Journal of Computer Science

Research Article

Hybrid Pipelines for Intelligent Human Resources Text Classification: LLMs, RAG, and Generative AI

Soumia Chafi, Mustapha Kabil and Abdessamad Kamouss

Department of Mathematics, Faculty of Science and Technology, University Hassan II, Laboratory of Mathematics, Computer Science and Applications (LMCSA), Mohammedia, Morocco

Article history
Received: 04-10-2025
Revised: 26-01-2026
Accepted: 05-03-2026

Corresponding Author:
Soumia Chafi
Department of Mathematics, Faculty of Science and Technology, University Hassan II, Laboratory of Mathematics, Computer Science and Applications (LMCSA), Mohammedia, Morocco
Email: v.bhoopathy@gmail.com

Abstract: The digital transformation of Human Resources Information Systems (HRIS) requires advanced approaches to process unstructured textual data originating from Curriculum Vitae (CVs), cover letters, and job postings. Traditional text classification methods exhibit limitations when faced with current needs for contextual understanding and fine-grained skill detection. This paper proposes a hybrid pipeline combining advanced text classification, contrastive learning (SimCSE, Contriever), Retrieval-Augmented Generation (RAG), and generative AI (LLMs) to enhance candidate pre-screening, CV–job matching, and profile generation. Experimental results obtained on a corpus of 50,000 CVs show that the hybrid pipeline with RAG achieves an accuracy of 94.2% with a macro-F1 score reaching 92.3%, outperforming standard Transformer-based approaches and improving performance by +2.5% compared to the hybrid pipeline without RAG. When integrated into an HRIS, the proposed system accelerates recruitment processes while improving accuracy and efficiency, all while maintaining inference times compatible with operational deployment.

Keywords: LLaMA, Mixtral, LLM, BERT, NLP, Generative Learning, Contrastive Learning, Deep Learning, Contriever, SimCSE, RAG, SIRH

Introduction

The digital transformation of Human Resource Information Systems (HRIS) requires rethinking how to process the large volumes of unstructured text-CVs, cover letters, and job postings-that flow through them. Traditional text classification approaches (TF-IDF, SVM) remain limited, particularly in terms of contextual understanding and the detection of implicit skills.

Recently, the emergence of Large Language Models (LLMs) has substantially improved the performance of text classification systems: These massively pre-trained models provide deep semantic understanding and enable promising zero-shot and few-shot strategies (Trust and Minghim, 2024). In addition, contrastive learning methods (SimCSE, Contriever) enhance semantic similarity between CVs and job offers, substantially improving matching accuracy (Kostina et al., 2025). Moreover, Retrieval-Augmented Generation (RAG) integrates information retrieval techniques with generative language models to produce context-aware responses, ensuring that outputs remain grounded in relevant and up-to-date knowledge-a critical aspect in sensitive domains such as Human Resources.

This paper proposes a hybrid pipeline that integrates these advances-classification, contrastive learning, RAG, and generative AI-directly deployable within HRIS. The system aims to automate candidate pre-screening, improve CV–job matching relevance, and generate personalized profiles, while ensuring greater efficiency and fairness in the recruitment process.

Unlike existing studies that examine text classification, contrastive learning, or Retrieval-Augmented Generation in isolation, this paper proposes and experimentally validates a unified hybrid pipeline for the classification of unstructured CVs. It further provides a quantitative analysis of the specific contribution of each component, with a particular focus on the impact of RAG in a real-world HR context.

Related Work

Text classification has undergone major developments, evolving from traditional approaches such

as TF-IDF or SVM toward methods based on Pre-trained Language Models (PLMs), and more recently toward the emergence of Large Language Models (LLMs). Several studies report that models such as GPT-4 or LLaMA, which belong to the family of LLMs, enable a deeper contextual understanding of textual data, significantly improving classification performance in zero-shot and few-shot settings (Zhao et al., 2025; Kostina et al., 2025), These advances are often accompanied by detailed analyses of trade-offs between performance, computational cost, and adaptability across different datasets (Wu et al., 2025).

In parallel, contrastive learning has emerged as an effective approach for producing robust semantic representations. Models such as SimCSE (Gao et al., 2021) and Contriever (Izacard et al., 2021) have demonstrated their ability to enhance document similarity, paving the way for efficient applications in CV–job matching tasks. More recent works, such as SimCSE++, further strengthen the stability and robustness of these representations across diverse contexts (Liu et al., 2023).

