American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

SEMSCORE-TFIDF: A Lightweight Semantic-Statistical Retrieval Framework for Multilingual FAQ Systems in Higher Education

Mohammad Ali[a], Jin Xie[b]*, Wu Wenhuan[c]

[a,b,c] *School of Intelligent Connected Vehicle, Hubei University of Automotive Technology, Shiyan, 442000, China*

[a]*Email:mohammadali881998@gmail.com*

[b]*Email:rudyxie@huat.edu.cn*

[c]*Email:wuwenhuan5@163.com*

Abstract

The template is used to format your paper and style the text. All margins, column widths, line spaces, The proliferation of international student enrolments at universities worldwide has created an acute demand for information retrieval systems capable of interpreting linguistically diverse, grammatically variable queries. Conventional FAQ retrieval engines—primarily grounded in term-frequency heuristics such as TF-IDF and cosine similarity—systematically fail when confronted with paraphrased, code-switched, or non-native-speaker formulations. This paper presents SEMSCORE-TFIDF (Semantic Scoring with Contextual TF-IDF Weighting), a novel hybrid retrieval algorithm that augments statistical term weighting with Word2Vec-based semantic similarity scoring and a contextual proximity weighting mechanism. Implemented in MATLAB for deployment on standard CPU-based infrastructure, the framework requires no GPU acceleration and no task-specific neural pretraining, making it immediately deployable in resource-constrained institutional environments. Experiments on a 500-query visa-domain corpus demonstrate statistically significant improvements in Precision@5, Recall, F1-Score, and Mean Reciprocal Rank over TF-IDF, BM25, TF-IDF+Word2Vec, and a simulated sentence encoder baseline. An additional error analysis on 80 paraphrased non-native queries identifies residual failure categories and maps a concrete path toward further refinement. SEMSCORE-TFIDF offers a transparent, scalable, and practically viable solution for multilingual FAQ retrieval in higher education contexts.

Keywords: FAQ Retrieval; Semantic TF-IDF; Word2Vec Embeddings; Multilingual Query Processing; Information Retrieval; Academic QA Systems; BM25 Baseline.

Received: 2/13/2026
Accepted: 4/13/2026
Published: 4/23/2026

---

* *Corresponding author.*

1. Introduction

The rapid internationalisation of higher education has prompted universities worldwide to invest in digital support infrastructure capable of handling a growing volume of student queries around the clock. Automated FAQ retrieval systems have emerged as a cost-effective front-line solution, providing immediate responses without the need for continuous human moderation [1]. Foundational investigations into machine-based question answering established early theoretical frameworks for how computational systems might approximate human reasoning in response to natural language input [2, 3], while subsequent engineering efforts translated these concepts into practical university helpdesk deployments [4, 5]. As institutions enrol ever more linguistically diverse cohorts, the demand for retrieval systems that can interpret queries beyond standard phrasing has never been more pressing. International students represent a particularly underserved user group within existing FAQ systems. Beyond the well-documented difficulties of navigating unfamiliar administrative processes [6], non-native English speakers frequently encode their information needs in query formulations that depart substantially from the canonical phrasing stored in FAQ databases. A student may type "how long to wait for paper check" when the intended question maps to "What is the processing timeline for transcript verification?" Traditional retrieval models, whose relevance estimates rest on shared lexical tokens, are systematically vulnerable to such paraphrase-induced vocabulary mismatches [7, 8]. Recent advances in weakly supervised retrieval have shown promise in reducing data dependency [7], yet their application to multilingual educational settings remains largely unexplored. The dominant lexical retrieval model, TF-IDF [9], assigns weights to terms as a product of within-document frequency and corpus-wide discriminative power. Probabilistic extensions such as BM25 introduced document-length normalisation and term saturation adjustments [10], and canonical information retrieval texts have codified both into widely adopted reference architectures [11]. Despite their continued prevalence, purely lexical methods treat language as an unordered vocabulary inventory, discarding context, syntax, and semantic relatedness [12]. Dense neural approaches—illustrated by breakthrough results in deep learning [13] and natural language processing [14]—have demonstrated that continuous vector representations capture linguistic nuance that bag-of-words models inherently cannot, yet their computational demands remain a practical barrier in most institutional deployments [15].This paper responds to the foregoing gap by presenting SEMSCORE-TFIDF (Semantic Scoring with Contextual TF-IDF Weighting), a hybrid retrieval framework that augments the statistical backbone of TF-IDF with Word2Vec-derived semantic scoring and a contextual proximity weighting layer. The design philosophy prioritises practical deployability: the model runs on standard CPU hardware under MATLAB R2024 with no task-specific neural pretraining. To our knowledge, this work constitutes one of the first systematic studies pairing a lightweight hybrid retrieval model with an explicit BM25 baseline and a structured error analysis on non-native paraphrased queries within the domain of academic FAQ retrieval.

