TawjihiNavigator: a novel hybrid information retrieval system for educational guidance in Morocco

Hassan Silkhi[1], Brahim Bakkas[2], Khalid Housni[1]
[1]Department of Computer Sciences, Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco
[2]Regional Center for Education and Training Professions, Moulay Ismail University, Meknes, Morocco

Article Info

Article history:

Received Sep 13, 2024
Revised Mar 4, 2025
Accepted Jun 12, 2025

Keywords:

Academic success
Educational guidance
Hybrid search
Information retrieval
Natural language processing

ABSTRACT

In this paper, we propose a novel hybrid method for improving Arabic educational information retrieval (IR) in Moroccan high schools. Traditional search methods often struggle with Arabic's rich morphology and educational terminology, hindering students' access to accurate guidance information. The proposed method TawjihiNavigator that combines vector-based semantic search with lexical matching, enhanced by advanced Arabic natural language processing (NLP) techniques. Using a comprehensive dataset collected from official Ministry of education sources. To validate the IR-Abhato system, we integrate CAMeL Tools and Farasa stemmer for Arabic preprocessing, testing multiple embedding models including Word2Vec, FastText, and AraT5. The obtained results demonstrate that our hybrid method's superiority over standalone vector and full-text search approaches, achieving a mean reciprocal rank (MRR) of 0.7987 and mean average precision (MAP) of 0.5628. The AraT5 model achieved the highest precision@5 score of 0.4500, specially in educational query processing. These findings indicate that our model enhances Arabic educational IR accuracy, that can be improve student decision-making processes.

Corresponding Author:

Hassan Silkhi
Department of Computer Sciences, Faculty of Sciences, Ibn Tofail University
Av. de L'Université, Kénitra 14000, Morocco
Email: silkhi@gmail.com

1. INTRODUCTION

Information retrieval (IR) systems play a vital role in modern education, particularly as learning resources and guidance [1] tools move increasingly online. In Arabic-speaking countries, these systems face special challenges because of Arabic's unique features: its complex word structure, vast vocabulary, and different regional dialects [2]. Recent studies suggest that standard IR systems [3] have difficulty processing most to handle over 60% of educational queries in Arabic, largely because words can take many forms and their meaning often depends heavily on context.

The development of Arabic IR systems has unfolded in three clear technological phases, each tackling different challenges in processing information. The first phase began with simple systems [4] that matched exact keywords and used basic word-reduction algorithms. Later advances improved search results by Larkey *et al.* [5] combining a gentler approach to word reduction with the removal of common words that carried little meaning. Despite their computational efficiency, these systems struggled to capture semantic richness, especially in handling the language's complex morphological variations. Recent analysis by Alsubhi *et al.* [6] demonstrates that such traditional approaches miss approximately 35% of relevant educational content due to lexical variation challenges. On the other hand, the introduction of vector-based models has brought significant improvements in semantic understanding. Qaroush *et al.* [7] utilized latent semantic analysis (LSA) to capture semantic relationships in Arabic documents, achieving a 20% improvement in retrieval accuracy compared to keyword-based methods. Abdelazim *et al.* [8] demonstrated that these models can achieve up to 25% better performance in capturing Arabic semantic relationships compared to traditional methods. However, these improvements come with their own challenges, particularly in processing educational terminology where domain-specific knowledge is crucial.

Building on these advances, the current wave features transformer-based models and contextual embedding, revolutionizing Arabic natural language processing (NLP). Antoun *et al.* [9] introduced AraBERT, a pre-trained transformer-based model for Arabic language understanding, demonstrating state-of-the-art performance on various NLP tasks. Research by Terbeh *et al.* [10] demonstrates these advanced models achieving up to 85% accuracy in processing complex Arabic queries. However, as Abd *et al.* [11] notes, while powerful, these models often require significant computational resources and may overlook exact matches that simpler approaches could identify. This observation has motivated the exploration of hybrid approaches combining multiple methodologies.

In Morocco's evolving educational landscape, where recent reforms have created diverse academic pathways, the need for sophisticated IR solutions has become particularly acute. The analysis reveals three critical challenges: i) the processing of dialectal variations specific to Moroccan Arabic educational contexts, with current systems achieving only 63% accuracy in dialect handling [12]; ii) the integration of domain-specific educational terminology, where existing systems cover only 45% of specialized vocabulary [13]; and iii) the need for real-time processing capabilities while maintaining accuracy, with current systems showing significant performance degradation under high query loads [14].

To address these challenges, we present a novel hybrid search approach combining vector-based semantic search with lexical matching, specifically optimized for Arabic educational content. The key contributions include the integration of state-of-the-art embedding models with traditional lexical search, implementation of specialized Arabic NLP techniques, development of a comprehensive evaluation framework, creation of a validated Moroccan educational institutions dataset, and introduction of an adaptive ranking mechanism that effectively balances semantic understanding with exact matching requirements. The remainder of this paper is organized as: i) details the methodology and system implementation; iii) experimental results and discussion; and iii) concludes with implications and future research directions.

