

Comparative Study of TF-IDF and SBERT Feature Representations for Random Forest-Based Resume Classification

Fathya Fathimah Azzahra¹⁾, Roni Andarsyah²⁾*, Mohamad Nurkamal Fauzan³⁾

^{1,2,3)} School of Information Technology, Universitas Logistik dan Bisnis Internasional, Bandung, Indonesia

¹⁾ 714220012@std.ac.ulbi.id, ²⁾ roniandarsyah@ulbi.ac.id, ³⁾ m.nurkamal.f@ulbi.ac.id

Submitted : Aug 18, 2026 | Accepted : Sep 7, 2026 | Published : Oct 1, 2026

Abstract: Manual resume screening in high-volume recruitment is time-consuming and prone to inconsistent evaluation, motivating automated screening using Machine Learning (ML) and Natural Language Processing (NLP). Prior comparisons of TF-IDF and Sentence-BERT (SBERT) often differ in classifiers or task formulations, limiting direct comparison of feature representations. This study compares TF-IDF and SBERT using an identical Random Forest classifier for resume-to-role classification across 32 role categories. The dataset contains 10,102 resume-job description pairs from a public Hugging Face corpus. Three sources of data leakage were identified and mitigated, followed by five-fold StratifiedGroupKFold cross-validation and Bonferroni-corrected paired statistical testing. TF-IDF Standard achieved a mean macro F1 of 0.9953, TF-IDF Enriched 0.9964, and SBERT 0.9816. The difference between the two TF-IDF configurations was not significant ($p = 0.1367$), whereas both significantly outperformed SBERT ($p = 0.0021$ and $p = 0.0015$). Component-wise ablation showed that resume-only input retained high performance (macro F1 up to 0.9977), while job-description-only input achieved at most 0.2339, indicating that resume content provides the primary predictive signal. Equalizing the input budget to 256 tokens also preserved the TF-IDF advantage. Role-mismatch results further showed no performance degradation when job descriptions were replaced with descriptions from unrelated roles. These findings indicate that TF-IDF provides strong lexical representations for resume-to-role classification under the evaluated protocol, while highlighting the importance of leakage auditing and controlled input conditions.

Keywords: Data Leakage; Feature Representation; Random Forest; Resume Classification; Sentence-BERT; TF-IDF

INTRODUCTION

In Indonesia, the labor market continues to face a persistent gap between job seekers and available formal openings: the open unemployment rate stood at 4.85% in August 2025, corresponding to approximately 7.46 million unemployed individuals within a labor force of 154 million (Badan Pusat Statistik, 2025). This gap manifests directly at the recruiter's desk: a single job posting in Indonesia typically attracts 500 to 600 applications, with popular openings at large companies drawing several thousand (Nadira, 2026), a pattern echoed across many labor markets globally. Resume screening is therefore one of the most resource-intensive stages in human resource management, particularly in high-volume recruitment where organizations may receive hundreds to thousands of applications for a single opening. The growth of online recruitment platforms has made manual filtering and classification of such large applicant pools increasingly costly and time-consuming for hiring organizations, motivating the recruitment industry to pursue automated screening and classification solutions (Surendiran et al., 2023). Traditional workflows that rely heavily on manual resume parsing and matching further result in long cycle times, high cost, and low match accuracy (Wang & Zhao, 2025). Machine Learning (ML) and Natural Language Processing (NLP) have therefore been widely adopted to automate this process, with a clear technical trajectory from classical feature representations such as TF-IDF, achieving up to 91.38% accuracy (Surendiran et al., 2023) and 0.857 accuracy (Beristain et al., 2024), toward Transformer-based sentence embeddings such as SBERT and large language models achieving up to 92.0% accuracy on large-scale benchmarks (Heakl et al., 2024).

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Despite this progression, whether TF-IDF or SBERT provides the stronger feature representation for resume classification, when classifier architecture, dataset, and evaluation protocol are held constant, remains empirically unresolved. Prior comparative studies either vary the classifier alongside the feature representation or evaluate SBERT only as a similarity-ranking mechanism rather than as classifier input features, making it impossible to isolate feature representation's independent contribution; the Literature Review elaborates this gap with supporting citations.

This study presents a controlled empirical comparison of TF-IDF and SBERT as feature representations for an identical Random Forest classifier in a resume+job description-to-Role classification across 32 role categories using a public Hugging Face resume and job description corpus.

During pipeline development, an unexpected observation revealed that the initial classification performance was dominated by data leakage instead of genuine predictive capability. Initial classification runs achieved a near-perfect macro F1-score exceeding 0.999 across all three feature conditions. Inspection of feature importance rankings revealed that the highest-contributing features were role name fragments rather than skill-related terms, indicating label leakage instead of genuine classification capability. Three compounding leakage sources were identified and corrected through a three-layer mitigation pipeline, detailed in the Method section. Because job description text is deliberately combined with resume text as model input, a dedicated ablation isolating each document type's independent contribution was conducted to verify that classification performance reflects resume-skill alignment rather than incidental job description leakage.

Beyond its methodological contribution, an accurate, computationally efficient resume classifier is directly relevant to practical recruitment workflows, where model output can support, rather than replace, a human reviewer's screening process. Under such use, the choice between TF-IDF and SBERT carries a concrete trade-off: TF-IDF is computationally lightweight, while SBERT demands greater overhead at inference time. An empirically grounded, statistically validated answer to whether this cost yields a meaningful performance gain is therefore of direct practical value.

This study makes three primary contributions to resume classification research. First, to the best of the authors' knowledge, it provides a controlled comparison between TF-IDF and SBERT under an identical Random Forest classifier, identical dataset, and identical evaluation protocol, thereby isolating the effect of feature representation. Second, it introduces a three-stage data leakage mitigation framework addressing role mentions, duplicated job descriptions, and resume self-identification. Third, it demonstrates through cross-validated statistical significance testing that apparent performance improvements observed from a single train-test split do not necessarily represent meaningful performance gains.

LITERATURE REVIEW

Previous studies have evaluated TF-IDF with classical classifiers such as Random Forest and Logistic Regression (Beristain et al., 2024; Surendiran et al., 2023), BERT and LLM architectures fine-tuned on resume text (Heakl et al., 2024), SBERT-based similarity matching (Kaya, 2023), and TF-IDF combined with Logistic Regression after summarization preprocessing (Maree, 2025). A direct comparison between BERT and SBERT has also been reported through similarity ranking (Deshmukh & Dahake, 2025). Other studies tested SBERT only as a similarity baseline (Ajjam & Al-Raweshidy, 2026), or combined TF-IDF with BERT features (Kumar et al., 2026), instead of isolating SBERT as a standalone feature representation for a classical classifier.