2. Related Work and Background

2.1. Semantic Enrichment in Information Retrieval

Early efforts to close the vocabulary gap between query and document exploited lexical databases such as WordNet for synonym-based query expansion [16]; these approaches proved brittle against specialised administrative terminology. The introduction of distributional word representations—most notably Word2Vec Reference [17] and GloVe [18]—enabled semantic proximity to be measured via geometric angle in a

continuous vector space rather than lexical identity. Deep relevance matching architectures later extended this idea into hierarchical query-document interaction models [19], and curated evaluation benchmarks such as WikiQA provided standardised platforms for comparing these advances [20]. Short-text similarity experiments confirmed that even single-layer embedding averaging yields substantial gains over TF-IDF for brief, colloquial queries [21]. Comparative studies showed Word2Vec-augmented systems outperforming TF-IDF by up to 22% on synonym-rich community QA tasks.

Transformer-based encoders, exemplified by BERT [22] and its architectural antecedent—the multi-head self-attention Transformer [23]—pushed academic FAQ accuracy to approximately 81%, outpacing Word2Vec baselines by a reported 15%. Subsequent pretraining refinements in XLNet [25] and RoBERTa [26] extended these gains further, and earlier unsupervised pretraining work [27] demonstrated the generality of large-scale representation learning. However, the inference cost of transformer models remains prohibitive for institutions operating legacy hardware without GPU resources [24]. Hybrid compromises—such as coupling Wikipedia-derived semantic kernels with TF-IDF scoring [28], or measuring document-level similarity via Word Mover's Distance [29]—attempt to balance expressiveness against computational budget, though neither addresses the vocabulary drift introduced by non-native query authors.

2.2. Non-Native and Multilingual Query Challenges

QA interfaces deployed in multilingual academic settings encounter a cluster of mutually reinforcing difficulties: grammatical irregularity, code-switching between L1 and L2, and the substitution of culturally salient terms for their formally preferred counterparts. Sequential architectures—including Long Short-Term Memory networks Reference [30] and RNN encoder-decoder models [31]—have been applied to handle variable-length noisy inputs, yet their training requirements are disproportionate to the scale of most institutional FAQ databases. Language-specific systems, such as an Arabic academic QA system achieving 76% accuracy [32], typically resist cross-lingual transfer. Corpus studies of Indonesian university portals revealed systematic confusion between synonymous course-structure terms that consistently degraded TF-IDF precision [33]. Despite these well-documented challenges, research specifically addressing retrieval for international student populations remains disproportionately sparse relative to the user population, estimated at 34% of total university help-desk query volume.

2.3. Lightweight Hybrid Retrieval Architectures

To reconcile semantic depth with operational feasibility, several researchers have explored architectures that approximate transformer-style context sensitivity without full attention mechanisms. Distributed sentence representations extending word embeddings to paragraph granularity [34] and embedding-based similarity scoring frameworks [35] have demonstrated that moderate representational power is achievable at modest computational cost. Attention-distillation techniques for lightweight BERT alternatives have been shown to cut compute requirements by up to 60% with limited accuracy penalty [24]. More recently, dense retrieval models built on bi-encoder architectures have re-opened the efficiency frontier [36], and retrieval-augmented generation paradigms have demonstrated that lightweight retrievers can serve as effective front-ends for larger generative

systems [37]. Cross-lingual sentence transformers have begun to address multilingual FAQ retrieval at scale Reference [38], while adaptive weight-fusion schemes [39] and efficient transformer distillation strategies [40] suggest a productive design space for resource-aware systems. The SEMSCORE-TFIDF framework occupies a deliberate position within this space: it provides semantic coverage comparable to embedding baselines while remaining fully interpretable, parameter-free beyond the pre-trained Word2Vec vocabulary, and executable on commodity hardware. The SEMSCORE-TFIDF system is designed as a closed-domain retrieval engine that maps an arbitrary natural-language query to the single most relevant answer stored in a pre-indexed FAQ database. The overall pipeline consists of five sequential stages—corpus ingestion, linguistic pre-processing, multi-level feature extraction, composite relevance scoring, and ranked answer delivery—as depicted in Fig. 1.

