Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.4, August 2026, pp. 3733~3739
e-ISSN: 2548-6861

Automatic Classification and Semantic Retrieval for Low-Resource Parliamentary Archives: A Case Study of the DRC Senate

Otshudiakoy Jean[1*], Djungu Ahuka Saint Jean[1,2]
[1] Department of Mathematics, Statistics and Computer Science, Faculty of Science and Technology, University of Kinshasa, Kinshasa, Democratic Republic of the Congo
[2] CRIA-Center for Research in Applied Computing, Kinshasa, Democratic Republic of the Congo; Department of Mathematics, Computer Science and Statistics, University of Kinshasa
otshudiakoy2011@gmail.com, john_otshudi@senat.cd, saintjean.djungu@unikin.ac.cd

Article Info

Article history:

Received 2026-06-14
Revised 2026-07-02
Accepted 2026-08-11

Keyword:

Automatic Classification, Parliamentary Archives, OCR, TF-IDF, CamemBERT, Probabilistic Fusion, Semantic Search, Low-Resource NLP.

ABSTRACT

Parliamentary archives preserve the administrative, legal and political memory of legislative institutions. At the Senate of the Democratic Republic of the Congo, many documents are scanned PDF files, which makes manual indexing, classification and retrieval slow. This study does not propose a new algorithm. Its contribution is the applied integration of automatic classification and hybrid semantic retrieval in a French, OCR-based and low-resource parliamentary archive. The corpus contains 420 documents classified into five categories: orders, decrees, laws, ordinances and minutes. The models include TF-IDF with linear SVM, logistic regression and Naive Bayes, CamemBERT embeddings with logistic regression, and a weighted probabilistic fusion: P_fusion = alpha P_TFIDF + (1-alpha) P_CamemBERT, with alpha = 0.55. On the fixed test set, the hybrid model obtains accuracy = 0.786 and macro F1 = 0.815. Under stratified five-fold cross-validation, TF-IDF + linear SVM is the most stable lexical baseline, with accuracy = 0.821 +/- 0.052 and macro F1 = 0.834 +/- 0.050. These protocols are interpreted separately. For thirteen curated archival queries, hybrid retrieval performs best at k = 10, while TF-IDF obtains the highest MRR. The results show that lexical models are strong for categories with stable legal markers, whereas hybrid retrieval is useful for broader semantic access. The proposed system is a human-validated archival workflow, not a fully autonomous decision tool.

I. INTRODUCTION

Parliamentary archives preserve the institutional memory of legislative work. They contain documents that record legal decisions, administrative procedures, political debate and internal governance. In the context of the DRC Senate, this memory is not only historical; it is also operational, because staff members need to retrieve documents quickly for legislative, administrative and public information purposes [1], [2], [3].

Automatic document classification is the supervised assignment of a text to a predefined category using statistical or machine-learning models. In this study, the concept is applied to the identification of parliamentary document types, such as orders, decrees, laws, ordinances and minutes, after text extraction by OCR. This contextualization is important because errors in classification can affect subsequent indexing, retrieval and institutional decision support [4], [5].

Semantic search is an information retrieval approach that ranks documents according to their meaning rather than only exact keyword overlap. In parliamentary archives, semantic search is useful because different documents can discuss the same legal or administrative issue using different formulations. It therefore complements automatic classification: classification organizes the archive, while semantic search supports access to the content [6], [7].

A low-resource context refers to an environment where annotated data, domain-specific language resources, computational infrastructure, budget and specialized human expertise are limited. In the present case, the limitation concerns the modest size of the labeled corpus, class imbalance, the scarcity of Congolese parliamentary datasets in French, OCR noise and the practical cost of manual annotation by domain experts. This definition matters because it explains why robust classical baselines remain relevant alongside transformer-based approaches [8], [9].

The novelty of this study is therefore not algorithmic. It lies in the applied integration of automatic classification and hybrid semantic retrieval within a real French-speaking parliamentary archive from the DRC Senate. Previous work on legal NLP, low-resource NLP and AI-supported records management has mostly focused on general legal benchmarks, government records or high-resource settings. By contrast, this study evaluates how established lexical, semantic and hybrid approaches behave when the corpus is small, OCR-derived, institution-specific and class-imbalanced. This contribution positions the work within applied parliamentary informatics, where the main challenge is not only prediction accuracy but also practical archive organization, retrieval support, human validation and risk control [1], [3], [4], [10].

