Received: 31 August 2025 | Revised: 3 Feburuary 2026 | Accepted: 24 April 2026 | Published online: 3 June 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–12
DOI: 10.47852/bonviewAIA62027530

BON VIEW PUBLISHING

Attention-Based Explainability for Cross-Lingual Neural Information Retrieval in Low-Resource Languages: A LaBSE Framework with a Bilingual Medical Context Case Study

Nasir Hamzah[1] **and Zico Pratama Putra**[1,*]

[1] *Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia*

Abstract: The growing demand for equitable access to medical knowledge in multilingual contexts highlights a critical gap: traditional information retrieval (IR) systems perform poorly in low-resource languages, limiting access to medical expertise. Neural IR models, while effective, often lack interpretability—a serious concern in clinical applications such as ICD-10 code retrieval. Existing cross-lingual IR systems for low-resource languages are similarly constrained by limited task-specific tuning and the absence of real-time user feedback. This paper proposes an attention-based explainable framework for cross-lingual neural IR, specifically designed for bilingual ICD-10 code retrieval in Indonesian and English. The framework leverages Language-agnostic BERT Sentence Embedding—well-suited for low-resource settings—combined with an attention-based explainability mechanism that highlights token-level contributions, enhancing transparency and interpretability in model decision-making. The system is deployed as an interactive web-based application, demonstrating practical usability in clinical contexts. On the Indonesian MIRACL dataset, the framework achieves a mean reciprocal rank (MRR) of 0.782 and Recall@5 of 91.4% without fine-tuning. After domain-specific adaptation on the Master ICD-10 dataset, performance improves to MRR of 0.9352 and Recall@5 of 95.73%, surpassing strong baselines. These results demonstrate the viability of trustworthy, interpretable, and multilingual IR systems for healthcare in low-resource settings.

Keywords: attention mechanisms, cross-lingual information retrieval, ICD-10 code retrieval, low-resource languages, neural IR

1. Introduction

As multilingual content has grown globally in recent years, the need for equitable access to information across languages has made cross-lingual information retrieval (CLIR) an increasingly vital problem [1, 2]. Major languages including English have been effectively supported with a sufficient amount of annotated data and potent Natural Language Processing (NLP) models; however, low-resource language like Bahasa Indonesia is still lacking in modern information retrieval (IR) systems [3, 4]. This discrepancy presents a considerable challenge, particularly in critical applications like healthcare, where precise and interpretable retrieval of medical knowledge across languages is essential [5, 6].

The primary issue this research addresses is the urgent need for reliable and accessible cross-lingual retrieval of medical information, known as Question-and-Answering Information Retrieval (Q&A-IR), due to its lack of availability in low-resource languages [7, 8]. Existing conventional IR systems exhibit significant limitations in handling multilingual contexts, particularly for Bahasa Indonesia [9]. Additionally, neural IR models based on pretrained transformer architectures such as Language-agnostic BERT Sentence Embedding (LaBSE) outperform other libraries in semantic retrieval but either have lower interpretability or act as a black box [10]. This black-box nature remains a significant barrier in domains like medical diagnosis code retrieval, where the rationale for retrieval outcomes is critical to clinician trust and decision-making [11]. In addition, existing cross-lingual IR models for low-resource languages that have not been tuned to a particular task achieve suboptimal performance, and generally, they are not coupled with a real-time user interaction environment and machine–human bilingual feedback mechanism that can be critical in practical medical decision support systems [12, 13].

In contrast to English biomedical IR, which benefits from large-scale annotated corpora and domain-specific pretrained models such as PubMedBERT and MedCPT, Indonesian medical IR remains severely under-resourced. Publicly available Indonesian clinical datasets are limited in size and annotation depth, and no specialized pretrained retrieval models currently exist for this language. This resource gap motivates the need for cross-lingual retrieval frameworks capable of transferring semantic knowledge

***Corresponding author:** Zico Pratama Putra, Faculty of Information Technology, Universitas Nusa Mandiri, Indonesia. Email: zico.zpp@nusamandiri.ac.id

from high-resource languages while maintaining interpretability in low-resource clinical environments.

To provide a high-level understanding of the proposed approach for readers outside the IR domain, Figure 1 illustrates the overall conceptual workflow of the cross-lingual retrieval and explainability process, from bilingual query input to interpretable ICD-10 code retrieval.