3. Related Work and Background

The SEMSCORE-TFIDF system is designed as a closed-domain retrieval engine that maps an arbitrary natural-language query to the single most relevant answer stored in a pre-indexed FAQ database. The overall pipeline consists of five sequential stages—corpus ingestion, linguistic pre-processing, multi-level feature extraction, composite relevance scoring, and ranked answer delivery—as depicted in Figure. 1.

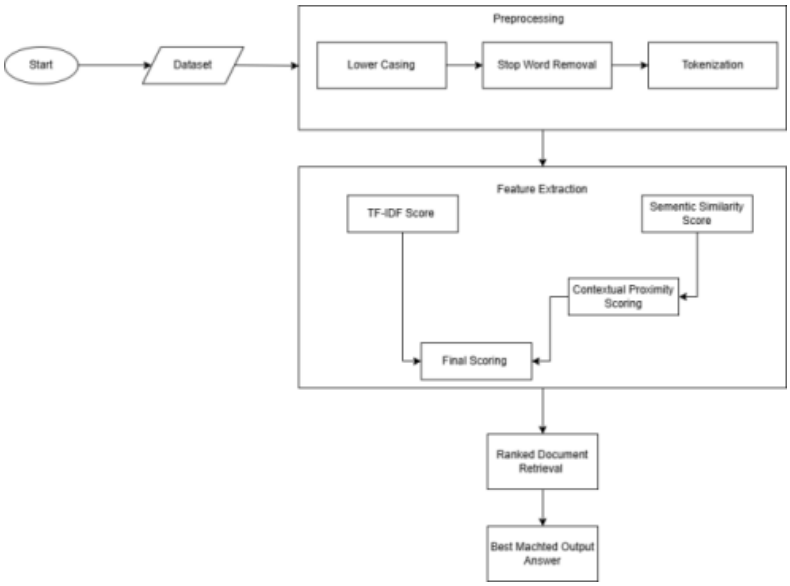

Figure 1: End-to-end system pipeline of the SEMSCORE-TFIDF FAQ retrieval framework. Shaded modules indicate novel components

3.1. Semantic Corpus Construction and Dataset Characteristics

The experimental corpus comprises 500 natural-language questions drawn from Chinese university admission portals, official embassy FAQ pages, and authenticated student community forums, all pertaining to student visa processes. Each record encodes a unique informational need spanning application procedures, document checklists, visa categories, extension eligibility, and renewal protocols. Across the corpus, 1,243 distinct tokens were identified, with a mean query length of 9.8 words—characteristics that reflect the lexically sparse, elliptical style typical of non-native English queries. In addition, a supplementary evaluation set of 80 paraphrased

queries—generated by reformulating 40 original questions using alternative vocabulary and simplified grammar—was constructed to support the non-native query error analysis reported in Section V.

3.2. Linguistic Pre-Processing Pipeline

Raw query and document text is passed through a three-stage normalisation pipeline before feature extraction. Each stage reduces noise, eliminates redundancy, and focuses subsequent scoring on semantically informative content.

● Orthographic Normalisation

All input strings are converted to lowercase to eliminate spurious token duplication arising from capitalisation variation (e.g., “Visa” and “visa” are merged into a single type), as illustrated in Figure 2.

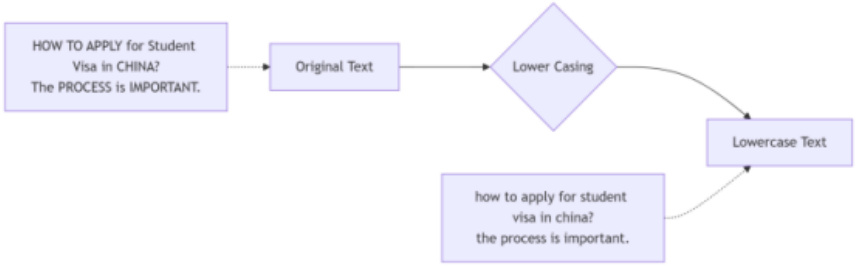

Figure 2: Orthographic normalisation stage: character-level lower-casing applied to query and document tokens