Recent dense retrieval and multilingual sentence-embedding models, including Dense Passage Retrieval, Contriever-style unsupervised dense retrieval and weakly supervised text-embedding models, have improved semantic matching by learning vector representations optimized for retrieval rather than simple keyword overlap. These approaches are relevant for parliamentary archives because users may formulate queries that paraphrase the wording of legal documents. However, they often require larger evaluation collections, domain adaptation, retrieval-oriented fine-tuning or computational resources that may not be immediately available in a low-resource institutional setting. For this reason, the present study uses CamemBERT only as a frozen French pretrained embedding extractor and treats modern multilingual dense retrievers as an important comparative baseline for future work rather than as the main object of the current feasibility study [21], [22], [23].

Integrated view of classification and semantic search.

In the proposed archival workflow, classification and semantic search are not independent tasks. Classification first assigns each incoming document to an institutional category, which supports metadata enrichment, filtering and archive organization. Semantic search then ranks documents within the archive according to the meaning of a user query. The classification output can therefore serve as a structural filter or metadata signal for retrieval, while semantic search improves access to the content when users do not know the exact document type or terminology. The complete workflow is: digitization, OCR, text cleaning, automatic pre-classification, human validation, metadata storage, lexical/semantic indexing and query-based retrieval.

II. METHOD

A. Corpus and Classification Task

The dataset contains 420 OCR-processed parliamentary and legal documents. Each document is assigned to one and only one category, making the task a single-label multiclass classification problem with five categories: orders, decrees, laws, ordinances and minutes. The corpus is intentionally not artificially balanced because it reflects the distribution of an actual institutional archive. This choice improves ecological validity but creates methodological constraints: the minority class, minutes, contains only 25 documents, and this can increase the variance of evaluation scores.

The dataset size limits model stability and external generalization. With only 420 documents, the models may be sensitive to the train-test split, and transformer-based embedding approaches may not reach their full potential because they often benefit from larger labeled corpora, domain-specific adaptation or retrieval-oriented fine-tuning. In this study, CamemBERT was not trained or fine-tuned on the parliamentary corpus; it was used only as a frozen pretrained embedding extractor. The results should therefore be interpreted as evidence from a feasibility study on the studied DRC Senate corpus, not as universal evidence for all parliamentary or legal archives. This limitation also affects semantic search because thirteen curated queries cannot fully represent all real retrieval scenarios.

TABLE I.
DISTRIBUTION OF THE PARLIAMENTARY DOCUMENT CORPUS.

Class	Number of documents	Share
Orders	140	33.33%
Ordinances	122	29.05%
Decrees	68	16.19%
Laws	65	15.48%
Minutes	25	5.95%
Total	420	100%

B. OCR and Preprocessing

OCR is the process of converting scanned page images into machine-readable text. In this study, OCR is not a generic preprocessing detail; it is a central institutional constraint because many parliamentary archives are preserved as scanned PDF files. The extracted text was cleaned by removing empty outputs, correcting encoding problems, reducing repeated spaces and normalizing textual artifacts before vectorization.

No aggressive stopword removal was applied. In parliamentary and legal documents, recurrent formulaic expressions may contain useful information for distinguishing document types. This contextual choice is important because words that appear ordinary in general text may carry institutional meaning in legal or administrative documents [2], [11].

OCR quality was considered a source of noise for both classification and retrieval. The study removed empty OCR outputs and obvious duplicate outputs, but it did not perform a full character-level OCR error audit. Consequently, recognition errors, broken words, missing accents and page-layout artifacts may have weakened both TF-IDF term weighting and CamemBERT embeddings. This limitation is particularly important for scanned legal documents, where a single corrupted term may change the apparent category or reduce the match between a query and a relevant document.