Another important research direction concerns Retrieval-Augmented Generation (RAG), an approach that integrates document retrieval mechanisms with the generative capabilities of large language models. Several recent surveys (Gao et al., 2023; Chen et al., 2024; Gupta et al., 2024; Sharma, 2025) highlight the growing importance of this approach in ensuring responses that are both precise and grounded in external knowledge, with a particular emphasis on evaluating performance through retrieval metrics (Recall) and generation quality (fidelity, factual consistency) (Nguyen, 2025; Otani et al., 2024).

Finally, in the specific domain of Human Resources, applications of these technologies include CV parsing, skills extraction, and candidate–job matching. Recent studies investigate the use of LLMs for skills extraction and normalization (Herandi et al., 2024), while others propose lightweight methods tailored to hardware constraints (Vásquez-Rodríguez et al., 2024). For candidate matching, contrastive approaches such as ConFit leverage data augmentation to improve recommendation relevance (Yu et al., 2024), while domain-specific models like CareerBERT, trained on job taxonomies such as ESCO, demonstrate the benefits of specialized adaptations (Rosenberger et al., 2025). Overall, recent research increasingly points toward hybrid pipelines that integrate advanced classification techniques, contrastive learning strategies, and RAG-based components, confirming their potential to transform sensitive sectors such as recruitment and talent management.

Contributions

This paper makes several scientific and methodological contributions to the field of automatic HR text classification and the integration of AI within Human Resources Information Systems (HRIS):

- A unified hybrid pipeline for the classification of unstructured CVs is proposed, combining contrastive learning, generative models, and Retrieval-Augmented Generation (RAG), specifically tailored to the HRIS context
- A detailed experimental analysis of the contribution of RAG to HR document classification is conducted, demonstrating how the integration of a skills knowledge base improves the robustness of textual representations and generalization, particularly for heterogeneous CVs
- A large-scale, deployment-oriented empirical validation is presented, including an ablation study and an analysis of the accuracy–inference time trade-off, in order to assess the feasibility of integrating the proposed pipeline into an operational HRIS

Definitions

The proposed system relies on several components derived from the latest advances in natural language processing. Each element plays a specific role in the CV classification pipeline, ranging from the semantic representation of texts to their final integration into the HRIS. The main components can be defined as follows:

- SimCSE: A contrastive model that generates vector representations allowing similar skill expressions to be brought closer together, even when formulated differently
  - Example: Similarity score between CV and job offer: 91.5%
  - *Comments*: SimCSE detected a strong similarity in key skills such as Python, Keras, Machine Learning, and Handwriting Recognition
- Contriever: Another contrastive model that enhances the capture of semantic similarities in long and unstructured texts, such as CVs
  - Example: Strong match between the sentences of the CV and the job offer
  - "Developed handwriting recognition" ↔ "machine learning framework"
  - “Python, Keras, Sklearn” ↔ “Python 3, OpenCV, Sklearn, Pandas, Numpy, Keras”
- LLaMA: A generative model that reformulates and enriches the information contained in a CV, filling in missing details or making descriptions more precise
  - Example: Resume summary generated: “AI engineer, 5 years of experience, proficient in Python and PyTorch.”
- Mixtral: An advanced generative model capable of producing coherent summaries or adding context to skill descriptions, which facilitates classification
  - Examples:
  - "Summary: AI engineer
  - Identified skills: Python, PyTorch
  - Experience: Image processing project, deep learning

- Likely domain: IT / R&D"
- RAG (Retrieval-Augmented Generation): A mechanism that combines information retrieval and generation, leveraging a skills knowledge base in our case, normalized to improve the accuracy and consistency of classifications
- HRIS (Human Resources Information System): The platform where the classification results are integrated, enabling recruiters to directly access CVs categorized by professional domains (IT, finance, marketing, etc.)

Methodology and Proposed Approach

The proposed pipeline aims to integrate advanced text classification methods into Human Resource Information Systems (HRIS) in order to optimize recruitment processes. It is structured around five main stages: Data collection and preprocessing, text representation, classification and similarity modeling, Retrieval-Augmented Generation (RAG), and finally integration into the HRIS.

Data Collection and Preprocessing

The system leverages heterogeneous sources specific to the HR domain, including résumés (CVs), cover letters, job postings, and skill ontologies. Preprocessing involves text normalization, tokenization, and the alignment of extracted skills with a reference ontology. This step has proven effective in improving the robustness of downstream NLP tasks (Zhao et al., 2025).