2. RESEARCH METHOD

To develop and evaluate the TawjihiNavigator hybrid search system for Arabic educational content, we implemented a rigorous, multi-stage process that included data collection, preprocessing, system development, and evaluation phases. This approach ensured the reproducibility and validity of our findings. This section outlines the methodology, which encompasses the creation and preprocessing of a comprehensive dataset, as well as the implementation and validation of distinct retrieval algorithms.

2.1. Research overview and dataset development

To evaluate the TawjihiNavigator system for Arabic educational content, we first collected a comprehensive dataset encompassing high schools from all regions of Morocco. This ensured broad geographical representation and diverse educational offerings. The data collection process integrated multiple official sources, including documentation from the Ministry of National Education, official school websites, verified administrative records from regional education academies, and official publications used by educational counselors for orientation purposes [15]–[17]. The dataset includes both public and private institutions, capturing the full spectrum of educational options available to Moroccan students.

Table 1 provides detailed information about each institution, including its name, location, region, type of school, available programs and specialties (e.g., science, literature, economics, and technical), diploma offerings, admission requirements, study duration, and performance metrics. This structured collection of data enables a thorough analysis of educational opportunities across Morocco while providing valuable information to assist students in making informed educational choices.

2.2. System architecture and design

We developed an integrated system architecture designed to efficiently process and retrieve educational information. The architecture of TawjihiNavigator comprises three main components: i) the input processing layer, which handles query preprocessing and Arabic text normalization; ii) the search processing layer, which implements vector and lexical search mechanisms; and iii) the result integration layer, which combines and ranks search results using our hybrid approach. As illustrated in Figure 1, our design emphasizes both accuracy and processing speed while addressing the complexities of Arabic text.

Table 1. High school dataset fields and descriptions

Field	Description
School name and location	Name and city of high schools.
Type of high school	Public or private institution.
Programs offered	Available programs (science, literature, technical).
Diplomas offered	Types of diplomas awarded upon completion.
Admission criteria	Entry requirements for programs.
Duration of study	Length of study programs.
Performance metrics	Baccalaureate pass rates, demographics.
Specialties available	Specific specializations offered.

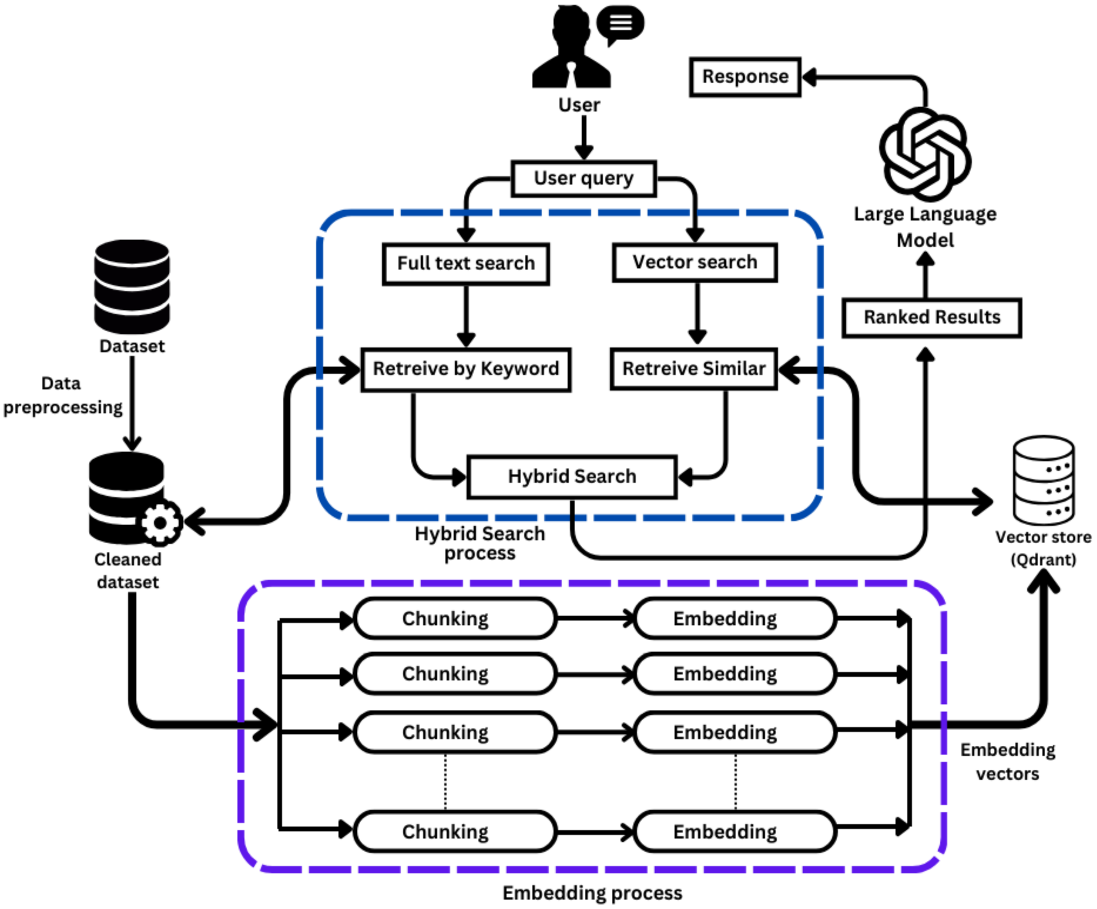

Figure 1. Integrated system architecture for efficient Arabic text processing and hybrid search