TF-IDF represents documents as sparse vectors weighted by term frequency and inverse document frequency, a representation that captures exact keyword and terminology matches without requiring model training, making it particularly effective for domain-dense technical text such as resumes (Ajjam & Al-Raweshidy, 2026). SBERT extends the BERT architecture with a siamese network structure to produce fixed-dimensional sentence embeddings independently for each input, enabling efficient semantic similarity computation compared to standard BERT, which requires joint processing of sentence pairs (Deshmukh & Dahake, 2025). Random Forest, an ensemble of decision trees trained through bootstrap aggregating with random feature subset selection at each split, was selected as the classifier across all conditions in this study due to its established use as a resume classification baseline (Surendiran et al., 2023) and its capacity to expose feature importance scores for verifying that predictions are not driven by residual label leakage (Alsubaie & Aleisa, 2025).

Table 1 summarizes this comparison against related studies, showing that previous studies investigated TF-IDF classification, SBERT-based matching, or hybrid feature fusion separately, with none directly comparing TF-IDF and SBERT under an identical Random Forest classifier and evaluation protocol.

Table 1 Comparison with Related Studies

Study	TF-IDF	SBERT	RF	Feature Repr. Comp.	Stat. Test	Feat. Import.	Resume Task
(Surendiran et al., 2023)	Yes	-	Yes	-	-	-	Yes
(Beristain et al., 2024)	Yes	-	Yes	-	-	-	Yes

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

(Heakl et al., 2024)	Yes	-	Yes	-	-	-	Yes
(Kaya, 2023)	-	-	Yes	-	-	-	Yes*
(Maree, 2025)	Yes	-	-	-	-	-	-
(Deshmukh & Dahake, 2025)	-	Yes	-	-	-	-	-
(Ajjam & Al-Raweshidy, 2026)	Yes	Yes	-	Yes**	Yes	-	-
(Kumar et al., 2026)	Yes	-	Yes	-	-	Yes	Yes
(Chafi et al., 2025)	Yes	-	-	-	-	-	Yes
(Alsubaie & Aleisa, 2025)	-	-	Yes	-	-	Yes	Yes***
Proposed	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* Resume-to-job similarity ranking using SBERT embeddings; no feature representation comparison. (Kaya et al., 2023)

** Compared TF-IDF and SBERT for cosine similarity retrieval, not as classifier feature representations. (Ajjam & Al-Raweshidy, 2026)

*** Resume hiring-decision classification from structured demographic features, not free-text resume content. (Alsubaie & Aleisa, 2025)

Beyond direct classification, related tasks have also been studied extensively. For resume parsing and named entity recognition, He (2023) applied a BiLSTM-CRF architecture, Liu et al. (2021) proposed a CRNN-based multi-label model, and Luo (2022) introduced a lexically enhanced LSTM approach, primarily for structured field extraction. For candidate-job matching, Łepicki et al. (2025), explored Siamese networks, Liu (2025), proposed a Dual-Tower BERT architecture, Baghbanzadeh and Wu (2025) combined graph neural networks with an LLM, and Rosenberger et al. (2025) developed CareerBERT, a domain-adapted embedding space. These parallel directions confirm the breadth of ML and NLP adoption in resume-related tasks, though none addresses the specific gap identified in Table 1.

This progression is echoed elsewhere in the literature. Mishra et al. (2021) reported 96.96% accuracy using Stacked KNN with hard voting for IT recruitment screening, and Ujlayan et al. (2023), proposed a TF-IDF-based ensemble framework combined with topic modeling for recruitment optimization, demonstrating that classical ML remains competitive. He (2023) and Luo (2022) advanced structured extraction further, achieving 92.98% and 96.26% F1-score respectively for Chinese resume named entity recognition using BiLSTM-CRF and lexically enhanced LSTM. At the Transformer and LLM tier, Chafi et al. (2025). combined SimCSE, LLaMa, and RoBERTa in a hybrid contrastive-generative pipeline, reaching 91.7% accuracy on automatic CV classification. Yet across this breadth, no study isolates feature representation choice, specifically TF-IDF versus SBERT, as the sole controlled variable within an identical classifier and evaluation protocol, the gap this study addresses.

METHOD

This study follows a controlled experimental pipeline designed to isolate the effect of feature representation on Random Forest classification performance, illustrated in Fig. 1.

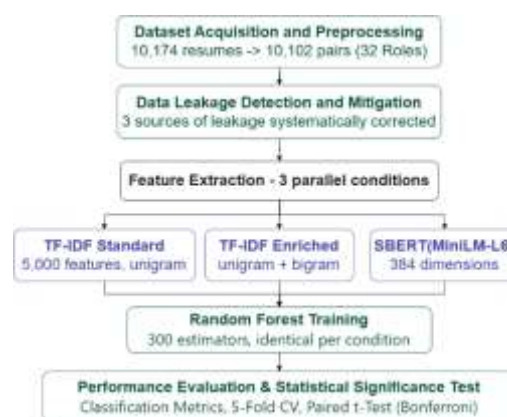


Fig. 1 Research Methodology Flowchart

Fig. 1 illustrates each stage of this pipeline. The pipeline begins with dataset acquisition and preprocessing, reducing the initial 10,174 resume-job description records to 10,102 pairs across 32 role categories. The cleaned data then passes through data leakage detection and mitigation, correcting three identified leakage sources through role-name masking and group-based train-test splitting. The masked text is subsequently transformed under three parallel feature-extraction conditions, TF-IDF Standard (5,000 features, unigram), TF-IDF Enriched (unigram and bigram), and SBERT (all-MiniLM-L6-v2, 384 dimensions), each feeding an identically configured Random Forest

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

classifier (300 estimators). Each condition is finally evaluated using classification metrics, five-fold cross-validation, and Bonferroni-corrected paired t-tests, isolating the effect of feature representation on classification performance while holding classifier architecture constant.

Dataset and Class Distribution

This study utilizes the Resume-Screening-Dataset, a publicly available corpus hosted on Hugging Face under the MIT license (AzharAli05, n.d.), accessed on April 2, 2026, comprising 10,174 resume-job description records at the time of access.

Table 2 Dataset Attribute Description

Attribute	Type	Description
Role	Categorical	Job role label; classification target (45 values before filtering, 32 after ≥ 20 -sample threshold, Table 3)
Resume	Free Text	Candidate resume text (29-4,680 characters)
Job_Description	Free Text	Job posting text; concatenated with Resume as model input (Eq. 1) (69-5,310 characters)
Decision	Categorical	Hiring outcome label ("select"/"reject"); not used in this study
Reason_for_decision	Categorical	Justification text for Decision (539 distinct values); not used in this study

Only Role, Resume, and Job_Description were used in this study, as classification target and model input, respectively; Decision and Reason_for_decision are reported here for completeness but were not part of the classification pipeline.