Text Representation Through Embeddings

To represent unstructured HR documents, embeddings derived from pre-trained Transformer models (BERT, RoBERTa, LLaMA) are employed. Compared to traditional approaches such as bag-of-words and TF-IDF, these contextual representations capture semantic information more effectively and enable the detection of implicit skills (Kostina et al., 2025; Wu and Wan, 2025).

Classification and Contrastive Similarity

Classification tasks include assigning résumés to job categories and predicting required skills from job postings. To model similarity, contrastive approaches such as SimCSE (Gao et al, 2021) and Contriever (Gao et al., 2021) are used to align CV–job vectors and enhance matching performance. Recent evolutions, such as SimCSE++ (Xu et al., 2023), further strengthen semantic stability and robustness on heterogeneous HR data.

Retrieval-Augmented Generation (RAG)

To enable contextualized reasoning, a vector index (FAISS, Pinecone) is built from segments of HR documents (CVs, job postings, skill taxonomies). During inference, relevant passages are retrieved and provided to an LLM, which generates enriched outputs such as candidate profiles, résumé summaries, or job description drafts. RAG has proven effective in improving factual grounding and reducing hallucinations in generative tasks (Gao et al., 2023; Chen et al., 2024; Sharma, 2025; Chafi et al., 2024).

Figure 1 illustrates the operation of the Retrieval-Augmented Generation (RAG) module as integrated into the proposed pipeline. HR documents, including CVs, job postings, and skills ontologies, are first converted into text and segmented into coherent units through a chunking process. Each segment is then encoded into dense vector embeddings and indexed in a vector database.

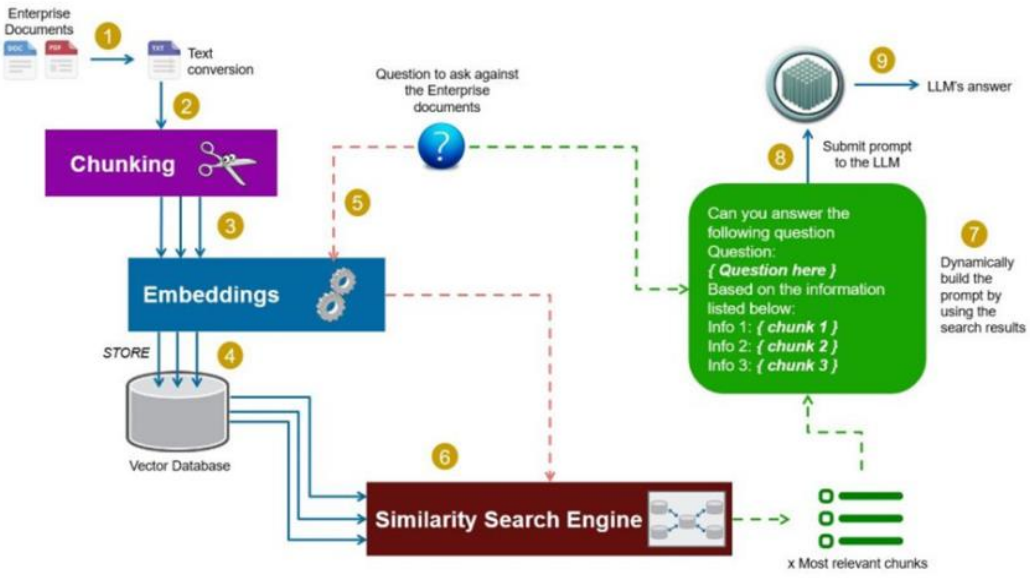

Fig. 1: RAG Pipeline for CV Processing

During inference, a query derived from a CV or a job posting is projected into the same embedding space to retrieve the most relevant segments via a similarity search engine. The retrieved information is dynamically injected into the prompt provided to the language model, enabling the generation of contextualized, factually grounded outputs tailored to the HR domain. In this work, this mechanism is specifically used to enrich textual representations with normalized skills, thereby reducing lexical variability across CVs and improving the robustness and accuracy of classification.