C. Stratified Sampling and Cross-Validation

The corpus was divided into training and test subsets using stratified sampling. The retained split was 80% for training and 20% for testing, corresponding approximately to 336 training documents and 84 test documents. Stratification was used to preserve the class proportions in both subsets, which is important because the corpus is imbalanced and minority classes such as minutes would otherwise be poorly represented.

In addition to the hold-out test split, stratified five-fold cross-validation was performed for the lexical models. The hold-out test and cross-validation results are reported for different purposes. The fixed test split provides one independent estimate on 20% of the corpus, whereas cross-validation provides a descriptive estimate of stability across five partitions. These results are therefore not compared as if they came from the same experimental protocol. Formal significance testing with repeated cross-validation remains a future extension.

D. Model Configurations

TF-IDF represents each document by weighting terms according to their frequency in the document and their relative rarity in the corpus. In the parliamentary context, this is useful because categories such as orders, decrees and minutes often contain recurrent institutional markers. The TF-IDF representation was combined with linear classifiers adapted to sparse high-dimensional text data [4], [12].

Logistic Regression was trained as a multinomial linear classifier with regularization and a high iteration limit to ensure convergence. LinearSVC was used as a maximum-margin classifier for sparse textual vectors. Multinomial Naive Bayes was included as a probabilistic baseline with additive smoothing. These configurations provide a fair comparison between discriminative linear models and a classical probabilistic model in a small OCR-based corpus [5], [12].

CamemBERT was selected because the parliamentary documents are mainly written in French and because it is a well-established pretrained French language model. In this study, CamemBERT was not trained, not adapted and not fine-tuned on the parliamentary corpus. It was used only as a frozen pretrained encoder to generate document embeddings. These embeddings were then used as input features for Logistic Regression in the classification experiment and for vector-based ranking in the semantic search experiment. This choice was motivated by the limited corpus size, the risk of overfitting and the objective of evaluating a realistic low-resource pipeline. However, it also limits the comparison with newer multilingual dense retrieval models and modern sentence-embedding architectures, which should be evaluated in future work.

E. Probabilistic Fusion

Probabilistic fusion combines the predicted class probabilities produced by two complementary models. In this study, the fusion combines the lexical TF-IDF probability vector and the CamemBERT probability vector using the rule P_fusion = alpha x P_TFIDF + (1 - alpha) x P_CamemBERT. The coefficient alpha controls the relative contribution of the lexical and semantic models. The strategy is simple, reproducible and suitable for a low-resource setting because it does not require training a new deep architecture.

The selected value was alpha = 0.55. This value gives slightly more weight to TF-IDF while preserving a substantial contribution from CamemBERT. The rationale is that parliamentary documents contain strong lexical markers, such as decree, order, ordinance, law, minutes, article, session and institutional references. TF-IDF captures these markers directly. CamemBERT, by contrast, can add semantic proximity when two documents express similar legal or administrative content with different formulations. The improvement of the hybrid model should therefore be interpreted as a complementarity effect rather than as evidence that the semantic component alone is superior [6], [13].

F. Semantic Search Evaluation

Semantic search was evaluated as the retrieval component of the archival workflow. Thirteen curated queries were associated with target document classes and evaluated using Precision@k, Recall@k, mean reciprocal rank and nDCG@k. The queries were constructed to cover common archive needs: searching for legal acts by type, retrieving administrative decisions, finding documents related to sessions or minutes and looking for documents expressed through broader institutional vocabulary. The small number of queries is a limitation because it may not represent the full diversity of real parliamentary information needs and may introduce selection bias [6], [7].

Three retrieval approaches were compared: lexical TF-IDF search, CamemBERT semantic search and hybrid TF-IDF + CamemBERT search. The hybrid retrieval setting reflects a realistic user scenario: archive users may combine exact institutional keywords with broader conceptual requests. In practice, classification metadata can also be used as a filter before retrieval; for example, a user may first restrict the search to laws or minutes and then use semantic ranking to identify the most relevant documents within that subset. To reduce evaluation bias, the queries were grouped into four needs: document-type search, administrative-action search, session/minutes search and broader conceptual legal search. This design remains limited because thirteen queries cannot cover all real parliamentary retrieval scenarios.