The dataset repository provides no documented collection methodology (its README is empty). However, the raw text shows converging signs of synthetic, LLM-assisted generation: most Resume entries begin with a generation artifact (e.g., "Here's a professional resume for [Name]:"), Job_Description entries recur through a small set of templated phrasings, and Reason_for_decision draws from only 539 distinct strings across 10,174 records. This pattern is inconsistent with authentic, independently written postings or resumes, and is reported here as an inference from the data rather than a disclosed methodology.

Prior to feature extraction, job description and resume text fields were concatenated into a single field following Equation (1):

$$\text{CombinedText} = \text{JD}_{\text{clean}} [\text{SEP}] \text{Resume}_{\text{clean}} \quad (1)$$

Job descriptions averaged approximately 52 words and were placed first to accommodate SBERT's 256-token limit, while resumes averaged approximately 350 words. Prior to filtering, the dataset contained 38 role categories. Six categories with fewer than 20 samples (network engineer, graphic designer, ai engineer, hr specialist, cybersecurity specialist, ui/ux designer) were excluded to ensure reliable per-class evaluation, yielding a final dataset of 10,102 pairs across 32 role categories.

As summarized in Table 3, the class distribution exhibits moderate imbalance (imbalance ratio 14.73), spanning three tiers: 6 large classes (507–825 samples), 24 mid-sized classes (220–270 samples), and 2 small classes (Project Manager: 56, Software Developer: 66). This distribution motivated the class-weighted Random Forest configuration described below.

Table 3 Class Distribution Summary (32 Final role categories)

Statistic	Value
Total samples	10,102
Largest class	Data Scientist (825 samples)
Smallest class	Project Manager (56 samples)
Imbalance ratio (max/min)	14.73
Classes with 500+ samples	6
Classes with 220–270 samples	24
Classes with under 100 samples	2

Data Leakage Mitigation

In initial runs, the model reached an unnaturally high macro F1-score above 0.999 because feature importance was heavily dominated by role-name words rather than actual skills. To fix this leakage, three sequential corrections were applied before feature extraction. First, the original filter requiring job descriptions to explicitly include the target role name was removed, as it directly leaked the answer into the input text. Second, duplicate

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

job descriptions (which made up 77.8% of records, or 3,374 unique descriptions across 10,102 rows) were handled using GroupShuffleSplit with the cleaned job description text as the grouping key, making sure identical descriptions never overlapped between training and test sets. Third, since 98.5% of resumes stated their target role name in the first sentence, a masking function stripped all 32 predefined role-name strings from the combined text before feature extraction.

These corrections were kept consistent across all experimental setups. Distribution checks confirmed that all 32 classes were still represented in the test set, with the smallest class retaining 18 samples. However, since the masking function strictly targeted the 32 exact role-name strings, it does not clean out semantically related role variants or close synonyms

Feature Representation Summary

Three feature representation conditions were constructed from the masked combined text and applied consistently across all experiments. Both TF-IDF conditions used scikit-learn's TfidfVectorizer with max_features=5000 and no upper document-frequency bound (max_df=1.0). Prior to vectorization, text was lowercased, non-alphabetic characters were removed, tokenized using NLTK's word_tokenize, lemmatized with WordNet, and filtered to tokens longer than two characters. TF-IDF Standard removed standard NLTK English stopwords and used ngram_range=(1, 1) (unigrams only), with no minimum document-frequency threshold (min_df=1) and no sublinear term-frequency scaling. TF-IDF Enriched extended this pipeline with an additional domain-specific stopwords list {experience, experienced, strong, results, driven, professional, skilled, motivated, detail, oriented, proven, track, record, summary, year, years, sep combined with ngram_range=(1, 2) (unigrams and bigrams), sublinear_tf=True, and min_df=2. SBERT used sentence embeddings from all-MiniLM-L6-v2 with a maximum sequence length of 256 tokens and no stopword removal or lemmatization, applied directly to the masked combined text.

The TF-IDF weight for term t in document d within corpus D follows Equation (2):

$$\text{TF-IDF}(t, d, D) = \text{tf}(t, d) \times \log \left(\frac{|D|}{\text{df}(t)} \right) \quad (2)$$

SBERT embeddings follow the sentence embedding approach established for semantic similarity tasks in prior resume screening work (Deshmukh & Dahake, 2025), used directly as numerical input features without intermediate vectorization.

Random Forest Classification

A Random Forest classifier was used across all three conditions to isolate the effect of feature representation while holding classifier architecture constant, consistent with its established use as a resume classification baseline (Surendiran et al., 2023). An identical configuration was applied across conditions: 300 estimators, random state 42, and balanced class weighting (class_weight='balanced'), applied given the 14.73 imbalance ratio reported above. An ablation comparing this configuration against an unweighted baseline, evaluated on the TF-IDF Enriched condition, showed a small performance difference (macro F1 without weighting: 0.9947; with balanced weighting: 0.9956; difference: +0.0009), with the largest effect observed for the smallest class, project manager, where recall improved from 0.94 to full agreement with weighting applied. This suggests that the domain-specific, largely non-overlapping technical vocabulary distinguishing most role categories, combined with the leakage mitigation described above, already yields a highly discriminative feature space largely independent of class weighting, though weighting still provides a measurable benefit for the smallest classes. No condition-specific hyperparameter optimization was performed to ensure a fair comparison across feature representations.

Performance Evaluation and Statistical Significance Testing

Model performance was evaluated on a held-out test set using accuracy, precision, recall, and F1-score, each macro-averaged across the 32 role categories to avoid bias toward more frequent classes.

Accuracy measures the proportion of correctly classified samples among all evaluated samples (3):

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \quad (3)$$

Precision measures the proportion of correctly predicted samples among all samples predicted as a given class(4):

$$\text{Precision}_{\text{macro}} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i} \quad (4)$$

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Recall measures the proportion of correctly predicted samples among all samples that actually belong to a given class(5):

$$\text{Recall}_{\text{macro}} = \frac{1}{N} \sum_{i=1}^N \frac{TP_i}{TP_i + FN_i} \quad (5)$$

F1-score is the harmonic mean of precision and recall(6):

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^N \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (6)$$

where TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative counts, respectively.

Beyond conventional single-split evaluation, this study assessed whether observed performance differences reflected genuine model superiority or sampling variation. Five-fold StratifiedGroupKFold cross-validation preserved both job-description grouping (preventing the leakage described above across folds) and class-proportional stratification. Paired t-tests were applied to the resulting per-fold macro F1-scores with Bonferroni correction for three pairwise comparisons (corrected $\alpha = 0.0167$). A Wilcoxon signed-rank test was additionally performed as a nonparametric cross-check; however, with only five folds, it is reported as a supporting rather than primary analysis due to its limited statistical power. A 32×32 confusion matrix was constructed for each condition to examine inter-class error patterns. Feature importance was analyzed to verify that predictions were driven by domain-relevant terminology instead of residual leakage (Alsubaie & Aleisa, 2025). Finally, the best-performing model was applied to one previously unseen resume-job description pair as an illustrative prediction.