In this paper, we seek to explicitly address and overcome these challenges by introducing an attention-based explainability framework for cross-lingual neural IR [14, 15]. Specifically, we aim to provide a solid CLIR system that utilizes LaBSE to effectively facilitate bilingual ICD-10 medical code retrieval for low-resource languages by semantically aligning queries written in either Indonesian or English with a bilingual medical code repository, integrate an attention-based explainability paradigm that adds transparency and interpretability elements by identifying the contributions of tokens from both query and retrieved document, and finally, make this tool available as an interactive web-based application, which we evaluate in various scenarios of medical IR [16, 17].

The contributions of this study are threefold and presented in an integrated narrative. First, we develop a CLIR framework that leverages LaBSE to effectively support low-resource languages, with a specific focus on bilingual retrieval of ICD-10 medical codes [18, 19]. This framework allows queries in either Indonesian or English to be semantically matched against a bilingual repository of medical codes [20, 21]. The initial training utilized the MIRACL dataset, and subsequent fine-tuning was performed on a comprehensive Master ICD-10 corpus, containing 10,469 documents with an average length of 6.36 words and a maximum length of 29 words [22, 23].

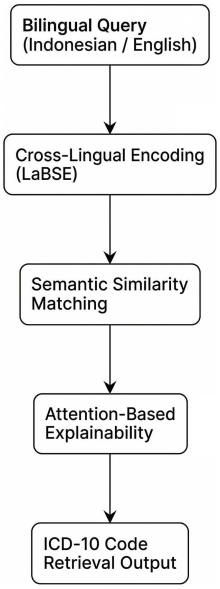
Second, the system uses an attention-based explainability mechanism that focuses on token-level contributions in both the query and the retrieved document [24, 25]. The feature improves both transparency and interpretability by allowing users, particularly those in clinical settings, to identify which segments of the input are responsible for the retrieved results [26, 27]. In the medical domain, attention to seemingly common words like "dengan" (with) or "yang" (which) is crucial for differentiating diagnoses and specific ICD-10 codes, as these terms often carry significant diagnostic implications.

Finally, we deploy the solution as an interactive web-based application, allowing users to enter medical queries and obtain ranked ICD-10 codes, complete with attention heatmaps and evaluation metrics such as mean reciprocal rank (MRR) and Recall@k, thereby demonstrating the system's usability and effectiveness in real-world medical IR scenarios [28, 29]. Our model demonstrates enhanced performance on the Master ICD-10 dataset compared to its initial training on MIRACL-ID, and it also outperforms the IndoBERT model on the same Master ICD-10 dataset.

The novelty of this work is the contextual synthesis and application of established components in neural IR (LaBSE and attention mechanisms) used to tackle specific challenges faced in low-resource clinical environments, an underexplored intersection in prior research. In contrast to general-purpose cross-lingual models or existing explainability methods, our solution represents three domain-specific improvements: a hand-crafted bilingual ICD-10 corpus that significantly expands the vocabulary gap of English–Indonesian terms; clinician-explainable attention visualizations that capture diagnostically relevant terms rather than syntactic patterns alone; and a deployable pipeline that optimizes healthcare systems in low-resource settings. This adjustment to real-world constraints represents an academic impact beyond the mere novelty of algorithms.

We thus show in this work that with the integration of both semantic retrieval and attention-based explainability, we can take strides in both understanding and interpreting cross-lingual IR, achieving greatly improved quality of output from these systems to help promote equality within inner-organizational learning environments across multilingual and lower-resource countries [30, 31]. The uniqueness of this work lies in the integration of attention-based explainability into a LaBSE-driven cross-lingual medical IR framework for bilingual ICD-10 code retrieval in a low-resource setting, combined with an interactive clinical application.

Figure 1
Conceptual overview of the proposed cross-lingual medical IR framework



2. Literature Review

2.1. Cross-lingual information retrieval in low-resource languages

CLIR aims to enable semantic retrieval across languages and has been widely studied using query translation, document translation, and multilingual representation learning approaches [1, 2]. Recent advances in multilingual transformer models and sentence embeddings have improved cross-lingual alignment, particularly for underrepresented languages [3, 4, 32]. However, most CLIR research still focuses on high-resource languages and general-domain datasets, while low-resource settings remain challenging due to linguistic diversity and data scarcity. In addition, interpretability is rarely addressed in CLIR systems, despite its importance in sensitive domains such as healthcare.