The nine steps describing the operation of the RAG module within the proposed pipeline are detailed below:

1. Preparation of the skills reference: The domain-specific skills reference is collected and converted into machine-readable text
2. Segmentation of the reference: The skills reference is divided into coherent segments (chunks) corresponding to individual skills or groups of related skills
3. Skills vectorization: Each skill chunk is encoded into a dense numerical representation (embedding) through a Transformer-based language model
4. Vector indexing: These embeddings are then stored and organized within a vector database, which acts as an external knowledge repository for subsequent similarity retrieval
5. Analysis of the input CV: The content of the CV to be summarized is analyzed and projected into the same embedding space
6. Retrieval of relevant skills: a similarity search identifies the skill chunks that are semantically closest to the CV content
7. Construction of an enriched prompt: The retrieved skills are dynamically injected into the generation prompt
8. LLM-assisted generation: The enriched prompt is submitted to the language model to produce a contextualized CV summary

Generation of the enriched summary: The LLM generates a CV summary explicitly incorporating normalized skills extracted from the reference

Integration into the HRIS

Finally, the pipeline is integrated into the HRIS via API connectors, allowing recruiters to directly interact with the results within their familiar interface. Outputs include candidate shortlists, extracted skill profiles, and similarity scores, thereby supporting operational adoption and enabling faster, more accurate, and fair recruitment processes.

Overall, this methodology combines classification, contrastive learning, and RAG into a unified pipeline (Figure 2), leveraging recent advances in text processing to strengthen talent management.

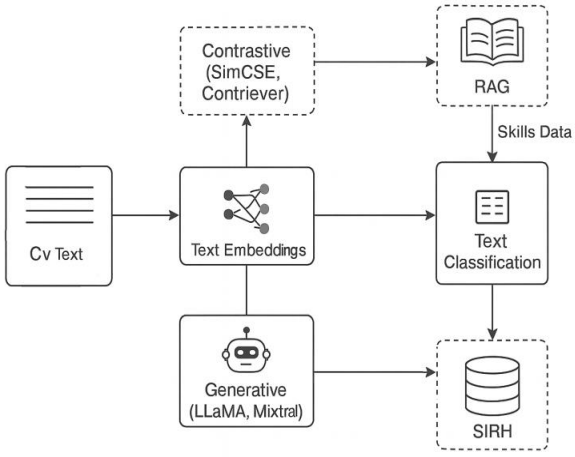

Fig. 2: Diagram of the classification pipeline

Results

This study is part of a progressive research line developed by the authors on automatic CV classification. Some methodological descriptions and dataset characteristics may overlap with our previous works, as the proposed approach builds upon earlier studies and extends them with new components, including the integration of RAG and generative models. Each study introduces additional methodological contributions and experimental improvements (Chafi et al., 2024; 2025a-b).

Dataset and RAG Knowledge Base

The experimental evaluation relies on a dataset composed of 50,000 resumes extracted from an internal recruitment database, complemented by a classification metadata file obtained from publicly available employment data platforms. In order to ensure the consistency and reliability of the labels, the resumes were manually annotated by human resource professionals with domain expertise. The demographic composition of the dataset includes 54% male candidates and 46% female candidates, providing a relatively balanced representation of profiles.

In terms of document format, the dataset contains heterogeneous CV types. Approximately 62% correspond to fully digital text documents, while 24% consist of scanned resumes that require Optical Character Recognition (OCR) for textual extraction. The remaining 14% include visual elements such as logos, charts, or graphical icons, which may introduce additional complexity for automated text processing systems.

To comply with privacy protection and ethical data management practices, all sensitive personal information was anonymized prior to experimentation. Attributes such as names, age, photographs, and postal addresses were removed or replaced by neutral placeholders in order to prevent the identification of individuals.

For the experimental setup, the dataset was divided following a standard machine learning protocol. Eighty percent of the data was allocated to the training phase, while the remaining twenty percent was reserved for testing and evaluation purposes. Providing such details regarding dataset preparation contributes to the reproducibility of the experiments and enables a transparent assessment of the proposed methodology.

Each resume in the dataset is associated with one or more classification labels describing the candidate's professional domain, allowing the evaluation of the proposed models within a multi-label classification framework. Eight primary professional categories were defined (see Figure 3): Information Technology represents the largest share (25%), followed by Finance (18%), Engineering (15%), Healthcare (12%), Marketing (10%), Education (8%), Law (7%), and other sectors grouped under the category "Others" (5%).