Component-wise Ablation: Resume-only vs. Job Description-only

To verify that classification performance reflects resume-skill alignment rather than job description leakage, the masked combined text was split at its [SEP] separator into a job-description-only segment and a resume-only segment. Each segment was independently vectorized under the same TF-IDF Standard, TF-IDF Enriched, and SBERT configurations, classified using an identically configured Random Forest, and evaluated under the same GroupShuffleSplit and five-fold StratifiedGroupKFold protocol as the combined-input experiments.

Token-Length Equalization Ablation

Because SBERT embeddings are generated under a 256-token input limit while TF-IDF processes the full combined text, a further ablation equalized the input token budget across representations. The masked combined text was truncated to the first 256 tokens using the same tokenizer employed by the SBERT model (all-MiniLM-L6-v2), after which TF-IDF Standard and Enriched vectorizers were re-fit on the truncated text and evaluated under the identical Random Forest configuration and five-fold cross-validation protocol. SBERT results under this condition are identical to the combined-input SBERT results reported above, since SBERT embeddings are generated under the same 256-token limit regardless of condition.

Role-Mismatch Ablation

To further isolate the causal contribution of job description content beyond the resume-only/job-description-only decomposition, a role-mismatch ablation was conducted. Each resume retained its original role label, but its paired job description was replaced with a job description drawn from a different role category, using a fixed one-to-one, without-replacement assignment (random seed 42) generated once and reused across all folds. This assignment guaranteed that no job description was paired with its original role and that every job description was used exactly once. The mismatched job description and original resume were concatenated following the same [SEP] format, masked, and vectorized using the same TF-IDF Standard, TF-IDF Enriched, and SBERT configurations, then classified using an identically configured Random Forest and evaluated under the same five-fold StratifiedGroupKFold protocol, grouped by the original (matched) job description text to preserve consistency with the combined-input experiments.

RESULT

Effect of Data Leakage Correction

Prior to correction, preliminary role classification achieved a near-perfect macro F1-score above 0.999, with feature importance dominated by role name tokens rather than domain-relevant skills. Table 4 compares the highest-contributing TF-IDF features before and after leakage correction to illustrate this shift.

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Table 4 Top Features Before/After Correction

Rank	TF-IDF Standard (Before)	TF-IDF Standard (After)	TF-IDF Enriched (Before)	TF-IDF Enriched (After)
1	engineer	cloud	engineer	cloud
2	architect	database	game developer	database
3	cloud	data	highly devops	data
4	data	design	software engineer	design
5	developer	marketing	analyst	unity

After masking all 32 role name strings (not only each sample's own role label), the top-ranked features shifted entirely to domain-relevant skill terminology, with no residual role-name fragments detected in either condition.

Classification Performance Comparison

On the single held-out test split, TF-IDF Enriched reached the highest macro F1-score (0.9947), followed by TF-IDF Standard (0.9937) and SBERT (0.9834). Table 5 presents accuracy, macro precision, macro recall, and macro F1-score across the three conditions.

Table 5 Single-Split Classification Performance

Condition	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
TF-IDF Standard	0.9946	0.9946	0.9930	0.9937
TF-IDF Enriched	0.9960	0.9957	0.9940	0.9947
SBERT	0.9852	0.9887	0.9795	0.9834

Statistical Significance Testing

To evaluate model stability, a five-fold StratifiedGroupKFold cross-validation was performed. Table 6 summarizes the macro F1-scores alongside paired t-test results under Bonferroni correction.

Table 6 Cross-Validation Results and Statistical Significance Testing

Fold	Standard	Enriched	SBERT
1	0.9948	0.9961	0.9853
2	0.9966	0.9959	0.9848
3	0.9957	0.9987	0.9851
4	0.9946	0.9959	0.9765
5	0.9950	0.9956	0.9763
Mean $\pm$ SD	0.9953 $\pm$ 0.0007	0.9964 $\pm$ 0.0011	0.9816 $\pm$ 0.0043

As shown in Table 6, the performance difference between TF-IDF Standard and Enriched ($p = 0.1367$) is not statistically significant, indicating sampling variation. However, both TF-IDF variants significantly outperform SBERT ($p < 0.01$). A supplementary Wilcoxon signed-rank test aligns with these findings, though it lacked significance due to the small fold sample size ($n=5$).

Component-wise and Token-Length Ablation

The component-wise ablation (Table 7) shows that resume-only classification achieved macro F1-scores comparable to, and for TF-IDF Enriched slightly exceeding, the combined-input results (0.9977 vs. 0.9964), confirming that resume content alone carries sufficient discriminative signal. In contrast, job-description-only classification collapsed to macro F1 scores between 0.2158 and 0.2339, an order of magnitude below both combined and resume-only results, though still above the random-guess baseline for 32 classes (≈ 0.031). This large and statistically consistent gap rules out job-description-driven leakage as an explanation for the reported performance.

Notably, TF-IDF Enriched significantly outperformed TF-IDF Standard under resume-only input ($p = 0.0051$), a difference not significant under combined input ($p = 0.1367$) or job-description-only input ($p = 0.1384$). This suggests Enriched preprocessing's bigram and domain-specific stopwords configuration provides a genuine advantage specific to resume text, one diluted when shorter, lexically generic job description text is concatenated into the same input.

To address the asymmetric text budget between representations, since SBERT truncates input at 256 tokens while TF-IDF processes the full combined text, an additional ablation truncated the combined text to the same 256-token window using the SBERT tokenizer prior to TF-IDF vectorization. 97.5% of samples exceeded this limit, confirming that most inputs are affected by truncation. Even under this equalized token budget, TF-IDF Standard-256 (mean F1 = 0.9946) and TF-IDF Enriched-256 (0.9945) both significantly outperformed SBERT

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

(0.9816; $p = 0.0017$ and $p = 0.0023$, respectively), confirming that TF-IDF's advantage over SBERT is not attributable to its access to longer text.