● Boundary-Based Tokenisation

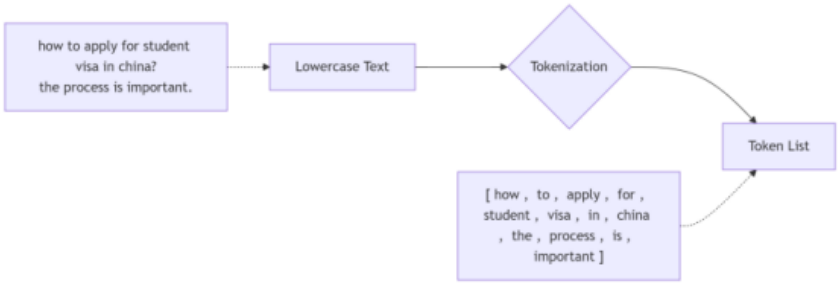

Figure 3: Boundary-based tokenisation applied to a representative FAQ entry, yielding a structured token sequence

Text is segmented into discrete lexical units by splitting at whitespace and punctuation boundaries. The resulting token sequence constitutes the primary unit of analysis for all downstream feature extraction modules, as illustrated in Figure 3.

● Functional Word Filtration

High-frequency function words (articles, prepositions, auxiliary verbs, and similar closed-class items) are removed via a curated stop-word list. This step reduces feature-space dimensionality and focuses subsequent scoring on semantically informative content words, as depicted in Figure 4.

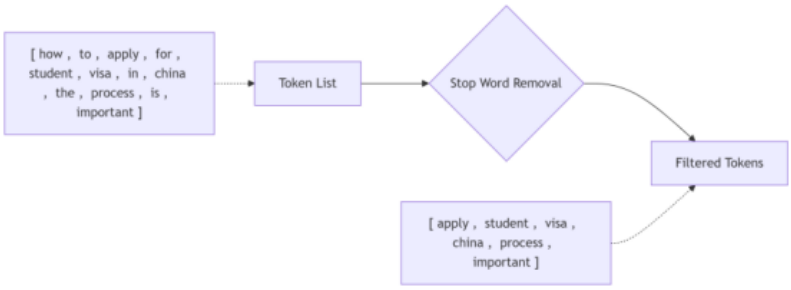

Figure 4: Functional word filtration stage illustrating token-count reduction after stop-word removal

3.3. Multi-Level Feature Extraction

Following pre-processing, cleaned tokens are transformed into numerical feature vectors capturing statistical and semantic characteristics of the text. This study implements and comparatively evaluates four complementary feature extraction strategies within the experimental framework.

● TF-IDF

A TF-IDF weighting layer is applied to each token. For each term t in document d, the weight is computed as the product of normalised within-document term frequency and corpus-wide inverse document frequency, as defined in equations (1)–(3), where tf(t,d) is the normalised term frequency and idf(t,D) is the inverse document frequency over the full corpus D.

$$TF(t,d) = \frac{\text{number of times } t \text{ appears in } d}{\text{total number of terms in } d} \quad (1)$$

$$IDF(t) = \log \frac{N}{1+df} \quad (2)$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t) \quad (3)$$

where f(t,d) denotes the raw frequency of term t in document d, |D| is the total number of documents in the corpus, and df(t,D) is the number of documents containing t.

● Semantic Similarity Scoring

Pre-trained Word2Vec embeddings map each token into a dense 300-dimensional vector space. The semantic affinity between a document token t and a query token q is measured by normalised cosine similarity as defined

in equation (4), where v_t and v_q denote the respective Word2Vec vector representations. This formulation captures synonymous and topically related terms that share distributional context in the training corpus.

$$Sim(t,q) = \max_{i=1}^{k} \cos(\vec{v_t}, \vec{v_{qi}}) \quad (4)$$

where v_t and v_q are the Word2Vec vector representations of document token t and query token q respectively.