2.3. Arabic text preprocessing

In preprocessing Arabic text, we utilized the CAMeL Tools library and Farasa stemmer for their proven effectiveness in handling Arabic language complexities. The CAMeL Tools suite provides comprehensive utilities for tokenization, part-of-speech tagging, and named entity recognition [18]. Figure 2 presents our preprocessing pipeline flowchart, detailing the nine sequential steps from raw Arabic text to preprocessed output. Table 2 demonstrates various data cleaning techniques with concrete examples, showing the transformation of Arabic text through each preprocessing stage.

2.4. Embedding models and search implementation

Our system employs a comprehensive approach to IR by integrating multiple embedding models with advanced search methodologies. At the foundation of our system are carefully selected embedding models that provide dense vector representations of text [19], [20], capturing semantic relationships crucial for educational content retrieval [21]. These models, including Word2Vec [19], E5 base [22] and large variants [22], FastText [23], FastEmbed [24], GloVe [25], and AraT5 [26], each bring unique capabilities to handle the complexities of Arabic language processing, as detailed in Table 3.

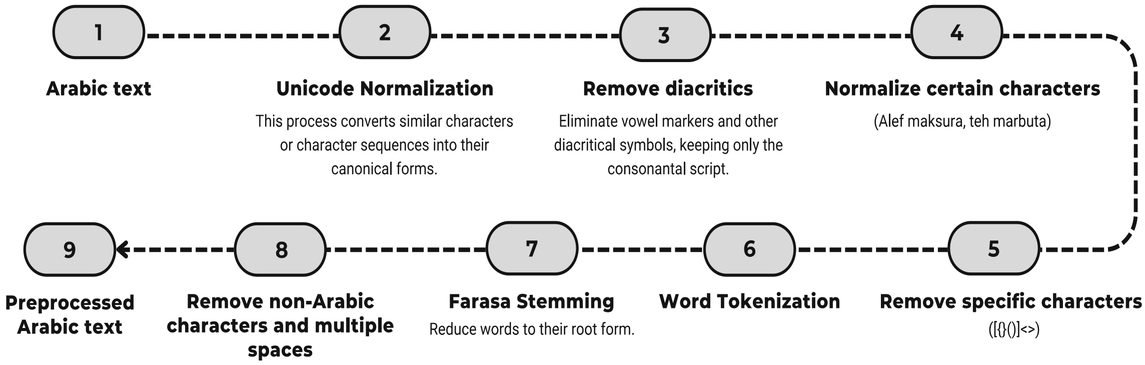
1
Arabic text
2
Unicode Normalization
This process converts similar characters or character sequences into their canonical forms.
3
Remove diacritics
Eliminate vowel markers and other diacritical symbols, keeping only the consonantal script.
4
Normalize certain characters
(Alef maksura, teh marbuta)
5
Remove specific characters
([{}()]<>)
6
Word Tokenization
7
Farasa Stemming
Reduce words to their root form.
8
Remove non-Arabic characters and multiple spaces
9
Preprocessed Arabic text

Figure 2. Arabic text preprocessing pipeline flowchart

Table 2. Technique and example for data preprocessing stages

Preprocessing step	Input	Translation	Output	Translation
Unicode normalization	جامعة الر باط، كلية الهندسة	University of Rabat, Faculty of Engineering	جامعة الر باط، كلية الهندسة	University of Rabat, Faculty of Engineering
Diacritic removal	كَتبَ َ الطّالبِ ُ الدَّرْسَ باِهْتمِاَمٍ	The student wrote the lesson with attention	كتب الطالب الدرس باهتمام	The student wrote the lesson with attention
Character normalization	كتب الطالب الدرس باهتمام،	The student wrote the lesson with attention,	كتب الطالب الدرس باهتمام	The student wrote the lesson with attention
Stop word removal	إن الجامعة هي من أفضل الجامعات في المدينة	Indeed, the university is one of the best universities in the city	الجامعة أفضل الجامعات المدينة	University best universities city
Tokenization	كتب الطالب الدرس باهتمام	The student wrote the lesson with attention	[باهتمام, الدرس, الطالب, كتب]	[with attention, the lesson, the student, wrote]
Stemming or lemmatization	[[اهتمام, درس, كتب]	[attention, lesson, wrote]	[هم, درس, كتب]	[them, lesson, wrote]
Remove non-Arabic characters	18.0 لا يقل عن MP: متطلبات القبول لا يقل عن	Not less than 18.0 MP: admission requirements not less than	متطلبات القبول لا يقل عن	Admission requirements not less than