Table 7 Ablation Study: Cross-Validated Performance and Significance vs. SBERT

Input Source	Condition	Mean F1 ($\pm$ Std.)	p vs. SBERT ($\alpha=0.0167$)	Significant
Resume-only	TF-IDF Standard	0.9957 (± 0.0012)	0.0024	Yes
Resume-only	TF-IDF Enriched	0.9977 (± 0.0014)	0.0018	Yes
Resume-only	SBERT	0.9811 (± 0.0047)	-	-
JD-only	TF-IDF Standard	0.2339 (± 0.0173)	0.0003	Yes
JD-only	TF-IDF Enriched	0.2321 (± 0.0182)	0.0006	Yes
JD-only	SBERT	0.2158 (± 0.0156)	-	-
256-token budget	TF-IDF Standard-256	0.9946 (± 0.0019)	0.0017	Yes
256-token budget	TF-IDF Enriched-256	0.9945 (± 0.0014)	0.0023	Yes
256-token budget	SBERT-256 ^a	0.9816 (± 0.0043)	-	-

^a Identical to combined-input SBERT (Table 6), since SBERT operates under a fixed 256-token limit regardless of condition.

Role-Mismatch Ablation Results

If job description content leaked substantial role information, pairing a resume with a job description from an unrelated role should introduce conflicting label signal and degrade classification performance. Table 8 shows that this was not observed: macro F1 did not decrease under role-mismatch for any condition and instead increased significantly for TF-IDF Standard (0.9953 to 0.9978, $p = 0.0102$) and SBERT (0.9816 to 0.9892, $p = 0.0002$) after Bonferroni correction ($\alpha = 0.0167$), while the increase for TF-IDF Enriched (0.9964 to 0.9987) did not reach significance ($p = 0.0221$).

Table 8 Role-Mismatch Ablation: Matched vs. Mismatched Job Description Pairing

Condition	Matched Mean F1	Mismatch Mean F1	Δ	p-value	Significant ($\alpha=0.0167$)
TF-IDF Standard	0.9953	0.9978	+0.0025	0.0102	Yes
TF-IDF Enriched	0.9964	0.9987	+0.0023	0.0221	No
SBERT	0.9816	0.9892	+0.0076	0.0002	Yes

This pattern is inconsistent with job-description-driven leakage: a genuinely leaking job description would be expected to confuse the classifier when deliberately mismatched, whereas performance instead remained stable or modestly improved. A plausible explanation is that the job description contributes comparatively little discriminative signal relative to the resume regardless of role alignment, so that its replacement with an unrelated but still technically-worded description does not introduce meaningful conflicting signal; the resume-only ablation reported above, in which resume content alone reproduced or exceeded combined-input performance, is consistent with this interpretation. This finding should be read as complementary rather than causally definitive, as the exact mechanism behind the small performance increase was not further isolated in this study.

Confusion Matrix Analysis

Fig. 2 and Fig. 3 present the 32×32 confusion matrices for the three conditions, with predicted labels on the horizontal axis and true labels on the vertical axis.

Fig. 2 (a) Confusion Matrix, Condition 1: TF-IDF Standard + RF and Condition 2: (b) TF-IDF Enriched + RF Both conditions show a strongly dominant diagonal across all 32 role categories, consistent with the high macro F1-scores reported in Table 5. Misclassifications are sparse and concentrated among roles sharing overlapping technical vocabulary, most visibly between Data Engineer and Data Scientist, and between Cloud Architect and Cloud Engineer, reflecting genuine skill overlap between these roles rather than systematic model bias.

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

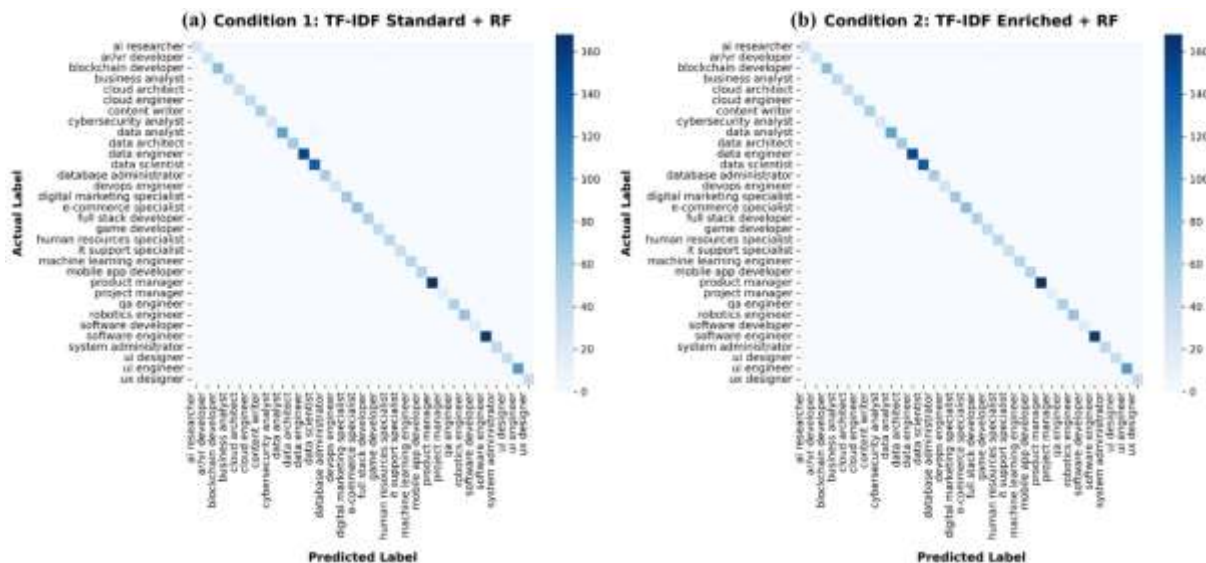


Fig. 2 (a) Confusion Matrix, Condition 1: TF-IDF Standard + RF and Condition 2: (b) TF-IDF Enriched + RF

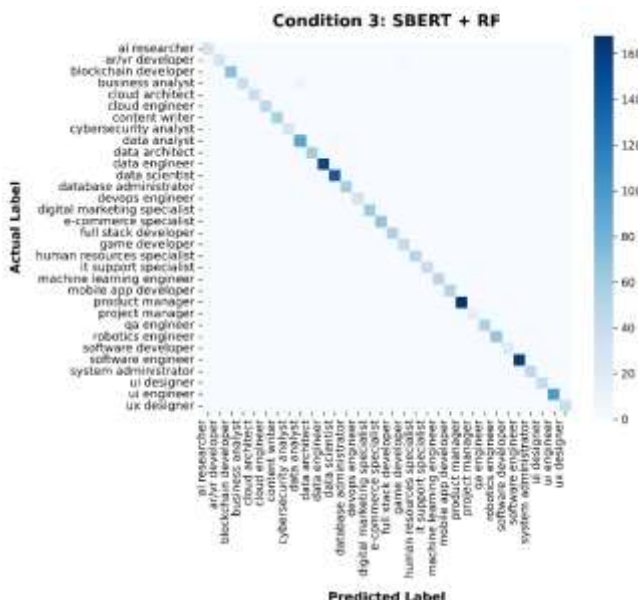


Fig. 3 Confusion Matrix, Condition 3: SBERT + Random Forest

Fig. 3 presents the confusion matrix for the SBERT condition. While the diagonal remains dominant, off-diagonal errors are visibly more dispersed than in the TF-IDF conditions (Fig. 2), consistent with SBERT's lower macro F1-score (Table 5). This pattern aligns with the interpretation that SBERT's dense semantic embeddings capture broader contextual similarity but are comparatively less sensitive to the discrete technical vocabulary that most strongly differentiates closely related roles.