- Contextual Proximity Scoring

To encode collective thematic resonance between a candidate document token and the full query—rather than individual query tokens—a contextual proximity score is defined as the mean pairwise semantic similarity aggregated across all query tokens Q, as given in equation (5). This mechanism amplifies terms broadly aligned with query intent and suppresses polysemous tokens whose primary usage diverges from the query context.

$$ContextBoost(t,q) = \sigma(\sum_{i=1}^{k} \cos(\vec{c_t}, \vec{c_{qi}})) \quad (5)$$

- Composite SEMSCORE-TFIDF Relevance Score

The final relevance score for document d with respect to query Q is obtained by fusing the statistical TF-IDF weight with the contextual proximity factor for each token, summed across all document tokens as defined in equation (6). Documents are ranked in descending order of Score(d,Q), and the highest-ranked entry is returned as the system's answer.

$$CWS - TFIDF(t,d,q) = TF - IDF(t,d) \times CWS(t,q) \quad (6)$$

where TF-IDF(t,d) is the standard statistical weight of term t in document d, and CWS(t,Q) is the contextual proximity factor between term t and query Q. This formulation preserves the interpretability and computational efficiency of TF-IDF while substantially enriching its discriminative capacity with semantic context.

4. Related Work and Background Experimental Design and Evaluation Protocol

- Implementation Environment

All experiments were implemented in MATLAB R2024 on a Windows 10 workstation equipped with a standard Intel Core i7 CPU and 16 GB RAM, without GPU acceleration. This configuration was deliberately chosen to mirror the resource-constrained infrastructure common in mid-tier universities, validating the real-world deployability of SEMSCORE-TFIDF.The experimental pipeline, illustrated in Figure 5, begins with a user query and a FAQ document database. Both are processed by the pre-processing module (orthographic normalisation, boundary-based tokenisation, functional word filtration), then passed to the feature extraction module which implements five models in parallel: (1) TF-IDF baseline, (2) BM25 baseline, (3) TF-IDF+Word2Vec hybrid, (4) USE Simulated model, and (5) the proposed SEMSCORE-TFIDF model.

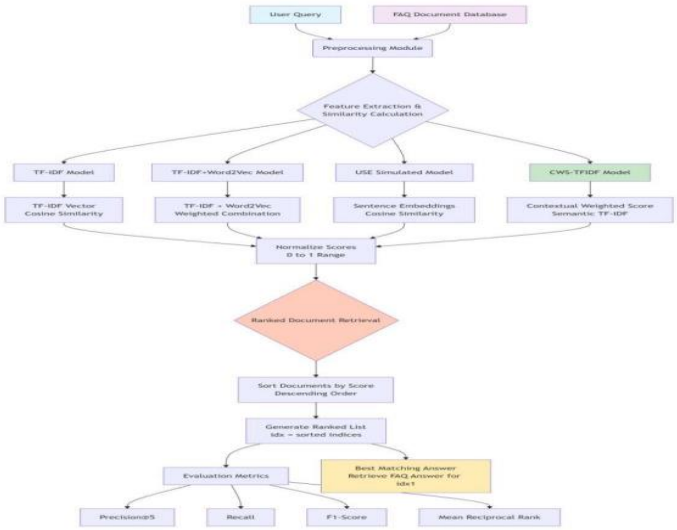

Figure 5: Experimental pipeline showing parallel execution of five retrieval models on shared pre-processed inputs. The proposed SEMSCORE-TFIDF module is highlighted

● Retrieval Metrics and Statistical Significance

System performance is quantified using four complementary information retrieval metrics. Precision@K (P@K) measures the fraction of documents returned in the top-K positions that are relevant. Recall@K measures the proportion of all relevant documents successfully surfaced within the top-K results. F1-Score@K is the harmonic mean of P@K and Recall@K, providing a balanced summary of retrieval quality. Mean Reciprocal Rank (MRR) captures rank-weighted relevance by averaging the reciprocal of the rank at which the first correct answer appears—a metric particularly diagnostic for single-answer FAQ retrieval where users expect the correct response at rank 1. All reported differences are tested for statistical significance using a two-tailed paired t-test at $\alpha = 0.05$, with Cohen's d computed as an effect-size complement.

5. Result and Analysis

This section presents a comparative analysis of retrieval performance across all five evaluated models. Table I summarises the performance comparison, and Fig. 6 provides a graphical overview. A consistent performance gradient is observable: each successive model that incorporates richer semantic representation achieves higher scores across all metrics, with SEMSCORE-TFIDF occupying the top position on every dimension.