2.2. Medical and ICD code retrieval systems

Medical IR systems play a critical role in supporting clinical decision-making, including diagnosis and coding tasks. Prior studies on ICD code assignment predominantly formulate the problem as supervised multi-label classification using deep learning and transformer-based architectures [14, 18, 33]. While

biomedical retrieval models trained on large-scale English corpora, such as PubMed-based representations, demonstrate strong performance in general biomedical search tasks, their effectiveness often degrades in multilingual or low-resource medical contexts without domain-specific adaptation. Recent studies have also explored synthetic data generation using machine learning (ML) techniques to support radioactive waste management in Indonesia, demonstrating the potential of data augmentation strategies in resource-constrained healthcare and safety domains [34]. Retrieval-oriented approaches for medical concepts remain comparatively less explored than classification-based methods, particularly in bilingual settings. Recent efforts in Indonesia have demonstrated the integration of ML models to enhance radioactive waste management of disused sealed radioactive sources, showcasing the value of ML-driven approaches in specialized, safety-critical domains with limited data [35].

2.3. Explainability in medical NLP and information retrieval

Explainability has become a central concern in medical NLP to enhance transparency, trust, and accountability in clinical applications [11, 26, 27]. Existing explainable medical NLP studies commonly employ attention mechanisms and attribution-based techniques to interpret model predictions in classification tasks, including explainable ICD coding and clinical text analysis. Although explainability has also been investigated in multilingual and cross-lingual NLP tasks, its integration into cross-lingual medical IR systems remains limited, especially for retrieval-based ranking scenarios in low-resource languages.

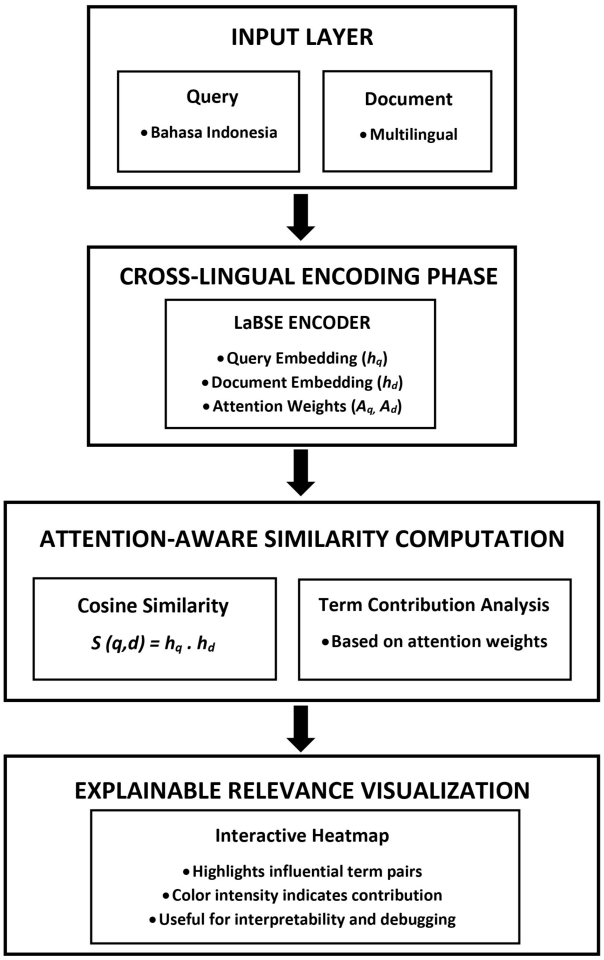
2.4. Research gap and motivation

Despite these advances, existing research predominantly focuses on monolingual or high-resource scenarios and emphasizes classification-based medical NLP rather than retrieval-oriented ranking in bilingual clinical environments. Complementary works in related low-resource Indonesian contexts have applied supervised ML for categorization and prediction tasks in hazardous material management, highlighting the importance of domain-specific modeling [36]. Moreover, explainability mechanisms are rarely incorporated into cross-lingual medical IR systems for low-resource languages, even though interpretability is essential for clinical trust and decision support. These limitations highlight a clear research gap, motivating the proposed attention-based explainable framework for bilingual ICD-10 code retrieval that aims to improve both retrieval effectiveness and model interpretability.