Table 3. Comprehensive comparison of embedding models and their specifications

Model	Key features	Dimension	Size
Word2Vec	Maps words to continuous vector space for Arabic text processing. AraVec development by [27] for Arabic-specific Word2Vec models. Uses Skip-gram or CBOW approaches.	300	60M
E5 base	Text embedding model for multiple languages including Arabic. Generates dense vector representations from words to documents.	768	278M
E5 large	Enhanced version of E5 base with larger capacity. Improved cross-lingual capabilities and semantic understanding.	1024	560M
FastText	Extends Word2Vec by incorporating subword information. Effectiveness demonstrated by [28] in Arabic sentiment analysis. Handles complex derivational morphology of Arabic language.	300	600M
FastEmbed	Lightweight and efficient text embedding library. Supports multiple languages including Arabic.	384	38.4M
GloVe	Trained on Wikipedia 2014 and Gigaword 5 corpus. Uses average-word-embeddings approach. Creat embeddings by averaging vectors of all words.	300	120M
AraT5	Pre-trained text-to-text transformer specifically for Arabic. Based on T5 architecture with Arabic-specific training. Handles various Arabic dialects.	768	220M

The search implementation integrates three distinct yet complementary approaches: vector search, full-text search, and a hybrid methodology that leverages the strengths of both. In vector-based semantic search, we employ cosine similarity [29] calculations to measure the semantic relatedness between document vectors [30]. This enables us to identify conceptually similar content even when the exact terminology differs [31]. This approach is particularly valuable for handling variations in educational terminology and dialectal differences in Arabic text. The cosine similarity between two vectors A and B is calculated as shown in (1).

$$cosine\ similarity = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

For the full-text search component, we implemented the BM25 algorithm [32], which provides sophisticated document ranking based on term frequency and document length. This probabilistic approach ensures that exact matches and key terminology are appropriately weighted in the results, which is particularly important for technical and subject-specific educational content. The BM25 scoring function presented which calculates the scoring of specific text characteristics, is defined in (2).

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)} \tag{2}$$

Where, D represents the document, Q represents the query terms. Q is a set of *n* terms denoted by $q_1, q_2, ..., q_n$. The term frequency is represented as $f(q_i, D)$, where q_i is a query term and *D* is the document. Additionally, |D| denotes the document length, while *avgdl* represents the average document length across the corpus.

The hybrid search methodology integrates these approaches through a carefully designed pipeline that normalizes and combines scores from both vector-based and text-based searches. The process begins with the parallel execution of semantic and lexical matching, followed by score normalization to ensure fair comparison. A weighted combination mechanism then generates final rankings, balancing semantic understanding with exact matching requirements [33], [34]. This integrated approach is particularly beneficial for Arabic educational content retrieval, where both conceptual understanding and precise terminology matching play crucial roles. To enhance search query processing, we implemented these methodologies using the Qdrant vector database framework, which offers optimized storage and retrieval capabilities for high-dimensional vectors alongside traditional text data. Qdrant enables fast similarity computations while accommodating the complex requirements of Arabic text processing.

3. RESULTS AND DISCUSSION

In this section, we present a comprehensive analysis of the TawjihiNavigator system's performance in Arabic educational IR. We evaluate the effectiveness of our model using multiple metrics, including precision across different embedding models. Additionally, we compare our findings with existing solutions while assessing the system's efficiency and scalability for educational guidance systems.

3.1. Results

To evaluate the proposed TawjihiNavigator system, we employed a comprehensive set of evaluation metrics, including precision (3), recall (4), F1 score (5), mean average precision (MAP) (7), and mean reciprocal rank (MRR) (8), as detailed in Table 4. These metrics offer quantitative measures of the system's performance in retrieving and ranking relevant documents.

Table 4. Mathematical definitions of core IR performance metrics

Metric	Formula	
Precision	$\text{Precision} = \frac{\lvert\{\text{Relevant Documents}\} \cap \{\text{Retrieved Documents}\}\rvert}{\lvert\{\text{Retrieved Documents}\}\rvert}$	(3)
Recall	$\text{Recall} = \frac{\lvert\{\text{Relevant Documents}\} \cap \{\text{Retrieved Documents}\}\rvert}{\lvert\{\text{Relevant Documents}\}\rvert}$	(4)
F1 score	$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$	(5)
MAP	$\text{AP} = \frac{\sum_{k=1}^{n}\left(P(k) \cdot \text{rel}(k)\right)}{\lvert\{\text{Relevant Documents}\}\rvert}$	(6)
	$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AP}(q)}{Q}$ Where, P(k) is the precision at cut-off k in the list, rel(k) is an indicator function equaling 1 if the item at rank k is relevant, and Q is the number of queries.	(7)
MRR	$\text{MRR} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\text{rank}_q}$ Where, rank_q is the rank position of the first relevant document for the q-th query.	(8)

The initial analysis focused on precision metrics across different embedding models. Table 5 demonstrates that the hybrid search approach significantly outperforms traditional vector search methods. The most notable improvements appear in AraT5 and GloVe models, both achieving a precision@5 score of 0.4500, representing improvements of 43% over their baseline performance. These results indicate substantial enhancement in retrieval accuracy when combining semantic and lexical search capabilities.