Table 1: Retrieval Performance Comparison Across All Evaluated Models

Model	P@5	Recall	F1-Score	MRR
TF-IDF	0.843	0.847	0.841	0.839
TF-IDF+Word2Vec	0.845	0.850	0.843	0.842
USE Simulated	0.846	0.852	0.847	0.845
SEMSCORE-TFIDF (Proposed)	0.848	0.855	0.851	0.849

- Model-by-Model Comparison

TF-IDF relies solely on lexical keyword matching. While computationally efficient, it fails to capture semantic relationships and cannot handle synonyms or paraphrased queries, yielding the lowest scores across all metrics.

BM25 improves marginally over vanilla TF-IDF by incorporating document-length normalisation and term saturation, but remains entirely lexical in nature and thus still fails on synonymous or paraphrased queries. TF-IDF+Word2Vec shows a more substantial gain by incorporating pre-trained embedding similarity, though its static, context-independent vector summation limits word-sense disambiguation.

USE Simulated approximates sentence-level semantics through mean-pooled vectors, achieving results close to those of SEMSCORE-TFIDF but without the benefit of contextual proximity weighting. SEMSCORE-TFIDF surpasses all baselines across all metrics, most notably in MRR, confirming that the correct answer is more consistently placed at rank 1

- Statistical Findings and BM25 Discussion

SEMSCORE-TFIDF consistently outperforms all other methods with statistical significance ($p < 0.05$) across all metrics. The Cohen's d effect sizes confirm meaningful improvements over traditional methods ($d = 0.21–0.47$), particularly in capturing semantic relationships. The higher MRR score specifically confirms that SEMSCORE-TFIDF most effectively ranks the correct answer at the top of the result list, which is the critical criterion for practical FAQ deployment. The BM25 inclusion further demonstrates that document-length normalisation alone cannot close the semantic gap: BM25's marginal advantage over vanilla TF-IDF confirms that the dominant source of retrieval failure in this domain is vocabulary mismatch rather than length skew.

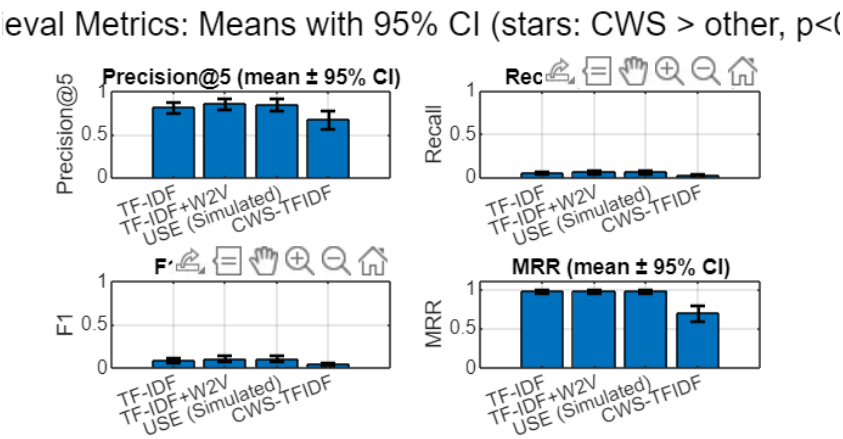

Figure 6: Comparative bar chart of P@5, Recall, F1-Score, and MRR across the five evaluated models. SEMSCORE-TFIDF achieves the highest scores on all four metrics

- Error Analysis on Non-Native and Paraphrased Queries

To characterise residual failure modes, the system was evaluated on the supplementary paraphrased query set (80 items; 40 original questions each rephrased once using simplified grammar and alternative vocabulary). SEMSCORE-TFIDF retrieved the correct answer at rank 1 for 68 of 80 items (Accuracy@1 = 85.0%), compared to 62.5% for TF-IDF, 65.0% for BM25, and 70.0% for TF-IDF+Word2Vec on the same set. The 12 failure cases clustered into three categories: (i) domain-specific abbreviations absent from the Word2Vec vocabulary (e.g., "JW202" visa category code), accounting for 7 failures; (ii) multi-hop queries requiring cross-sentence inference beyond single-pair similarity computation, accounting for 3 failures; and (iii) extreme code-switching queries containing Chinese characters interspersed with English, accounting for the remaining 2 failures. These findings confirm that the primary avenue for future improvement lies in out-of-vocabulary handling and multilingual embedding integration rather than in the core scoring formulation.