III. RESULTS AND DISCUSSION

A. Comparative Classification Performance

TABLE II.
TEST PERFORMANCE OF THE CLASSIFICATION MODELS.

Rank	Model	Acc.	Macro Prec.	Macro Rec.	Macro F1
1	Hybrid TF-IDF + CamemBERT	0.786	0.860	0.799	0.815
2	TF-IDF + linear SVM	0.774	0.831	0.786	0.802
3	TF-IDF + logistic regression	0.762	0.823	0.774	0.791
4	CamemBERT embeddings + logistic regression	0.643	0.654	0.685	0.660
5	TF-IDF + Naive Bayes	0.548	0.625	0.453	0.433

On the fixed hold-out test set, the best result is obtained by the hybrid TF-IDF + CamemBERT fusion, with accuracy = 0.786 and macro F1-score = 0.815. TF-IDF + linear SVM remains the strongest isolated lexical classifier, with accuracy = 0.774 and macro F1-score = 0.802. The difference between the hybrid model and TF-IDF + SVM is modest, but it suggests that the semantic component can contribute additional information when combined with lexical probabilities. The result should not be interpreted as a new algorithmic breakthrough; it is an applied finding showing that a simple weighted fusion can be useful in a low-resource parliamentary archive.

The lower performance of CamemBERT embeddings alone should be interpreted in relation to corpus size, document length and OCR noise. Transformer embeddings can be powerful in large and well-annotated corpora, but here CamemBERT was used as a frozen pretrained embedding extractor and was not trained or fine-tuned on the parliamentary corpus. The archive also contains long, formulaic and sometimes noisy OCR-derived texts. TF-IDF captures explicit legal and administrative markers more directly, while CamemBERT embeddings contribute mainly when lexical overlap is insufficient. This explains why the semantic representation is weaker as a standalone classifier but useful when combined with lexical evidence.

B. Cross-Validation Stability

TABLE III.
STRATIFIED FIVE-FOLD CROSS-VALIDATION RESULTS.

Model	Accuracy mean +/- SD	Macro F1 mean +/- SD	Weighted F1 mean +/- SD
TF-IDF + linear SVM	0.821 +/- 0.052	0.834 +/- 0.050	0.817 +/- 0.054
TF-IDF + logistic regression	0.783 +/- 0.048	0.799 +/- 0.045	0.778 +/- 0.050
TF-IDF + Naive Bayes	0.593 +/- 0.012	0.523 +/- 0.015	0.525 +/- 0.014

The five-fold results confirm that TF-IDF + linear SVM is the most stable lexical model. Its cross-validation accuracy of 0.821 +/- 0.052 should not be directly compared with the hybrid hold-out accuracy of 0.786 because the evaluation protocols are different. Cross-validation averages performance across five partitions, while the hold-out score reflects one fixed test set. The standard deviation also indicates that performance varies across folds, which is expected with only 420 documents and an imbalanced minority class. The main conclusion is therefore that TF-IDF + SVM is a robust lexical baseline, while hybrid fusion is the best model under the fixed test protocol used for the full lexical-semantic comparison.

C. Per-Class Analysis and Confusion Matrices

TABLE IV.
PER-CLASS RESULTS FOR TF-IDF + LINEAR SVM.

Class	Precision	Recall	F1-score	Support
Orders	0.74	0.93	0.83	28
Decrees	0.82	0.64	0.72	14
Laws	0.90	0.69	0.78	13
Ordinances	0.70	0.67	0.68	24
Minutes	1.00	1.00	1.00	5

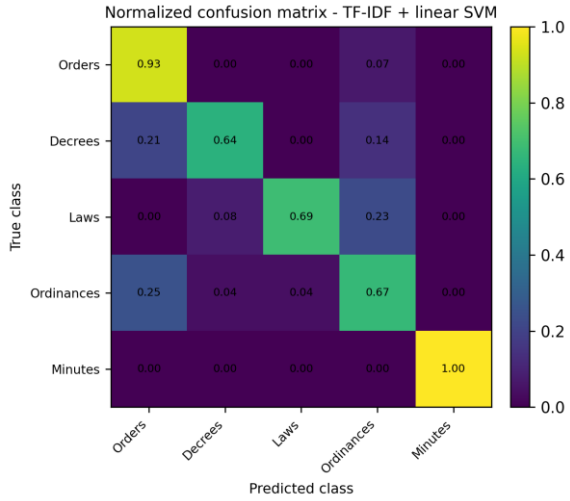

Figure 1. Normalized confusion matrix of TF-IDF + linear SVM.