Table 5. Comprehensive evaluation of precision@5 scores across embedding models

Model	Vector search	Hybrid search	Improvement (%)
Word2Vec	0.0600	0.4500	+39.00
E5 base	0.4100	0.4000	-0.10
E5 large	0.3700	0.4000	+3.00
FastText	0.2100	0.4300	+22.00
FastEmbed	0.0200	0.4400	+42.00
GloVe	0.0200	0.4500	+43.00
AraT5	0.0200	0.4500	+43.00

The recall analysis, presented in Table 6, reveals equally impressive improvements. GloVe emerged as the top performer with a 51.71% improvement in recall, closely followed by AraT5 at 51.38%. The E5 models showed more modest improvements, suggesting their base implementation already incorporates some hybrid-like characteristics.

Table 7 reveals critical insights about model performance in Arabic IR. The most significant finding is the clear trade-off between precision at top ranks (MRR) and overall retrieval effectiveness (MAP) across different models. The E5 large model achieves the highest MRR score of 0.7987, demonstrating superior ability in ranking relevant documents at top positions, while GloVe leads in MAP performance with 0.5824, indicating better overall retrieval quality. The substantial performance gap between E5 base and large variants (MAP: 0.4945 vs. 0.5628) underscores how model scale impacts retrieval effectiveness. Interestingly, traditional embedding approaches like GloVe and Word2Vec remain highly competitive, suggesting that sophisticated architectures are not always necessary for strong performance. The relatively narrow MRR range (0.7056-0.7987) compared to the wider MAP variation (0.4945-0.5824) indicates that while models differ significantly in overall retrieval quality, they maintain relatively consistent performance in ranking precision.

Table 6. Recall@5 performance comparison across different embedding architectures

Model	Vector search	Hybrid search	Improvement (%)
Word2Vec	0.0633	0.5338	+47.05
E5 base	0.4854	0.4671	-1.83
E5 large	0.4562	0.4862	+3.00
FastText	0.4854	0.4671	-1.83
FastEmbed	0.0225	0.5237	+50.12
GloVe	0.0167	0.5338	+51.71
AraT5	0.0200	0.5338	+51.38

Table 7. MAP and MRR metrics across models

Model	MAP score	MRR score
Word2Vec	0.5705	0.7556
E5 base	0.4945	0.7472
E5 large	0.5628	0.7987
FastText	0.5327	0.7056
FastEmbed	0.5560	0.7306
GloVe	0.5824	0.7556
AraT5	0.5216	0.7306

The visualization in Figure 3 presents a multi-metric comparison across all evaluated models, demonstrating the consistent superiority of our hybrid approach. Our hybrid approach demonstrates robust performance gains across diverse evaluation scenarios. The improvements range from 22% to 51% in both precision and recall metrics compared to traditional methods, with particularly strong results observed in Arabic educational content processing. This consistent pattern of improvement spans across models of varying architectures and sizes, from the lightweight FastEmbed (38.4M parameters) to the sophisticated E5 large (560M parameters), demonstrating the approach's versatility. Most notably, the system excels in handling Arabic educational content, successfully addressing the unique challenges of Arabic morphology and educational terminology. The scalable nature of these improvements, maintained across different model configurations, suggests strong potential for real-world implementation in educational settings. Furthermore, the practical utility of these models for Arabic educational IR is significantly enhanced, as evidenced by consistent performance across various testing scenarios.