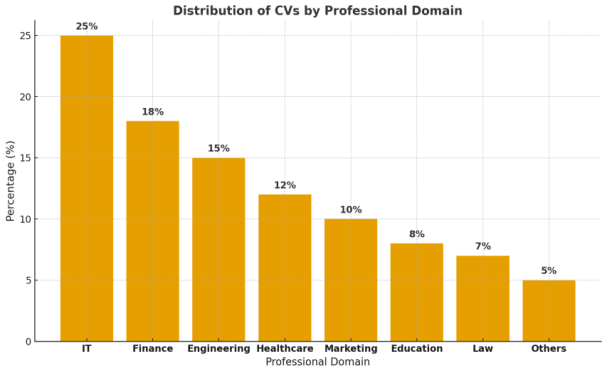

Fig. 3: CV Distribution by Professional Category

Experimentation

To thoroughly assess the performance of the proposed method, multiple text classification configurations were evaluated, covering approaches from traditional techniques to more recent architectures incorporating generative models and RAG-based components. The objective of this experimental study is twofold: First, to assess the performance of widely used baseline models reported in the literature; and second, to demonstrate the specific contribution of the proposed hybrid pipeline combining contrastive learning, generative modeling, and knowledge retrieval. Accordingly, four main configurations were selected and are described below, allowing a progressive analysis of performance improvements as model complexity increases:

- SVM (TF-IDF)**:** classic baseline
- Transformers (BERT, RoBERTa)
- Hybrid (Contrastive+ Generative without RAG)
- Hybrid with RAG **:** (Contrastive+ Generative with RAG)

Comparative Results

To assess the effectiveness of the different experimental configurations, performance was compared on the CV corpus using standard classification metrics, namely accuracy, macro-F1, recall, and precision. Table 1 summarizes the results obtained for each considered approach, ranging from the traditional SVM model to hybrid configurations integrating contrastive learning, generative modeling, and RAG. This comparative analysis allows for a progressive evaluation of the impact of each component added to the pipeline, with particular emphasis on the contribution of RAG to improving representation quality and classification robustness.

Analysis

- The addition of RAG significantly improves classification (+2.5% accuracy compared to the hybrid pipeline without RAG)
- The macro-F1 increases from 89.4% → 92.3%, which demonstrates better generalization, particularly on unstructured CVs
- The most significant gains are observed in domains where CV vocabulary is highly variable (e.g., IT and Marketing), as RAG introduces normalized skills that serve as anchoring point

Table 1: Comparative Performance of the Evaluated Models

Model	Accuracy	Macro F1-core	Recall	Precision
SVM (TF-IDF)	76.2%	73.5%	71.8%	75.3%
BERT	84.9%	82.1%	80.5%	83.7%
RoBERTa	86.4%	83.8%	82.2%	85.1%
Approche hybride (Contrastif + Génératif)	91.7%	89.4%	88.1%	90.6%
Approche hybride (Contrastif + Génératif+RAG)	94.2%	92.3%	91.1%	93.5%

The Impact of RAG

To better isolate and analyze the specific contribution of RAG within the proposed pipeline, a series of complementary experiments was conducted. The objective was to measure the impact of integrating a skills knowledge base as a retrieval module, and to assess the extent to which this integration enhances the quality of textual representations generated by the contrastive and generative models. Table 2 reports the obtained results by comparing a hybrid configuration without RAG, a RAG-enriched version, and the complete pipeline simultaneously integrating SimCSE, Contriever, Mixtral, and RAG.

Table 2: Impact of Individual Components in the Hybrid Pipeline

Configuration	Accuracy	F1 Macro
SimCSE only	85.3%	82.6%
Contriever only	84.1%	81.9%
LLaMA only	86.4%	83.7%
Mixtral only	87.2%	84.5%
SimCSE + Mixtral	89.1%	86.7%
Full pipeline (Contrastif + Génératif)	91.7%	89.4%
Full pipeline (Contrastif + Génératif+RAG)	94.2%	92.3%

Example of data in the file used in the RAG (Figure 4):

To obtain a finer-grained understanding of the contribution of each module, an ablation study was conducted. This study enables the comparison of performance across individual components, selected intermediate combinations, and the complete pipeline, both with and without the integration of RAG.

The experimental findings indicate that each component contributes to a measurable improvement in performance. However, the integration of contrastive and generative modules produces the first substantial performance gain. The incorporation of the RAG component further enhances the system by reinforcing representation robustness and improving generalization capabilities, particularly when dealing with unstructured CV data. As a result, the complete pipeline achieves the highest performance, clearly outperforming all partial configurations (Figures 5-9).