SBERT Embedding Visualization

To examine how SBERT organizes the feature space independently of classifier performance, Fig. 4 visualizes a two-dimensional PCA projection of SBERT embeddings for the first 10 role categories. Roles with unique technical vocabularies, such as Blockchain Developer and Data Architect, form distinct clusters, while semantically adjacent roles, notably Business Analyst-Data Analyst and AI Researcher-Cybersecurity Analyst, show partial overlap. This overlap offers a structural explanation for the SBERT misclassification pattern in Fig. 3: conceptually similar roles are more easily confused in SBERT's semantic space than in TF-IDF's exact-match representation.

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

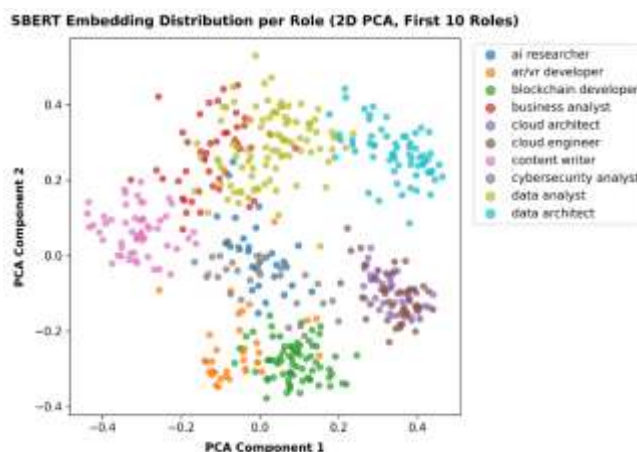


Fig. 4 SBERT Embedding Distribution per Role (2D PCA, First 10 Roles)

Out-of-Sample Validation

An illustrative out-of-sample evaluation using the TF-IDF Standard condition, selected because its performance was statistically equivalent to the Enriched configuration while using the simpler standard preprocessing configuration, was conducted on a previously unseen Blockchain Developer resume-job description pair (Table 9). The model ranked Blockchain Developer as the highest-probability role among the evaluated candidates. This example is consistent with the model's ability to identify role-relevant technical information in an unseen instance, but it should be interpreted as an illustrative prediction rather than evidence of generalization across a broader population.

Table 9 Out-of-sample Top-5 Performance

Rank	Predict Role	Probability
1	Blockchain Developer	77.00%
2	Software Engineer	4.67%
3	Data Engineer	2.67%
4	Game Developer	1.67%
5	Product Manager	1.67%

The model correctly ranked Blockchain Developer highest, while the remaining candidate roles shared overlapping technical infrastructure. This prediction provides an illustrative example of how the model distinguishes among technically related roles, although a single instance cannot establish genuine skill-based reasoning or broad generalization.

DISCUSSIONS

Interpretation of Comparative Results

Beyond the primary comparison, two additional ablations address plausible alternative explanations for the observed results. First, the resume/job-description component-wise ablation provides strong evidence against job-description-driven leakage as the primary source of model performance: resume-only classification achieved macro F1 scores of up to 0.9977, whereas job-description-only classification achieved macro F1 scores of no more than 0.2339. This indicates that the high classification performance is primarily supported by information contained in resumes rather than by job descriptions alone. Second, the token-length equalization ablation addresses the potential effect of unequal input lengths between TF-IDF and SBERT. Even when TF-IDF input was truncated to the same 256-token budget used for SBERT, TF-IDF continued to outperform SBERT, suggesting that its advantage cannot be explained solely by access to longer input text.

The comparative results indicate that, under the controlled and cross-validated evaluation protocol, both TF-IDF configurations significantly outperform SBERT for resume-to-role classification, while the enriched configuration does not provide a statistically significant improvement over the standard configuration. Across five folds, TF-IDF Standard achieved a mean macro F1 of 0.9953, compared with 0.9964 for TF-IDF Enriched and 0.9816 for SBERT. The paired t-test with Bonferroni correction found no significant difference between Standard and Enriched ($p = 0.1367$), whereas both Standard ($p = 0.0021$) and Enriched ($p = 0.0015$) significantly outperformed SBERT at the corrected significance level of 0.0167.

This aligns with Ajjam and Al-Raweshidy (2026), who found TF-IDF outperforming an SBERT baseline (all-MiniLM-L6-v2) in resume-job retrieval, attributed to TF-IDF's sensitivity to exact keyword matches salient in

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

technical resume text. Beristain et al. (2024) similarly reported TF-IDF with Random Forest as one of their strongest word-level configurations.

The results suggest that role-relevant signals in this dataset are concentrated in specific technical skill terms that TF-IDF can capture directly through lexical weighting. In contrast, SBERT represents text through dense contextual embeddings, which may not preserve discrete technical terminology as explicitly for this classification task. This interpretation is consistent with the feature importance patterns observed in the experiments, although the feature importance analysis does not by itself establish the underlying mechanism. The role-mismatch ablation reinforces this interpretation: because deliberately mispairing resumes with unrelated job descriptions did not degrade performance, the dominant predictive signal is attributable to resume content rather than to the job description matching the correct role.

This contrasts with Deshmukh and Dahake (2025), where SBERT outperformed BERT in similarity ranking; that comparison evaluates similarity-ranking mechanisms rather than classifier input features, and is therefore methodologically distinct rather than contradictory.

Practical Implications, Limitations, and Future Research

TF-IDF-based representations have a simpler computational pipeline than SBERT because they do not require pretrained neural embedding generation. However, this study did not directly benchmark runtime, memory usage, model size, or inference latency; therefore, computational efficiency should be interpreted as an architectural consideration rather than a measured finding. In terms of classification performance, the TF-IDF configurations achieved higher macro F1 scores than SBERT under the evaluation protocol used in this study. The magnitude of the performance change observed before and after leakage correction suggests that similar issues, particularly literal role-label mentions in input text and duplicate job descriptions spanning training and testing sets, may affect other resume classification studies that do not explicitly audit for these patterns.