3. Research Method

Our framework leverages the language-independent informative character of LaBSE based on the attention code-based explainability ability for the cross-lingual neural information capture process. To accomplish this, the textual input was cross-lingually encoded (phase 1), followed by the computation of similarity utilizing attention (phase *n*), and concluded with visual explanations identified that corresponded to relevance score generation (See Figure 2). By utilizing this end-to-end approach, we not only enhance the precision of our retrieval process but also gain valuable insights into how the model arrives at its conclusions. To enable usability in real-world environments, the whole pipeline is being deployed as a web-based application using

Figure 2
Framework architecture for cross-lingual information retrieval with explainability

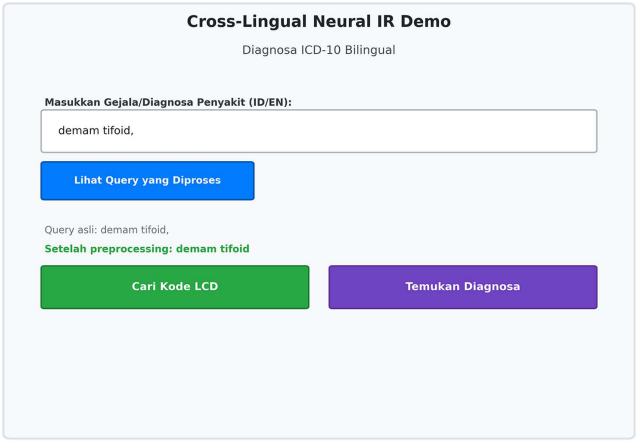


Streamlit. This is illustrated in Figure 3, where users enter queries in Indonesian or English that are normalized by a domain-specific preprocessor designed using several medical terminologies (e.g., "nyeri" → "sakit", "fever" → "pyrexia"). The structured queries are then matched with a bilingual ICD-10 code database, allowing retrieval in both directions (EN ↔ ID) and facilitating inclusivity in multilingual clinical contexts.

3.1. Query preprocessing

The system starts by preprocessing the raw user queries using a specific pipeline, such as shown in the demo interface (Figure 2), where an input of "demam tifoid" is trimmed to become "demam tifoid." Further, this approach helps maintain linguistic consistency and clinical relevance by implementing a specific set of text normalization rules for the medical domain. Punctuation and noise are also filtered out, and synonymous words (e.g., regional or colloquial) are mapped with bilingual medical lexicon to standard ICD-10 keywords. For example, "typhoid" and "tifoid" are harmonized to a form that we will call canonical, and case differences are normalized with lowercase conversion. This step improves the LaBSE encoder's generation of semantically accurate embeddings directly by reducing noise and unifying vocabulary. Not only is the preprocessed end product consistent with the structured ICD-10 corpus, but it also enables

Figure 3
Preprocessing pipeline demo interface



transparent retrieval, as demonstrated by who can easily differentiate between raw and processed queries in the user-friendly demo.

There was also no explicit stemming or rule-based medical tokenization in this study. Instead, the framework uses LaBSE's subword-level encoding, which is robust to morphological variance and compound medical terms between languages. Although domain-specific tokenization and entity-aware preprocessing can further boost retrieval performance, especially for longer clinical narratives, this work concerns itself with short, structured ICD-10 descriptions where subword modeling suffices.

3.2. Cross-lingual encoding with LaBSE

We employ LaBSE as our backbone encoder due to its proven effectiveness in multilingual semantic tasks. Given a query–document pair $(q,\ d)$ in a low-resource language (e.g., Indonesian), LaBSE generates contextualized embeddings as shown in Equation (1):

$$h_q = LaBSE\,(q),\ h_d = LaBSE\,(d) \tag{1}$$

where h_q and $h_d \in \mathrm{R}^{768}$ are sentence-level embeddings. To preserve fine-grained term interactions, we extract token-level attention weights A_q and A_d from LaBSE's final transformer layer, which later fuel our explainability module.

Additionally, the ICD-10 document corpus used for retrieval is stored in a bilingual format, combining Indonesian and English descriptors for each code. This corpus is dynamically queried in real time from a structured SQL database, ensuring extensibility and rapid lookup. This setup enables the system to support clinical code search across different linguistic preferences, aligning well with real-world deployment in multilingual healthcare environments.

The bilingual ICD-10 corpus construction involved three systematic steps: (1) extraction of official ICD-10 codes and English descriptions from WHO specifications, (2) professional translation of descriptions into Indonesian by certified medical translators, and (3) validation through cross-verification with existing Indonesian medical terminology databases. For triplet loss optimization, negative samples were generated using a hard negative mining strategy, where for each positive query–document pair, we selected the top-3 semantically similar but irrelevant documents based on initial LaBSE, ensuring challenging yet meaningful negative examples. The training employed 5-fold cross-validation to ensure robust performance estimation, with each fold maintaining the 70/15/15 train/validation/test split while preserving query–document pair integrity across linguistic variants.