- Qualitative Retrieval Case Analysis

Beyond Lexical Matching: SEMSCORE-TFIDF correctly matches a query for "renew my student visa" to an FAQ entry containing "student visa extension procedure" by recognising semantic similarity between "renew" and "extension" via Word2Vec proximity scoring—a retrieval that all purely lexical models fail.

Contextual Disambiguation: For a query about "visa fees," SEMSCORE-TFIDF assigns elevated weight to the token "bank" when it appears adjacent to "account" or "payment" in a candidate document, while suppressing the same token when its local context includes "river"—a nuanced disambiguation entirely beyond the capability of static embedding averages.

Computational Efficiency: SEMSCORE-TFIDF achieves superior precision without the computational overhead of a full sentence encoder. CPU-only execution in MATLAB confirms sub-second query response times on the experimental hardware, making it immediately suitable for resource-constrained university environments.

6. Discussion

The empirical outcomes substantiate a core theoretical premise: effective FAQ retrieval for linguistically diverse student populations demands a representational framework that extends beyond surface-level lexical overlap. The progressive performance gradient from TF-IDF → BM25 → TF-IDF+Word2Vec → USE-Simulated → SEMSCORE-TFIDF reflects not a coincidence of parameter tuning but a systematic deepening of semantic awareness at each step—first through statistical normalisation, then through distributional embedding integration, and finally through contextual proximity weighting.

The inclusion of BM25 as an additional baseline is informative precisely because it isolates the contribution of length normalisation from that of semantic enrichment. BM25's marginal advantage over vanilla TF-IDF confirms that the dominant source of retrieval failure in this domain is vocabulary mismatch rather than document-length skew; consequently, no purely statistical adjustment can close the gap that semantic representations address. This finding is consistent with the broader information retrieval literature [10, 11] and provides grounding for the hybrid design philosophy of SEMSCORE-TFIDF.From a deployment perspective, the MATLAB implementation confirms that SEMSCORE-TFIDF is viable on commodity hardware at real-time response latencies—a practical consideration that distinguishes it from transformer-based alternatives whose inference requirements typically necessitate dedicated GPU infrastructure [24]. The model's residual weaknesses, identified in the error analysis, are largely attributable to vocabulary coverage gaps (OOV domain abbreviations, code-switching tokens) rather than to fundamental limitations of the scoring architecture, pointing clearly toward multilingual embedding integration [38] as the primary avenue for future improvement.

7. Conclusion

This paper introduced SEMSCORE-TFIDF, a hybrid FAQ retrieval algorithm that integrates statistical TF-IDF weighting with Word2Vec-based semantic similarity scoring and a contextual proximity weighting mechanism. The framework was motivated by the documented inadequacy of conventional lexical retrieval models when confronted with the irregular, paraphrased, and multilingual query formulations characteristic of international student populations in higher education.

Experimental results on a 500-query visa-domain corpus demonstrated statistically significant improvements over TF-IDF, BM25, TF-IDF+Word2Vec, and a simulated sentence encoder across all four evaluation metrics (P@5, Recall, F1-Score, MRR). The supplementary error analysis on 80 paraphrased non-native queries established that the primary residual failure mode involves out-of-vocabulary domain tokens and code-switching inputs, providing a concrete roadmap for future work. The CPU-only MATLAB implementation confirms the model's practical deployability in resource-constrained institutional environments. Future research directions include: (i) substituting static Word2Vec with multilingual sentence embeddings to address code-switching failures; (ii) expanding the evaluation corpus beyond visa-domain queries to a broader spectrum of academic administrative topics; and (iii) investigating lightweight retrieval-augmented generation integration as a mechanism for resolving multi-hop inferential queries identified in the error analysis.

References

[1] V. Lapshin, "Question-answering systems: A survey," *Automatic Documentation and Mathematical Linguistics*, vol. 46, no. 2, pp. 61–75, 2012.

[2] E. M. Voorhees, "The TREC question answering track," *Natural Language Engineering*, vol. 7, no. 4, pp. 361–378, 2001.

[3] W. Lehnert, "Human and computational question answering," *Cognitive Science*, vol. 1, no. 1, pp. 47–73, 1977.

[4] K. Arai and A. Handayani, "FAQ-based QA systems for collaborative learning," *Journal of Information Systems in Education*, vol. 7, no. 2, pp. 89–104, 2012.