Six limitations apply. First, the dataset appears, based on the textual patterns discussed above, to be synthetically LLM-generated rather than compiled from authentic postings or resumes; generalizability to real-world or non-English corpora remains untested, and reported performance should be read as an upper bound under templated, high-signal text. Second, the out-of-sample evaluation comprises a single illustrative case and therefore does not provide sufficient evidence to establish generalization across a broader set of unseen samples. Third, algorithmic fairness auditing was outside this study's scope; prior work documents persistent gender and disability-related biases in AI-based hiring tools even after mitigation (Glazko et al., 2024; Tyagi, 2024; Tyagi et al., 2024), alongside evidence that applicants and recruiters perceive AI-driven screening as less procedurally fair than human-led screening (Cai et al., 2024; Lacroux & Martin-Lacroux, 2022; Ling et al., 2025; Noble et al., 2021), underscoring this as an important direction for future systematic auditing. Fourth, a single Random Forest configuration was applied identically across all three feature representations to isolate the effect of feature representation; while this controls for classifier architecture, it does not guarantee that each representation was evaluated at its individually optimal hyperparameter setting, since sparse TF-IDF and dense SBERT feature spaces may favor different configurations. Representation-specific tuning, for example through nested cross-validation, was outside this study's scope and is a direction for future work. Fifth, role-name masking relied on exact case-insensitive string matching against the 32 known role labels; it does not detect variant forms such as abbreviations, pluralization, hyphenation, or synonymous job titles, so residual indirect role information cannot be fully ruled out despite the feature-importance and role-mismatch checks reported above. Sixth, computational efficiency claims for TF-IDF relative to SBERT are based on architectural characteristics rather than measured runtime, memory, or latency benchmarks, as noted above.

Beyond classical ML pipelines combining TF-IDF-derived features with ensemble methods (Mishra et al., 2021; Ujlayan et al., 2023), the present findings further support the competitiveness of lexical feature representations for resume classification. They also point to promising directions for future research, including broader corpora, more rigorous out-of-sample validation, and fairness-aware model auditing.

Beyond the immediate resume classification task, this study's methodology offers two transferable contributions. First, the three-layer leakage detection procedure provides a replicable diagnostic framework that other resume classification researchers can use to audit their pipelines before reporting results. The procedure considers literal role-label mentions, duplicate job descriptions across training and testing sets, and residual role-related information in the input text. Second, the comparison between the single-split and cross-validated results illustrates the importance of statistical validation. Although TF-IDF Enriched achieved a slightly higher single-split macro F1 than TF-IDF Standard (0.9947 vs. 0.9937), the difference was not statistically significant across five folds ($p = 0.1367$). This finding suggests that small performance differences observed from a single train-test split should be interpreted cautiously and should not automatically be treated as evidence that one representation is superior to another.

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

CONCLUSION

This study evaluated the effect of TF-IDF and SBERT feature representations on Random Forest performance for resume-to-role classification. Using 10,102 resume-job description pairs spanning 32 role categories, three sources of data leakage were identified and systematically corrected during pipeline development. Under the same Random Forest configuration, five-fold cross-validation showed that TF-IDF Standard achieved a mean macro F1 of 0.9953, TF-IDF Enriched achieved 0.9964, and SBERT achieved 0.9816. The Bonferroni-corrected paired t-tests found no statistically significant difference between TF-IDF Standard and Enriched ($p = 0.1367$), while both TF-IDF configurations significantly outperformed SBERT ($p = 0.0021$ and $p = 0.0015$, respectively). The component-wise ablation further showed that high performance was retained when only resume text was used, whereas job-description-only classification produced substantially lower macro F1 scores, providing evidence that resume content was the primary source of predictive signal. The token-length equalization experiment also showed that TF-IDF retained its advantage when its input was restricted to the same 256-token budget as SBERT, reducing the likelihood that the observed difference was caused solely by unequal input length. A role-mismatch ablation further supported this conclusion: replacing each job description with one drawn from an unrelated role did not degrade, and in two of three conditions significantly improved, classification performance, a pattern inconsistent with job-description-driven leakage.

Overall, the results provide cross-validated evidence that TF-IDF representations are highly effective for resume-to-role classification on the evaluated dataset, with no statistically significant performance advantage from the enriched configuration over the standard configuration. However, the findings should be interpreted within the scope of the dataset, evaluation protocol, and five-fold validation design. The single out-of-sample example provides only an illustrative prediction rather than a comprehensive test of generalization. Future work should evaluate the models on broader and organization-specific datasets, including non-English corpora, conduct systematic computational and fairness benchmarking, and further investigate input-pairing strategies to distinguish resume-based classification from information derived from job descriptions.

DECLARATION OF GENERATE AI USE

During the preparation of this manuscript, the authors used Claude (Anthropic) to assist with drafting and refining Python code and ChatGPT (OpenAI) for language editing and manuscript revision. All AI-assisted content and code were reviewed, verified, and validated by the authors. The research design, experiments, data analysis, model training, interpretation of results, and final conclusions were conducted and verified by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the manuscript.

REFERENCES

- Ajjam, M. H., & Al-Raweshidy, H. S. (2026). AI-driven semantic similarity-based job matching framework for recruitment systems. *Information Sciences*, 724(September 2025). <https://doi.org/10.1016/j.ins.2025.122728>
- Alsubaie, N., & Aleisa, N. (2025). Mitigating Bias in AI Model Using eXplainable AI in Terms of Hiring Process in the Industry. *IEEE Access*, 13(August), 147218–147241. <https://doi.org/10.1109/ACCESS.2025.3599947>
- AzharAli05. (n.d.). Resume-Screening-Dataset [Data set]. Hugging Face. <https://huggingface.co/datasets/AzharAli05/Resume-Screening-Dataset>
- Badan Pusat Statistik. (2025). *Tingkat Pengangguran Terbuka (TPT) sebesar 4,85 persen. Rata-rata upah buruh sebesar 3,33 juta rupiah.* <https://www.bps.go.id/id/pressrelease/2025/11/05/2479/tingkat-pengangguran-terbuka--tpt--sebesar-4-85-persen--rata-rata-upah-buruh-sebesar-3-33-juta-rupiah-.html>
- Baghbanzadeh, A., & Wu, D. (2025). Resume-Job Compatibility Scoring Using Graph Neural Networks and Large Language Models. In *2025 The 13th International Conference on Information Technology: IoT and Smart City (ICIT 2025), December 05â•fi07, 2025, Shanghai, China* (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3787330.3787359>
- Beristain, S. I., Barbosa, R. R. L., & Barriocanal, E. G. (2024). Improving jobs-resumes classification: a labor market intelligence approach. *International Journal of Information Technology & Decision Making*, 23(04), 1509–1525. <https://doi.org/10.1142/S0219622023500013>
- Cai, F., Zhang, J., & Zhang, L. (2024). The impact of artificial intelligence replacing humans in making human resource management decisions on fairness: A case of resume screening. *Sustainability*, 16(9), 3840. <https://doi.org/10.3390/su16093840>
- Chafi, S., Kabil, M., & Kamouss, A. (2025). Optimizing Automatic CV Classification with Contrastive and Generative Learning. *Procedia Computer Science*, 265, 342–349. <https://doi.org/10.1016/j.procs.2025.07.190>
- Deshmukh, A., & Dahake, A. R. (2025). Comparing bidirectional encoder representations from transformers and sentence-BERT for automated resume screening. *IAES International Journal of Artificial Intelligence*, 14(4), 3404–3411. <https://doi.org/10.11591/ijai.v14.i4.pp3404-3411>