3.3. Attention-aware similarity scoring

The relevance score $S(q,\ d)$ is computed via cosine similarity between query and document embeddings, as defined in Equation (2):

$$S(q,\ d) = \frac{h_q \,.\, h_d}{||h_q||\ ||h_d||} \tag{2}$$

To enhance transparency, we derive term contribution scores by aggregating cross-attention weights between query terms q_i and document terms d_j(3):

$$Contribution\,(q_i, d_j) = \sum_{l=1}^{L} A_q^{(l)}\,(i)\,.\,A_d^{(l)}[j] \tag{3}$$

where L is the number of attention heads. This formulation quantifies how strongly each term pair influences the final similarity score—offering a novel interpretability advantage in low-resource IR contexts.

This attention-aware formulation also enables downstream utility such as explainability filtering, where document–query pairs with low contribution variance can be deprioritized or reranked, supporting user-centered decision refinement in clinical applications.

3.4. Computational complexity analysis

From a computational perspective, the dominant cost of the proposed framework arises from the LaBSE encoding stage, which scales linearly with the input length and embedding dimension, that is, $O = n \cdot d$, where n denotes the number of tokens and d the embedding size. The similarity computation between query and document embeddings is performed using cosine similarity with a complexity of $O(d)$. The attention-based explainability mechanism operates at the token level and introduces an additional cost of $O(n^2)$ due to pairwise token interactions. However, given the short and structured nature of ICD-10 descriptions (average length 6.36 tokens), this overhead remains negligible in practice. Consequently, the overall framework maintains computational efficiency and is well-suited for real-time clinical retrieval in resource-constrained healthcare environments.

3.5. Explainability module

To address the need for transparency and interpretability in neural IR—especially in clinical code retrieval—our system incorporates an attention-based explainability module that provides multilevel visual explanations for each retrieval result. For every user-submitted query, the system not only returns a top-k ranked list of ICD-10 codes but also visualizes the internal decision signals that led to these rankings. The interface presents the top-5 most relevant ICD-10 codes alongside their semantic similarity scores. Each code entry includes bilingual label descriptions (English and Indonesian), allowing users to verify relevance across linguistic boundaries. For example, given a query such as "demam tifoid," the system retrieves codes like A01.0 (typhoid

fever/demam tifoid) and A75.9 (typhus fever, unspecified/demam tifus, tidak spesifik), ordered by computed similarity scores.

To highlight which terms in the query contributed most strongly to document matching, token importance is visualized using a heatmap and a query-side word cloud. These are generated from the attention vectors extracted from LaBSE's final transformer layers. The word "demam" dominates the attention distribution, indicating it as the principal semantic driver in the retrieval process.

Complementing the query-side attention, we also visualize document-side attention contributions to provide a complete explanation of the matching process. For each top-ranked document, a token importance heatmap and a document-side word cloud are generated. The heatmap highlights token-wise attention intensities within the document, revealing which terms received the highest focus from the model during the relevance computation. Meanwhile, the word cloud emphasizes the most influential tokens in a more intuitive and user-friendly manner. As illustrated, terms such as "fever," "typhoid," and subwords like "hoid" appear prominently—indicating their semantic alignment with the original query.

Training and Evaluation

The model was initially trained and evaluated using the Indonesian subset of the MIRACL dataset, which comprises 10,000 query–passage pairs annotated with human-assessed relevance labels. The dataset is partitioned into training (70%), validation (15%), and test (15%) sets. Subsequently, the model underwent fine-tuning using a comprehensive Master ICD-10 dataset. For fine-tuning, the LaBSE model is optimized using triplet loss, as defined in Equation (4):

$$\mathscr{L} = \max\left(0, \alpha + S(q, d^{-}) - S\left(q, d^{+}\right)\right) \tag{4}$$

where d^{+} and d^{-} denote positive and negative passages, respectively, and $\alpha = 0.2$ represents the margin hyperparameter.

Triplet loss is chosen over traditional classification loss because it directly optimizes for ranking relevance, a critical requirement in IR tasks. This objective ensures that relevant documents are consistently ranked higher than nonrelevant ones, even when semantic similarity is subtle or multilingual gaps exist.