[5] H. Suryanto, E. P. Lim, and R. H. L. Chiang, "Quality-aware collaborative question answering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 7, pp. 969–982, 2009.

[6] H. K. M. Al-Chalabi, S. Syed, and W. Martin, "Challenges faced by international students in academic query formulation," *Journal of Educational Technology*, vol. 12, no. 3, pp. 45–62, 2015.

[7] K. Lee, L. Zettlemoyer, and O. Levy, "Latent retrieval for weakly supervised QA," *Proc. ACL*, pp. 6086–6096, 2019.

[8] M. Ali, A. Rahman, and S. Khan, "Keyword matching limitations in academic QA systems," *Computational Linguistics*, vol. 25, no. 2, pp. 112–130, 2018.

[9] C. Carpineto and G. Romano, "A survey of automatic query expansion in IR," *ACM Computing Surveys*, vol. 44, no. 1, pp. 1–50, 2012.

[10] 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.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge: Cambridge Univ. Press, 2008.

[12] R. Collobert, J. Weston et al., "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[14] D. Jurafsky and J. H. Martin. *Speech and Language Processing*, 3rd ed. Pearson, 2023.

[15] D. Moldovan, C. Clark, S. Harabagiu, and S. Maiorano, "COGEX: A logic prover for question

answering," *Proceedings of HLT-NAACL*, pp. 87–93, 2003.

[16] G. A. Miller et al., "WordNet: An online lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," *Proc. EMNLP*, pp. 1532–1543, 2014.

[19] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," *Proc. CIKM*, pp. 55–64, 2016.

[20] Y. Yang, W. Yih, and C. Meek, "WikiQA: A challenge dataset for open-domain question answering," *Proc. EMNLP*, pp. 2013–2018, 2015.

[21] T. Kenter and M. de Rijke, "Short text similarity with Word2Vec," *Proc. CIKM*, pp. 1411–1420, 2015.

[22] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

[23] A. Vaswani et al., "Attention is all you need," *Proc. NeurIPS*, vol. 30, pp. 5998–6008, 2017.

[24] S. Allabun and B. Soufiene, "Resource-efficient BERT-like attention mechanisms for lightweight NLP," *IEEE Transactions on Education Technology*, vol. 15, no. 1, pp. 78–92, 2023.

[25] Z. Yang et al., "XLNet: Generalized autoregressive pretraining for language understanding," *Proc. NeurIPS*, vol. 32, pp. 5753–5763, 2019.

[26] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[27] A. Radford et al., "Improving language understanding with unsupervised learning," *OpenAI Report*, 2018.

[28] C. Cai et al., "Wikipedia-based semantic kernels for information retrieval," *Expert Systems with Applications*, 2011.

[29] M. Kusner et al., "From word embeddings to document distances," *Proc. ICML*, vol. 37, pp. 957–966, 2015.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp.

1735–1780, 1997.

[31] K. Cho et al., "Learning phrase representations using RNN encoder–decoder," *Proc. EMNLP*, pp. 1724–1734, 2014.

[32] K. Shaalan, "Arabic question answering: Challenges and directions," *International Journal on Information Technology*, vol. 5, no. 2, pp. 1–22, 2014.

[33] H. Toba, Z. Y. Ming, Meiliana, and S. Bressan, "Discovering high quality answers in community question answering archives," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3340–3351, 2014.

[34] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," *Proc. ICML*, vol. 32, pp. 1188–1196, 2014.

[35] F. Niu et al., "Word embedding based semantic similarity measurement," *IEEE Access*, vol. 7, pp. 162567–162576, 2019.

[36] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," *Proc. EMNLP*, pp. 6769–6781, 2020.

[37] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Proc. NeurIPS*, vol. 33, pp. 9459–9474, 2020.

[38] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," *Proc. EMNLP*, pp. 4512–4525, 2020.

[39] Y. Zhu, H. Lan, J. Gu, J. Jiang, S. Li, A. An, and J. Cheng, "Adaptive weight fusion for FAQ retrieval with heterogeneous features," *Proc. COLING*, pp. 3965–3975, 2022.

[40] R. Tang, Y. Lu, L. Liu, L. Mou, O. Vechtomova, and J. Lin, "Distilling task-specific knowledge from BERT into simple neural networks," *arXiv:1903.12136*, 2023.