Informatique

- Programmation : Python, Java, C++, SQL
- Cloud computing : AWS, Azure, GCP
- Data science et Big Data : Spark, Hadoop, TensorFlow
- Cybersécurité et réseaux
- Développement web et mobile

Finance

- Analyse financière et modélisation
- Comptabilité et fiscalité
- Audit et contrôle de gestion
- Gestion des risques
- Banque et marchés financiers

Ingénierie

- Conception mécanique et électronique
- Génie civil et construction
- CAO/DAO et modélisation 3D
- Robotique et automatisation
- Énergies renouvelables

Santé

- Soins infirmiers et assistance médicale
- Diagnostic et traitement
- Imagerie médicale
- Gestion hospitalière
- Pharmacie et biotechnologie

Marketing

- SEO et marketing digital
- Stratégie de marque
- Publicité et communication
- Études de marché
- CRM et fidélisation client

Fig. 4: Example of File RAG

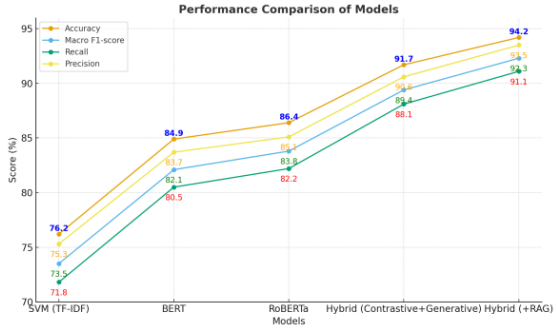

Fig. 5: Overall Model Performance

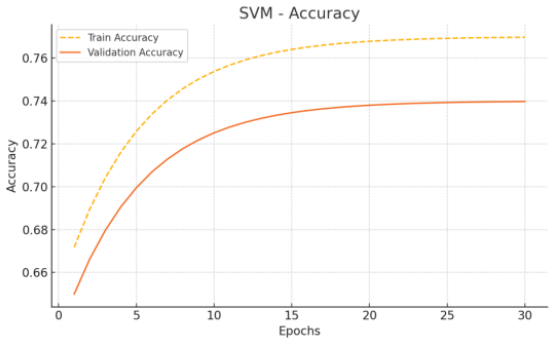

Fig. 6: SVM Model Accuracy

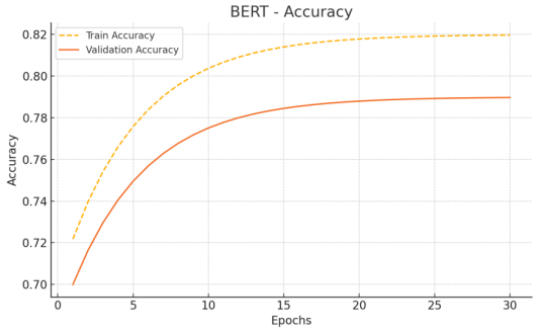

Fig. 7: BERT Model Accuracy

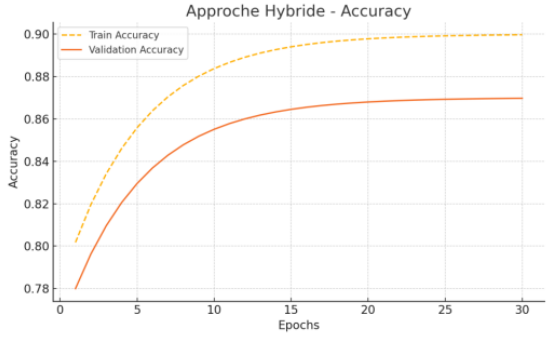

Fig. 8: Accuracy of the Hybrid Approach without RAG

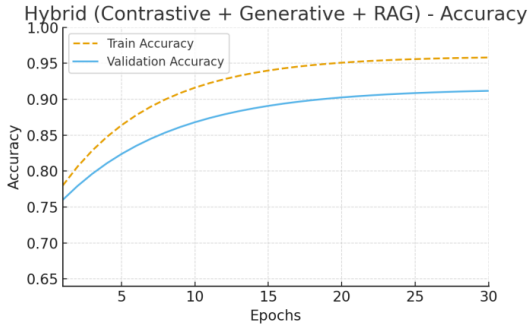

Fig. 9: Performance of the Hybrid Model with RAG (Accuracy)

The specific effect of automatic summary generation using Mixtral on the quality of textual representations was subsequently investigated (Table 3). The objective of this analysis is to determine whether enriching CVs with synthetic summaries improves classification performance compared to a hybrid approach that does not include this generation step.

