalize the provider’s historical fraud judgments, not every modern form of recruitment deception. The data omit later conversations, payment behavior, external company verification, and applicant outcomes. “Missing” combines true nulls, blank strings, and two binary absence flags; these meanings are related but not identical.

Internal validity: Exact duplicates were removed before splitting, but undetected templates and near duplicates may remain. The fixed short-description threshold and correlated aggregate features are simple design choices. Only one stratified holdout and limited hyperparameter settings were used; random variation is not quantified. Class weighting changes the operating point, and results would differ under threshold tuning or another cost model.

External validity: The corpus is historical, English-language, and platform-specific. Results may not generalize across countries, industries, languages, small-business practices, or contemporary form designs. Salary disclosure norms in particular vary. The use of one dataset and absence of prospective evaluation preclude production-level claims.

Temporal validity: Scammers adapt wording and completion strategies, and platforms alter fields. Concept-drift research shows why monitoring and adaptation are necessary when data-generating processes change [20]. A deployment study would need time-ordered validation, drift alarms for text and missingness, periodic re-annotation, recalibration, and rollback procedures. The present framework has not undergone these steps.

9. Conclusion and Future Work

This paper proposed and tested a transparent two-branch framework for fraudulent-job-advertisement detection. On a deduplicated EMSCAD holdout, missing-metadata and quality signals were predictive by themselves but substantially weaker than text. Fusion produced a modest increase in recall, balanced accuracy, and PR-AUC while decreasing precision and F1. The central conclusion is therefore conditional: completeness patterns can help a recall-sensitive triage system, but they are not a safe standalone heuristic and do not guarantee better overall classification.

Future work should evaluate repeated stratified and time-ordered splits, near-duplicate grouping, probability calibration, and explicit cost-sensitive thresholds. Cross-platform and multilingual validation should document field optionality and recruitment norms. Human-in-the-loop studies should measure reviewer burden, appeal outcomes, and subgroup error rates. These steps are prerequisites before any deployment claim or enforcement use.

Author Contributions

Karthik Juluri (K.J): Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft; Krishna Kant Pandey (K.K.P): Conceptualization, Validation, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The data that support the findings of this study are available from the corresponding author, K.K.P, upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors did not use any generative AI or AI-assisted technologies.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- [1] S. Vidros, C. Koliass, G. Kambourakis, and L. Akoglu, “Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,” *Future Internet*, vol. 9, no. 1, p. 6, 2017, doi: 10.3390/f9010006.
- [2] B. Alghamdi and F. Alharby, “An intelligent model for online recruitment fraud detection,” *Journal of Information Security*, vol. 10, no. 3, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.
- [3] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, “ORFDetector: Ensemble learning based online recruitment fraud detection,” in *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 2019, pp. 1–5, doi: 10.1109/IC3.2019.8844879.
- [4] J. Li, Y. Li, H. Han, and X. Lu, “Exploratory methods for imbalanced data classification in online recruitment fraud detection: A comparative analysis,” in *Proceedings of the 2021 4th International Conference on Computing and Big Data*, 2021, pp. 75–81, doi: 10.1145/3507524.3507537.
- [5] S. Mahbub, E. Pardede, and A. S. M. Kayes, “Online recruitment fraud detection: A study on contextual features in Australian job industries,” *IEEE Access*, vol. 10, pp. 82776–82787, 2022, doi: 10.1109/ACCESS.2022.3197225.

- [6] K. Nanath and L. Olney, “An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud,” *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100167, 2023, doi: 10.1016/j.jjime.2023.100167.
- [7] S. Seaman, J. Galati, D. Jackson, and J. Carlin, “What is meant by ‘missing at random’?” *Statistical Science*, vol. 28, no. 2, pp. 257–268, 2013, doi: 10.1214/13-STS415.
- [8] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, “A survey on missing data in machine learning,” *Journal of Big Data*, vol. 8, p. 140, 2021, doi: 10.1186/s40537-021-00516-9.
- [9] M. Saar-Tsechansky and F. Provost, “Handling missing values when applying classification models,” *Journal of Machine Learning Research*, vol. 8, pp. 1623–1657, 2007.
- [10] R. Mitra, S. F. McGough, T. Chakraborti, C. Holmes, R. Copping, N. Hagenbuch, S. Biedermann, J. Noonan, B. Lehmann, A. Shenvi, X. V. Doan, D. Leslie, G. Bianconi, R. Sanchez-Garcia, A. Davies, M. Mackintosh, E.-R. Andrinopoulou, A. Basiri, C. Harbron, and B. D. MacArthur, “Learning from data with structured missingness,” *Nature Machine Intelligence*, vol. 5, no. 1, pp. 13–23, 2023, doi: 10.1038/s42256-022-00596-z; arXiv:2304.01429.
- [11] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
- [12] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x; arXiv:1811.10154.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] P. Branco, L. Torgo, and R. P. Ribeiro, “A survey of predictive modeling on imbalanced domains,” *ACM Computing Surveys*, vol. 49, no. 2, pp. 31:1–31:50, 2016, doi: 10.1145/2907070; arXiv:1505.01658.
- [15] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.
- [16] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The balanced accuracy and its posterior distribution,” in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124, doi: 10.1109/ICPR.2010.764.
- [17] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daume III, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723; arXiv:1803.09010.
- [18] E. M. Bender and B. Friedman, “Data statements for natural language processing: Toward mitigating system bias and enabling better science,” *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018, doi: 10.1162/tacl_a_00041; ACL Anthology: Q18-1041.
- [19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 115:1–115:35, 2021, doi: 10.1145/3457607; arXiv:1908.09635.
- [20] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, “Learning under concept drift: A review,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2019, doi: 10.1109/TKDE.2018.2876857; arXiv:2004.05785.