To facilitate replication and extension of this work, all datasets and resources are publicly accessible. The Indonesian subset of the MIRACL dataset was obtained from Hugging Face Datasets[1], serving as the foundation for training and evaluation with predefined 70/15/15 splits. For domain-specific retrieval, we curated and open-sourced a bilingual ICD-10 corpus[2] (English–Indonesian) containing standardized medical terminology mappings. The complete implementation, including LaBSE fine-tuning scripts with triplet loss ($\alpha = 0.2$), attention visualization modules, and the Streamlit web interface, will be released upon publication completion to maintain anonymity. All experiments were conducted using PyTorch 2.0 and Hugging Face Transformers on NVIDIA A100 GPU infrastructure, with detailed configuration parameters documented for reproducibility.

Performance is assessed using standard ranking metrics, including MRR to measure ranking quality and Recall@k to quantify the proportion of relevant documents retrieved in the top-k results. These metrics are integrated into the web interface, enabling continuous monitoring of model behavior and comparative evaluation across model checkpoints or fine-tuning strategies. The training configuration and hardware environment are summarized in Table 1, ensuring reproducibility and methodological transparency.

The selected configuration provides stable convergence while maintaining efficiency for real-time retrieval. All experiments were conducted under identical conditions to ensure a fair and reproducible comparison across models.

Table 1
Model training configuration

Parameter	Value	Description
Learning rate	2e-5	AdamW optimizer with weight decay 0.01
Batch size	16	Number of samples per update
Epochs	10	Total training iterations
Margin α (triplet loss)	0.2	Ranking separation parameter
Optimizer	AdamW	Standard optimizer for transformer fine-tuning
Hardware	NVIDIA A100 (40 GB VRAM)	CUDA acceleration
Framework	PyTorch 2.0 + Hugging Face	Training and fine-tuning environment

4. Results and Discussion

This section presents the outcomes of the proposed cross-lingual neural retrieval system for ICD-10 code search, along with a discussion on the interpretability features enabled by attention-based visualization. The experiments were conducted using a comprehensive Master ICD-10 dataset, which was combined with a curated bilingual corpus of ICD-10 codes containing parallel descriptions in English and Indonesian. The evaluation focuses on two key aspects: retrieval performance and explainability effectiveness.

4.1. Retrieval performance

The model demonstrated strong performance in retrieving relevant ICD-10 codes based on user queries expressed in Bahasa Indonesia. We conducted a comparative evaluation of our fine-tuned LaBSE model against IndoBERT on the Master ICD-10 dataset. Our LaBSE model achieved robust performance, including high MRR and Recall@k scores, as presented in Table 2. This table allows for a direct numerical comparison of the performance metrics, illustrating our model's effectiveness across various recall depths. These results indicate the model's superior ability to consistently rank relevant ICD codes among the top retrieved items, which is also clearly depicted in Figures 4 and 5. Figure 4 provides a visual comparison of the MRR between the LaBSE and IndoBERT models, where higher bars signify better ranking quality. Similarly, Figure 5 illustrates the Recall@k performance

[1] https://huggingface.co/datasets/miracl/miracl

[2] https://github.com/fendis0709/icd-10

Table 2
Comparative performance evaluation of LaBSE fine-tuning with IndoBERT

Model	MRR	Recall@1	Recall@5	Recall@10	Recall@20
IndoBERT ICD-10	0.8945	0.8838	0.9064	0.9139	0.9217
LaBSE ICD-10	0.9352	0.9164	0.9573	0.9759	0.9843