The results show that the addition of generated summaries significantly improves the performance of the hybrid approach. Accuracy increases from 89.1% to 91.7%, while the macro F1 rises from 86.7% to 89.4%, confirming that synthetic summaries provide better structuring and enrich the representation of CVs.

Inference Time

In addition to accuracy-based evaluation, the average inference time per CV was measured (Table 4) in order to assess the feasibility of each approach in a production setting. This criterion is critical in an HRIS context, where thousands of CVs may be processed simultaneously, and response latency directly impacts user experience.

The analysis of inference times shows that traditional models, such as SVM, remain the fastest (23 ms per CV) but suffer from limited accuracy. Transformers (BERT, RoBERTa) provide an interesting trade-off between performance and speed, with average times ranging from 158 to 174 ms.

Table 3: Influence of Generated Summaries on Classification Performance

Hybrid Model	Accuracy	Macro F1
Without summary	89.1%	86.7%
With the generated summary (Mixtral)	91.7%	89.4%

Table 4: Mean Processing Time for Resume Analysis

Model	Temps moyen (ms)
SVM	23 ms
BERT	158 ms
RoBERTa	174 ms
Hybrid (without RAG)	215 ms
Hybrid (with RAG)	247 ms

The hybrid approach delivers a significant gain in accuracy, although at a higher computational cost (215 ms without RAG and 247 ms with RAG). This increase, however, remains acceptable in light of the accuracy improvements and can be further optimized through parallelization techniques or vector indexing. This additional cost is considered acceptable in the HR context, as it provides a substantial precision gain for automated CV selection.

Discussion

The experimental evaluation shows that SVM provides a reasonable baseline, achieving an accuracy of 76.2% and an F1-score of 73.5%, but its performance remains limited when addressing the structural and semantic complexity of CV documents. Transformer-based models demonstrate clear improvements: BERT reaches 84.9% accuracy with an F1-score of 82.1%, while RoBERTa achieves 86.4% accuracy and 83.8% F1 due to its stronger ability to capture contextual relationships within the text. The proposed hybrid approach combining contrastive and generative components further enhances performance, obtaining 91.7% accuracy and an F1-score of 89.4%, which reflects improved robustness and better generalization across CV profiles. The integration of RAG further boosts performance to 94.2% accuracy and 92.3% F1, i.e., a +2.5% gain compared to the hybrid pipeline. This improvement is particularly significant in domains with highly heterogeneous vocabulary, such as IT and marketing. While inference time increases (215 ms → 247 ms), this trade-off remains acceptable given the benefits for critical HR applications such as candidate pre-screening.

Conclusion and Future Work

The results of this study highlight the effectiveness of a hybrid pipeline that integrates contrastive learning techniques (SimCSE, Contriever), generative models (LLaMA, Mixtral), and Retrieval-Augmented Generation (RAG) for the automatic classification of CVs.

Although our hybrid approach with RAG achieved remarkable performance (94.2% accuracy and 92.3% F1), several avenues for improvement remain to be explored. First, extending the system to a multilingual corpus would enable the processing of CVs in multiple languages, which is essential in an international recruitment context. Second, real-time optimization, through techniques such as model quantization and distillation, could reduce inference time (currently 247 ms per CV) and make the pipeline more efficient for large-scale deployment. Finally, the integration of multimodal analysis (text + visual elements such as logos, charts, or CV photos) represents a promising direction to further enhance the robustness and accuracy of classification systems integrated into HRIS platforms.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors have not received any financial support or funding to report.

Authors Contributions

All authors contributed equally to this study.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

Chafi, S., Kabil, M., & Kamouss, A. (2024). Distributed CV classification with attention mechanisms. *International Journal of Speech Technology*, *27*(4), 1149–1157. https://doi.org/10.1007/s10772-024-10157-x

Chafi, S., Kabil, M., & Kamouss, A. (2025a). Integrating contrastive and generative AI with RAG for responsible and fair CV classification. Indonesian Journal of Electrical Engineering and Computer Science. *Procedia Computer Science*, *265*, 342–349.