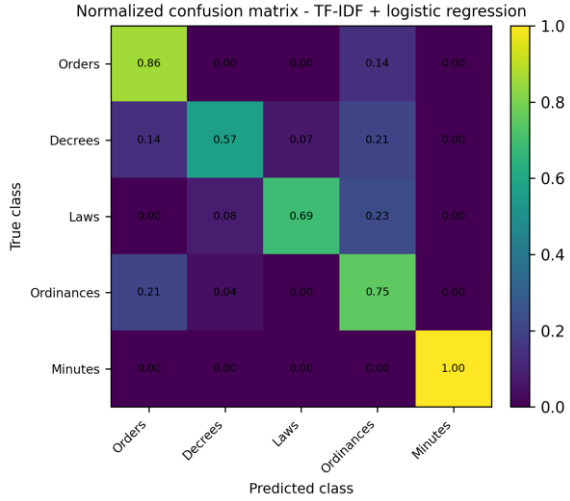

Figure 2. Normalized confusion matrix of TF-IDF + logistic regression.

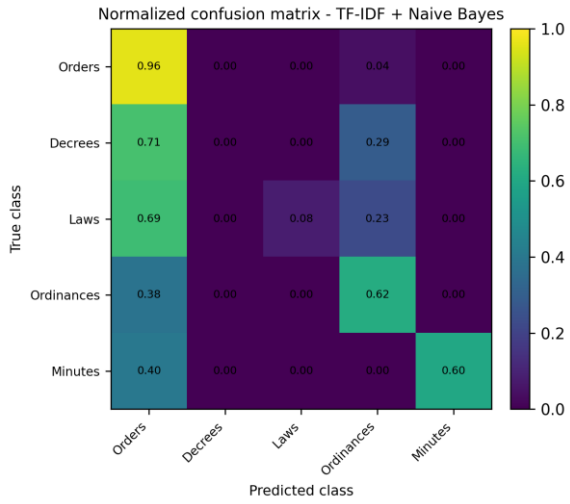

Figure 3. Normalized confusion matrix of TF-IDF + Naive Bayes.

The confusion matrices show that minutes are recognized with high stability, but this result must be interpreted cautiously because the test support is only five documents. Legal and regulatory categories remain more difficult. Decrees, laws and ordinances share institutional vocabulary, article-based structures, legal formulas, signatures and references to authorities. Misclassifications therefore appear to result mainly from three factors: lexical proximity between legal categories, OCR noise that modifies discriminative terms and document-series similarity when documents from the same administrative family have highly similar formulations.

Error analysis.

The most informative errors occur between decrees, laws and ordinances. These categories often contain the same legal style, such as preambles, numbered articles, enforcement clauses and references to institutional authority. A model may therefore give more importance to shared legal vocabulary than to the few terms that distinguish the document type. Naive Bayes is particularly affected because it assumes conditional independence between terms and is sensitive to frequent terms in short or noisy OCR outputs. By contrast, linear SVM separates classes more effectively in high-dimensional TF-IDF space, which explains its stronger stability.

For semantic search, the main failure cases are expected when queries are too broad, when relevant documents use very specific legal formulas, or when OCR errors break important terms. Lexical TF-IDF performs well when the query contains exact institutional markers. Hybrid search becomes more useful when the query is conceptual, for example when a user searches for an administrative action or legal issue without knowing the exact document title or category. This supports the practical recommendation to use lexical search for precise known-item queries and hybrid retrieval for broader exploratory archive search.

D. Semantic Search Results

TABLE V.
SEMANTIC SEARCH PERFORMANCE SUMMARY.