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

- Glazko, K., Mohammed, Y., Kosa, B., & Mankoff, J. (2024). Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, June 03â•fi06, 2024, Rio de Janeiro, Brazil (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3658933>
- He, W. (2023). BERT-BiLSTM-CRF Chinese Resume Named Entity Recognition Combining Attention Mechanisms. In *The 4th International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2023)*, November 1719, 2023, Dalian, China (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3652628.3652719>
- Heakl, A., Mohamed, Y., Mohamed, N., Elsharkawy, A., & Zaky, A. (2024). ResuméAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models. *Procedia Computer Science*, 244, 158–165. <https://doi.org/10.1016/j.procs.2024.10.189>
- Kaya, M. (2023). An Exploration of Sentence-Pair Classification for Algorithmic Recruiting. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23)*, September 18â•fi22, 2023, Singapore, Singapore (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3604915.3610657>
- Kumar, D., Verma, C., & Illes, Z. (2026). Optimizing student job placements with NLP and Explainable AI: A fair and transparent hiring framework. *Array*, 29(February), 100729. <https://doi.org/10.1016/j.array.2026.100729>
- Lacroux, A., & Martin-Lacroux, C. (2022). Should I Trust the Artificial Intelligence to Recruit? Recruiters' Perceptions and Behavior When Faced With Algorithm-Based Recommendation Systems During Resume Screening. *Frontiers in Psychology*, 13(July). <https://doi.org/10.3389/fpsyg.2022.895997>
- Lępicki, M., Latkowski, T., Antoniuk, I., Bukowski, M., Świdorski, B., Baranik, G., Nowak, B., Zakowicz, R., Dobrakowski, L., Act, B., & Kurek, J. (2025). Comparative Evaluation of Sequential Neural Network (GRU, LSTM, Transformer) Within Siamese Networks for Enhanced Job–Candidate Matching in Applied Recruitment Systems. *Applied Sciences (Switzerland)*, 15(11). <https://doi.org/10.3390/app15115988>
- Ling, B., Dong, B., & Cai, F. (2025). Applicants' Fairness Perception of Human and AI Collaboration in Resume Screening. *International Journal of Human-Computer Interaction*, 41(17), 10787–10798. <https://doi.org/10.1080/10447318.2024.2437235>
- Liu, J., Shen, Y., Zhang, Y., & krishnamoorthy, S. (2021). Resume Parsing based on Multi-label Classification using Neural Network models. *Proceedings of the 6th International Conference on Big Data and Computing*, 177–185. <https://doi.org/10.1145/3469968.3469998>
- Liu, X. (2025). Deep Learning-Based Intelligent Resume-Position Matching System : Semantic Understanding and Recommendation of BERT Model in Massive Recruitment Data. In *2025 International Symposium on Machine Learning and Social Computing (MLSC 2025)*, October 1618, 2025, Hongkong, China (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3778450.3778452>
- Luo, J. (2022). A Novel Chinese Resume Named Entity Recognition Model Based on Lexical Enhancement. In *2022 11th International Conference on Computing and Pattern Recognition (ICCPR) (ICCPR 2022)*, November 17â•fi19, 2022, Beijing, China (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3581807.3581856>
- Maree, M. (2025). Illuminating Summarization Efficacy: A Qualitative Analysis of Resume Condensation for Job Category Classification. *International Journal of Advances in Soft Computing and Its Applications*, 17(3), 275–287. <https://doi.org/10.15849/IJASCA.251130.16>
- Mishra, S., Mallick, P. K., Tripathy, H. K., Jena, L., & Chae, G. S. (2021). Stacked KNN with hard voting predictive approach to assist hiring process in IT organizations. *International Journal of Electrical Engineering Education*. <https://doi.org/10.1177/0020720921989015>
- Nadira, F. (2026, May 29). Cari Kerja Kian Mustahil, Satu Lowongan Diserbu Ribuan Orang. *CNBC Indonesia*. <https://www.cnbcindonesia.com/lifestyle/20260525164314-33-737899/cari-kerja-kian-mustahil-satu-lowongan-diserbu-ribuan-orang>
- Noble, S. M., Foster, L. L., & Craig, S. B. (2021). The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment*, 29(2), 139–153. <https://doi.org/10.1111/ijsa.12320>
- Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., & Zschech, P. (2025). CareerBERT: Matching resumes to ESCO jobs in a shared embedding space for generic job recommendations. *Expert Systems with Applications*, 275(October 2024), 127043. <https://doi.org/10.1016/j.eswa.2025.127043>
- Surendiran, B., Paturu, T., Chirumamilla, H. V., & Reddy, M. N. R. (2023). Resume Classification Using ML Techniques. *Proceedings of 2023 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication, IConSCEPT 2023*, 1–5. <https://doi.org/10.1109/IConSCEPT57958.2023.10169907>
- Tyagi, S. (2024). Promoting Gender Fair Resume Screening Using Gender-Weighted Sampling. In *International Conference on Computing, Machine Learning and Data Science (CMLDS 2024)*, April 12â•fi14, 2024,

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

- Singapore, Singapore* (Vol. 1, Issue 1). Association for Computing Machinery. <https://doi.org/10.1145/3661725.3661786>
- Tyagi, S., Anuj, Qian, W., Xie, J., & Andrews, R. (2024). Enhancing gender equity in resume job matching via debiasing-assisted deep generative model and gender-weighted sampling. *International Journal of Information Management Data Insights*, 4(2), 100283. <https://doi.org/10.1016/j.jjime.2024.100283>
- Ujlayan, A., Bhattacharya, S., & Sonakshi. (2023). A Machine Learning-Based AI Framework to Optimize the Recruitment Screening Process. *International Journal of Global Business and Competitiveness*, 18(1), 38–53. <https://doi.org/10.1007/s42943-023-00086-y>
- Wang, S., & Zhao, J. (2025). Artificial Intelligence-driven Enterprise Human Resource Recruitment Optimisation System Design. *Proceedings of the 9th International Conference on Electronic Information Technology and Computer Engineering*, 261–266. <https://doi.org/10.1145/3766671.3766717>