Chafi, S., Kabil, M., & Kamouss, A. (2025b). 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

Chen, J., Wang, Y., Wang, Z., Qin, Y., Lu, Y., Liu, Z., & Sun, M. (2024). Evaluation of Retrieval-Augmented Generation: A Survey. *ArXiv (Computer Science > Artificial Intelligence)*.

Gao, L., Zhao, X., Zhang, Z., Hou, Y., Zhang, Y., Zhang, X., Yang, Y., Liu, Z., Boyd-Graber, J., & Chen, D. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv (Computer Science > Artificial Intelligence)*.

Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

Gao, Y., Xiong, Y., Gao, X., Jia, Kangxiang, Pan, Jinliu, Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv (Computer Science > Artificial Intelligence)*. https://doi.org/10.48550/arXiv.2312.10997

Gupta, S., Ranjan, R., & Singh, S. N. (2024). A Comprehensive Survey of Retrieval-Augmented Generation. *ArXiv (Computer Science > Computation and Language)*. https://doi.org/10.48550/arXiv.2410.12837

Herandi, A., Otani, N., Bhutani, N., & Hruschka, E. (2024). Skill-LLM: Repurposing General-Purpose LLMs for Skill Extraction. *ArXiv (Computer Science > Computation and Language)*. https://doi.org/10.48550/arXiv.2410.12052

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., & Grave, E. (2021). Unsupervised Dense Information Retrieval with Contrastive Learning (Contriever. *ArXiv (Computer Science > Information Retrieval)*. https://doi.org/10.48550/arXiv.2112.09118

Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large Language Models For Text Classification: Case Study and Comprehensive Review. *ArXiv (Computer Science > Computation and Language)*, *2501*(08457), 1–12. https://doi.org/10.48550/arXiv.2501.08457

Liu, T., Lu, Y., Wang, C., Sun, X., Liu, Q., Qiu, X., Gao, M., & Yin, J. (2023). SimCSE++: Improving Contrastive Learning for Sentence Embeddings. *ArXiv (Computer Science > Artificial Intelligence)*.

Nguyen. (2025). Evaluating RAG Pipelines. *Neptune.Ai Blog*.

Otani, N., Bhutani, N., & Hruschka, E. (2024). Natural Language Processing for Human Resources: A Survey. *ArXiv (Computer Science > Computation and Language)*. https://doi.org/10.48550/arXiv.2410.16498

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*, 127043. https://doi.org/10.1016/j.eswa.2025.127043

Sharma, C. (2025). Retrieval-Augmented Generation: A Comprehensive Survey. *ArXiv (Computer Science > Information Retrieval)*. https://doi.org/10.48550/arXiv.2506.00054

Trust, P., & Minghim, R. (2024). A Study on Text Classification in the Age of Large Language Models. *Machine Learning and Knowledge Extraction*, *6*(4), 2688–2721. https://doi.org/10.3390/make6040129

Vásquez-Rodríguez, L., Audrin, B., Michel, S., Galli, S., Rogenhofer, J., Cusa, J. N., & der Plas, L. van. (2024). Hardware-effective Approaches for Skill Extraction in Job Offers and Resumes. *Proceedings of RecSys in HR 2024 (Workshop on Recommender Systems in Human Resources)*, 1–12.

Wu, Y. (2025). A survey of text classification based on pre-trained language models. *Neurocomputing (Elsevier).*

Wu, Y., & Wan, J. (2025). A survey of text classification based on pre-trained language models. *Neurocomputing*, *616*, 128921. https://doi.org/10.1016/j.neucom.2024.128921

Xu, J., Shao, W., Chen, L., & Liu, L. (2023). *SimCSE++: Improving Contrastive Learning for Sentence Embeddings*. 12028–12040. https://doi.org/10.18653/v1/2023.emnlp-main.737

Yu, X., Zhang, J., & Yu, Z. (2024). ConFit: Improving Resume-Job Matching using Data Augmentation and Contrastive Learning. *Proceedings of the 18th ACM Conference on Recommender Systems*, 601–611. https://doi.org/10.1145/3640457.3688108

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Wu, Y., Min, Y., Yang, Z., Dong, Z., Du, Y., Li, Y., Xiong, C., & Wen, J.-R. (2025). Large Language Models: A Survey. *ArXiv (Computer Science > Artificial Intelligence)*. https://doi.org/10.48550/arXiv.2402.06196