Metric	TF-IDF	CamemBERT	Hybrid
P@1	0.400	0.267	0.333
P@3	0.356	0.267	0.422
P@5	0.360	0.227	0.413
P@10	0.347	0.200	0.407
R@10	0.071	0.021	0.077
MRR	0.602	0.422	0.559
nDCG@10	0.351	0.220	0.400

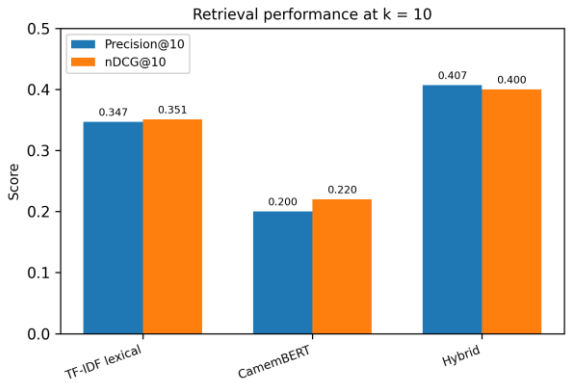

Figure 4. Retrieval performance at k = 10.

The search results show that hybrid search is the best method at k = 10, with Precision@10 = 0.407, Recall@10 = 0.077 and nDCG@10 = 0.400. Lexical TF-IDF search achieves the best MRR = 0.602, indicating that it often retrieves a relevant document very early when the query contains exact legal or institutional terms. CamemBERT embeddings alone perform less well, which may be explained by the fact that CamemBERT was used only as a frozen pretrained embedding extractor, without training or fine-tuning on parliamentary retrieval data, and by the small query set and OCR noise. The hybrid method improves the ranked list at larger cutoffs because it combines exact lexical evidence with semantic proximity.

E. Practical Implications for Parliamentary Archives

The achieved accuracy levels are practically meaningful only if the system is deployed with human validation. The classifier can pre-label incoming documents and reduce the time spent on initial sorting, while archivists retain the authority to confirm or correct the proposed class. This human-in-the-loop workflow is especially important for decrees, laws and ordinances, where legal proximity increases the risk of misclassification [3], [10].

The proposed workflow includes document digitization, OCR, text cleaning, automatic classification, metadata enrichment, semantic search, human validation and secure archival storage. This scope is broader than classification alone because archive modernization also requires retrieval, indexing, access control and knowledge management. In deployment, the classifier would pre-label documents, the validated class would be stored as metadata and the search engine would combine this metadata with lexical and semantic ranking.

F. Implementation and scalability considerations.

The present study did not include a formal system-level validation. In particular, it did not measure user experience, response time, indexing time, query latency, throughput, storage cost or scalability on a larger production archive. This limitation is important because practical archive modernization requires not only model accuracy but also acceptable operational performance and usability for archivists. Consequently, the present work should be interpreted as an algorithmic and workflow feasibility study, not as a complete deployment evaluation. From an implementation perspective, TF-IDF and linear classifiers are lightweight and easier to deploy on modest infrastructure, while CamemBERT-based embeddings require more memory and processing time. A future pilot deployment should therefore measure indexing time, retrieval latency, storage cost, user satisfaction, correction rates by archivists and scalability when the archive grows beyond the current corpus.

G. Data Leakage, Robustness and External Validity

Data leakage is a possible risk when highly similar documents or documents from the same administrative series are split between training and test sets. To reduce this risk, empty and obvious duplicate OCR outputs were removed before modeling. However, near-duplicate documents or series-based similarities may remain. Future work should apply document-series-aware splitting and temporal validation to better estimate deployment performance.

Robustness has been assessed descriptively rather than exhaustively. The models were tested on the available DRC Senate corpus, but not yet on future documents, new archive formats, larger collections or terminology from other institutions. The current results may therefore overestimate real deployment performance if future documents differ in layout, OCR quality or legal vocabulary. External validation should include temporal testing, document-series-aware splitting, additional parliamentary corpora and comparisons with recent multilingual dense retrieval models.

The conclusion that TF-IDF-based and hybrid models provide realistic solutions must therefore be limited to the studied corpus and research environment. External validity for other parliaments or legal archives requires additional empirical evaluation on new datasets, new time periods and other institutional workflows [1], [2], [10].