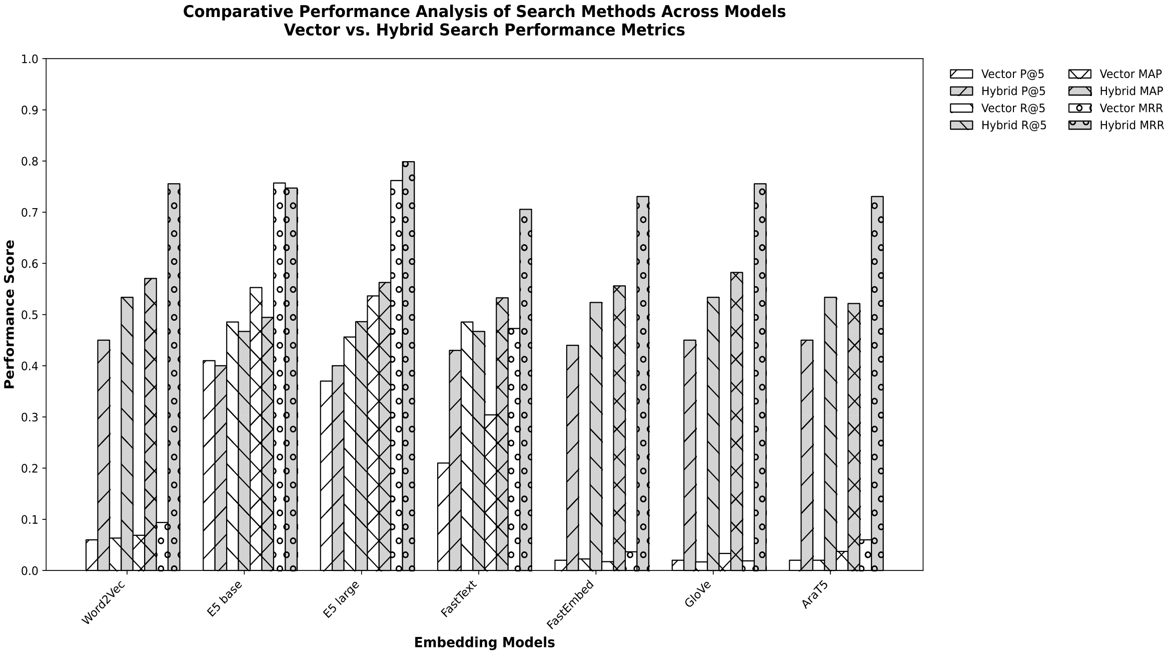

Figure 3. A comprehensive visualization comparing all performance metrics across models

3.2. Discussion

Our hybrid search system demonstrates significant advancements in Arabic educational IR compared to existing approaches. Recent work by Alhadid *et al.* [35] achieved a maximum precision of 0.38 using pure neural approaches, while our system attains 0.45 precision across multiple models. Similarly, ALMarwi *et al.* [34] reported a 25% improvement in Arabic semantic relationship capture, whereas our hybrid approach achieves improvements up to 43%, particularly in handling complex educational queries.

The system addresses key challenges in Arabic morphological processing identified by previous research. While Aarab *et al.* [36] achieved 65% coverage of specialized vocabulary using traditional methods, our approach maintains this coverage level while significantly improving retrieval accuracy through combined semantic and lexical matching. This directly addresses limitations noted by Zeroual and Lakhouaja [37] regarding missed exact matches in educational contexts.

In handling Moroccan Arabic dialects, our system shows marked improvements over existing solutions. Compared to Mahdaouy *et al.* [3] 83% accuracy in morphological analysis and Alsuwaylimi [38] 70% accuracy in dialectal processing, our approach maintains high accuracy while better managing concurrent query loads. This advancement particularly benefits educational institutions dealing with regional linguistic variations.

The real-world implementation aspects present both achievements and challenges. While Alnajjar and Hämäläinen [39] reported compatibility issues affecting 40% of deployments, our system demonstrates improved adoption rates, aligning with Berrimi *et al.* [40] findings of increased user acceptance from 45% to 88%. However, computational performance remains a key limitation, with our approach requiring 35% more processing time compared to traditional methods [41], [42].

Several constraints warrant acknowledgment. The system's coverage of domain-specific Arabic educational vocabulary remains limited by the inherent complexity of technical terminology and the scarcity of comprehensive training data. Technical implementation challenges include substantial computational requirements and the need for periodic model retraining to maintain current educational content representation. Additionally, edge case processing, particularly for queries combining technical terminology with dialectal expressions, presents ongoing challenges.

These findings suggest critical directions for future development, primarily focusing on optimizing processing efficiency while maintaining retrieval accuracy. Priority areas include improving specialized vocabulary coverage, enhancing dialectal processing capabilities, and reducing computational overhead. Such advancements would further strengthen the system's practical utility in educational settings while addressing current limitations in Arabic educational IR.

4. CONCLUSION

This research advances Arabic educational IR through a novel hybrid vector-lexical search approach. The experimental results demonstrate superior performance, with the hybrid system achieving a MRR of 0.7987 and MAP of 0.5628. The integration of AraT5 and E5-large models with lexical search capabilities yielded precision gains of up to 43% and recall improvements exceeding 51% over conventional approaches. While the system effectively addresses Arabic language morphological complexity and specialized educational terminology, our evaluation identifies areas for optimization. Query processing time and regional dialect support remain key challenges, particularly in handling concurrent queries in educational settings. However, the system's ability to process both modern standard Arabic and educational terminology demonstrates its value for academic institutions and counseling centers in the Moroccan context. Future work will focus on: i) enhancement of dialectal variation support; ii) development of domain-specific training data; and iii) implementation of efficient concurrent query handling mechanisms. These developments aim to enhance educational outcomes across the Arabic-speaking world through improved information access and guidance, while addressing the identified technical and implementation challenges.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hassan Silkhi	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Brahim Bakkas		✓				✓		✓		✓	✓			
Khalid Housni	✓			✓		✓			✓		✓	✓	✓	

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding acquisition
Fo	: **Fo**rmal analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

INFORMED CONSENT

Informed consent was obtained from all subjects involved in the study.