IV. CONCLUSION

This study evaluated an integrated workflow for automatic classification and semantic retrieval of parliamentary documents in a low-resource institutional context. The revised interpretation emphasizes that the contribution is applied and empirical rather than algorithmic: the work shows how established lexical, semantic and hybrid approaches can be organized into a human-validated archival workflow for the DRC Senate.

The results provide practical guidance for archive managers. TF-IDF-based approaches, especially TF-IDF + linear SVM, are preferable when document categories contain stable legal and administrative markers and when computational resources are limited. CamemBERT embeddings alone are less suitable in this setting because the corpus is small, OCR-derived and the CamemBERT model was not trained or fine-tuned on parliamentary data. The hybrid approach becomes most suitable when the goal is not only to classify documents but also to improve ranked retrieval for broader, paraphrased or semantically related archive queries. In all cases, deployment should remain human-validated, with archivists confirming or correcting automatic suggestions.

The study remains limited by the size of the corpus, the use of data from a single institution, the small set of thirteen curated search queries, the absence of detailed OCR quality measurement and the lack of system-level evaluation with real users. Future work should expand the corpus, construct a larger and more representative query set, test modern multilingual dense retrieval models, improve document-series-aware and temporal validation, measure indexing time, retrieval latency, scalability, user satisfaction and archivist correction rates, and integrate the classifier into a secure archival management platform.

ACKNOWLEDGMENT

The authors thank the academic supervisors and the institutional services that contributed to the reflection on parliamentary archive modernization and document digitization in the Democratic Republic of the Congo.

REFERENCES

[1] C. Sanderson and M. S. Irwin, "Transforming parliamentary libraries: Enhancing processes and services through artificial intelligence," IFLA Journal, 2025.
[2] F. Ariai and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," ACM Computing Surveys, 2025.
[3] A. Canning, "Artificial intelligence for reviewing government records and digital records management," Records Management Journal, 2025.
[4] M. A. Hedderich, L. Lange, H. Adel, J. Stroetgen, and D. Klakow, "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," NAACL-HLT, 2021.
[5] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," Information, vol. 10, no. 4, 2019.
[6] C. D. Manning, P. Raghavan, and H. Schuetze, Introduction to Information Retrieval. Cambridge University Press, 2008.
[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," EMNLP-IJCNLP, 2019.
[8] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," ACL, 2020.
[9] S. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," Natural Language Processing, 2025.
[10] M. Cueto Aparicio, "Guidelines for AI in Parliaments: best practices and governance principles," IFLAPARL / Westminster Foundation for Democracy, 2025.
[11] O. Q. Osagie, "The usefulness of artificial intelligence to safeguard records in libraries," South African Journal of Libraries and Information Science, 2024.
[12] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," European Conference on Machine Learning, 1998.
[13] L. Martin et al., "CamemBERT: A Tasty French Language Model," ACL, 2020.
[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.
[15] E. Chalkidis, I. Androutsopoulos, and N. Aletras, "Neural Legal Judgment Prediction in English," ACL, 2019.
[16] I. Chalkidis et al., "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," ACL, 2022.
[17] P. S. Nguyen, V. D. Nguyen, and M. Le Nguyen, "Legal domain knowledge acquisition for low-resource languages," arXiv preprint, 2023.
[18] S. Arimond, M. L. Tran, and M. Grabmair, "Transformer-based legal form classification," arXiv preprint, 2023.
[19] G. Krasadakis, A. Papadakis, and E. Papadakis, "Natural language processing challenges and advances in law consolidation," Electronics, vol. 13, no. 3, 2024.
[20] A. Principe, M. A. Fauzi, and A. D. Mauro, "Long document classification in the transformer era: challenges, models and evaluation," WIREs Data Mining and Knowledge Discovery, 2025.
[21] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," EMNLP, 2020.
[22] G. Izacard et al., "Unsupervised Dense Information Retrieval with Contrastive Learning," Transactions of Machine Learning Research, 2022.
[23] L. Wang et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training," arXiv preprint arXiv:2212.03533, 2022.