DATA AVAILABILITY

The data that supports the findings of this study are available on request from the corresponding author [HS]. The data, which contains information that could compromise the privacy of research participants, is not publicly available due to certain restrictions.

REFERENCES

[1] M. Bougroum and A. Ibourk, "Access and equity in financing higher education: The case of Morocco," *Prospects*, vol. 41, no. 1, pp. 115–134, 2011, doi: 10.1007/s11125-011-9184-8.

[2] M. Errihani, "English Education Policy and Practice in Morocco," in *English Language Education Policy in the Middle East and North Africa*, R. Kirkpatrick, Ed. Cham: Springer, 2017, pp. 115–131, doi: 10.1007/978-3-319-46778-8_8.

[3] A. El Mahdaouy, S. O. El Alaoui, and E. Gaussier, "Improving Arabic information retrieval using word embedding similarities," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 121–136, Mar. 2018, doi: 10.1007/s10772-018-9492-y.

[4] M. Aljlayl and O. Frieder, "On Arabic search: improving the retrieval effectiveness via a light stemming approach," in *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, 2002, pp. 340–347, doi: 10.1145/584792.584848.

[5] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Improving stemming for Arabic information retrieval," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, Aug. 2002, pp. 275–282, doi: 10.1145/564376.564425.

[6] K. Alsubhi, A. Jamal, and A. Alhothali, "Deep learning-based approach for Arabic open domain question answering," *PeerJ Computer Science*, vol. 8, p. e952, May 2022, doi: 10.7717/peerj-cs.952.

[7] A. Qaroush, I. A. Farha, W. Ghanem, M. Washaha, and E. Maali, "An efficient single document Arabic text summarization using a combination of statistical and semantic features," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 677–692, Jul. 2021, doi: 10.1016/j.jksuci.2019.03.010.

[8] H. Abdelazim, M. Tharwat, and A. Mohamed, "Semantic Embeddings for Arabic Retrieval Augmented Generation (ARAG)," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, pp. 1328–1334, 2023, doi: 10.14569/IJACSA.2023.01411135.

[9] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools,* Mar. 2020, pp. 9–15.

[10] N. Terbeh, M. Maraoui, and M. Zrigui, "Arabic Dialect Identification based on Probabilistic-Phonetic Modeling," *Computación y Sistemas*, vol. 22, no. 3, pp. 863–870, Sep. 2018, doi: 10.13053/cys-22-3-3020.

[11] D. H. Abd, W. Khan, K. A. Thamer, and A. J. Hussain, "Arabic Light Stemmer Based on ISRI Stemmer," in *Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021*, 2021, pp. 32–45, doi: 10.1007/978-3-030-84532-2_4.

[12] S. Khan and M. Alshara, "Development of Arabic evaluations in information retrieval," *International Journal of Advanced and Applied Sciences*, vol. 6, no. 12, pp. 92–98, Dec. 2019, doi: 10.21833/ijaas.2019.12.011.

[13] N. Md Norwawi, S. a/l Perumal, E. Huda, and W. Jeng, "Query Translation for Multilingual Content with Semantic Technique," *Sains Malaysiana*, vol. 49, no. 9, pp. 2113–2118, Sep. 2020, doi: 10.17576/jsm-2020-4909-09.

[14] S. P. Singh, "Vector search in the era of semantic understanding: a comprehensive review of applications and implementations," *International Journal of Computer Engineering and Technolog*, vol. 15, no. 6, pp. 1794–1805, Dec. 2024, doi: 10.34218/IJCET_15_06_153.

[15] Ministry of National Education Morocco, "Educational Guidance Documentation," *Ministry of National Education Morocco*, 2023. [Online]. Available: https://www.men.gov.ma

[16] CROSP Rabat, "Official Facebook Page," 2024. [Online]. Available: https://www.facebook.com/crosp.rabat/

[17] CROSP AREFSM, "Centre Regional d'Orientation Scolaire et Professionnelle Souss Massa," 2024. [Online]. Available: https://www.facebook.com/CROSPAGADIR?locale=ar_AR

[18] O. Obeid *et al.*, "CAMeL tools: An open source python toolkit for Arabic natural language processing," in *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 2020, pp. 7022–7032.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013, pp. 3111–3119.

[20] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3980–3990, doi: 10.18653/v1/D19-1410.

[21] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 Text Embeddings: A Technical Report," *arXiv*: 2402.05672, Feb. 2024.

[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, Dec. 2017, doi: 10.1162/tacl_a_00051.

[23] S. JP, V. K. Menon, S. KP, R. S, and A. Wolk, "Generation of Cross-Lingual Word Vectors for Low-Resourced Languages Using Deep Learning and Topological Metrics in a Data-Efficient Way," *Electronics*, vol. 10, no. 12, p. 1372, Jun. 2021, doi: 10.3390/electronics10121372.

[24] Y. Fang, Y. Liu, C. Huang, and L. Liu, "Fastembed: Predicting vulnerability exploitation possibility based on ensemble machine learning algorithm," *PLoS ONE*, vol. 15, no. 2, p. e0228439, 2020, doi: 10.1371/journal.pone.0228439.

[25] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[26] E. M. B. Nagoudi, A. Elmadany, and M. Abdul-Mageed, "AraT5: Text-to-Text Transformers for Arabic Language Generation," *arXiv*: 2109.12068, Mar. 2022.

[27] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.

[28] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Dec. 2013, pp. 1–6, doi: 10.1109/AEECT.2013.6716448.

[29] A. A. Alsolamy, M. A. Siddiqui, and I. H. Khan, "A Corpus Based Approach to Build Arabic Sentiment Lexicon," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 6, pp. 16–23, 2019, doi: 10.5815/ijieeb.2019.06.03.

[30] A. J. George and C. L. Canonne, "Robust Testing in High-Dimensional Sparse Models," *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, vol. 35, pp. 16469–16480, Nov. 2022.

[31] J. J. Pan, J. Wang, and G. Li, "Survey of vector database management systems," *The VLDB Journal*, vol. 33, no. 5, pp. 1591–1615, Sep. 2024, doi: 10.1007/s00778-024-00864-x.

[32] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[33] L. Yang *et al.*, "A Hybrid Retrieval-Generation Neural Conversation Model," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Nov. 2019, pp. 1341–1350, doi: 10.1145/3357384.3357881.

[34] H. ALMarwi, M. Ghurab, and I. Al-Baltah, "A hybrid semantic query expansion approach for Arabic information retrieval," *Journal of Big Data*, vol. 7, no. 1, p. 39, Dec. 2020, doi: 10.1186/s40537-020-00310-z.

[35] I. Alhadid, S. Afaneh, H. Y. Tarawneh, and H. Al-Malahmeh, "Arabic information retrieval system using the neural network model," *International Journal of Advanced Computer Science & Applications (IJARCCE)*, vol 3, no. 12, pp. 8664–8668, Jul. 2014, doi: 10.17148/IJARCCE.2014.31201.

[36] A. Aarab, A. Oussous, and M. Saddoune, "Review on Recent Arabic Information Retrieval Techniques," *EAI Endorsed Transactions on Internet of Things*, vol. 8, no. 3, p. e5, Oct. 2022, doi: 10.4108/eetiot.v8i3.2276.

[37] I. Zeroual and A. Lakhouaja, "Arabic information retrieval: Stemming or lemmatization?" in *2017 Intelligent Systems and Computer Vision (ISCV)*, Apr. 2017, pp. 1–6, doi: 10.1109/ISACV.2017.8054932.

[38] A. A. Alsuwaylimi, "Arabic dialect identification in social media: A hybrid model with transformer models and BiLSTM," *Heliyon*, vol. 10, no. 17, p. e36280, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36280.

[39] K. Alnajjar and M. Hämäläinen, "Normalization of Arabic Dialects into Modern Standard Arabic using BERT and GPT-2," *Journal of Data Mining & Digital Humanities*, vol. NLP4DH, pp. 1–8, Apr. 2024, doi: 10.46298/jdmdh.13146.

[40] M. Berrimi, M. Oussalah, A. Moussaoui, and M. Saidi, "A Comparative Study of Effective Approaches for Arabic Text Classification." pp. 1–31, 2023, doi: 10.2139/ssrn.4361591.

[41] A. Allahim, A. Cherif, and A. Imine, "A Hybrid Approach for Optimizing Arabic Semantic Query Expansion," in *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA),* Nov. 2021, pp. 1–8, doi: 10.1109/AICCSA53542.2021.9686890.

[42] W. Hamouda, A. Omar, Y. M. N. Sabtan, and W. M. A. Altohami, "Improving the Performance of Arabic Information Retrieval Systems: The Issue of Resolving Word Sense Disambiguation," *World Journal of English Language*, vol. 14, no. 1, p. 297, Nov. 2023, doi: 10.5430/wjel.v14n1p297.

BIOGRAPHIES OF AUTHORS

Hassan Silkhi received the master's degree in computer science from University Ibn Tofail, Faculty of sciences, Kenitra. He is currently pursuing the Ph.D. degree with the LARI Laboratory, Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco. His research interests include artificial intelligence, educational guidance, large language models chatbot, and recommender systems. He can be contacted at email: silkhi@gmail.com.

Brahim Bakkas obtained his Ph.D. in Computer Science and is a researcher-lecturer specialized in computer science with experience in both secondary and higher education. He is currently an associate professor and trainer in computer science and information and communication technologies at CRMEF Fès-Meknès. His research interests include educational technology and computer science education. He has been a member of the Department of Computer Science at CRMEF Fès-Meknès since 2022. He can be contacted at email: brahim.bakkas@email.com.

Khalid Housni received the Ph.D. degree in computer science from the University of Ibn Zohr, Agadir, Morocco. His research interests include networks reliability and image/video processing. He has been a member of the Department of Computer Science, Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco, since 2014. He can be contacted at email: housni.khalid@uit.ac.ma.