Figure 4
Mean reciprocal rank comparison between LaBSE and IndoBERT

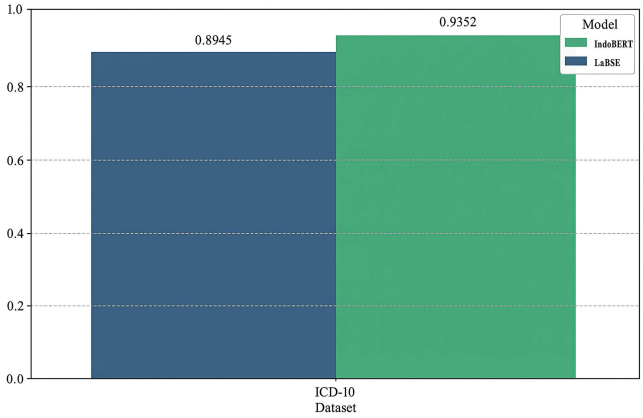
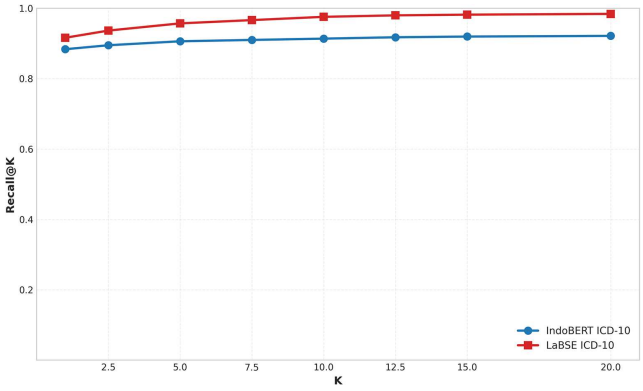


Figure 5
Recall@k comparison between LaBSE and IndoBERT



for both models across different "*k*" values (1, 5, 10, 20), allowing readers to understand how many relevant documents are retrieved within the top "*k*" results. In comparison, IndoBERT showed a comparatively lower performance on the same dataset, highlighting the advantage of our LaBSE-based framework.

The bilingual design of the ICD-10 corpus further contributes to this success. The Master ICD-10 corpus is composed of 10,469 documents with an average length of 6.36 words per document and a maximum length of 29 words. The distribution of document lengths (line graph in the image titled; ICD-10 Document Length Distribution) is heavily weighted toward short texts (5–10 words), which reflects the standardization inherent to ICD-10 code descriptions. This limited length not only maintains computational efficiency in real-time retrieval but also limits semantic noise with the ability for the LaBSE encoder to focus on clinically important terms. As shorter texts produce less noise and maintain better semantic similarity between queries and codes, high Recall@k values are primarily sustained by low document lengths.

This design choice is a strategic compromise between the need for clinical specificity and retrieval practicality, allowing the system to operate in high-fidelity but sufficiently fast response time, which is an essential consideration for deployment in low-resource health care environments. This distribution is also visualized in Figure 6—here the x-axis indicates the number of words per document, while the y-axis denotes the respective number of documents, giving a deeper understanding to readers regarding how frequent different levels of density are. Common terms of the corpus are listed in Figure 7, a word cloud showing that term, whose size is its repetitiveness or bigger meaning inside the corpus.

Beyond numerical improvements, the proposed framework makes three key contributions compared to existing approaches. First, it addresses an underexplored problem of cross-lingual ICD-10 retrieval in low-resource settings, rather than supervised code classification. Second, by leveraging multilingual sentence embeddings with domain-specific bilingual fine-tuning, the framework effectively aligns Indonesian and English medical terminology within a standardized coding system. Third, the integration of attention-based visualization provides transparency at the token level, supporting user understanding without sacrificing retrieval accuracy. The use of a static and short ICD-10 corpus reflects real-world clinical coding practices, where code descriptions are intentionally concise. Accordingly, the framework is optimized for standardized code retrieval rather than patient-specific clinical decision support, with extensions to longer clinical narratives and dynamic patient context left for future work.

To further illustrate the impact of transfer learning, Table 3 presents the performance comparison of the LaBSE model before and after fine-tuning with the Master ICD-10 dataset. This table explicitly shows the quantitative improvement in MRR and Recall@5 metrics, highlighting the benefits of adapting the LaBSE model to the domain-specific Master ICD-10 corpus through transfer learning.

In addition to numerical performance, the system provides visual feedback to enhance user interpretability. For example, a user query such as "demam tifoid" yields highly relevant results, including ICD code A01.0 (typhoid fever/demam tifoid), A75.9 (typhus fever, unspecified/demam tifus, tidak spesifik), and A01.1 (paratyphoid fever A/demam paratifoid), as shown in the ranked output in Figure 8. This figure displays the top-5 retrieved ICD-10 codes, their similarity scores, and bilingual labels (English and Indonesian), allowing users to visually verify the relevance of each result. These outputs reflect both high similarity scores and bilingual textual alignment, which help users validate the relevance of the retrieved codes.

The selection of baseline models was strategically motivated by their complementary strengths and relevance to cross-lingual medical retrieval. IndoBERT serves as the most appropriate monolingual Indonesian transformer baseline, having been

Figure 6
ICD-10 document length distribution (line graph)

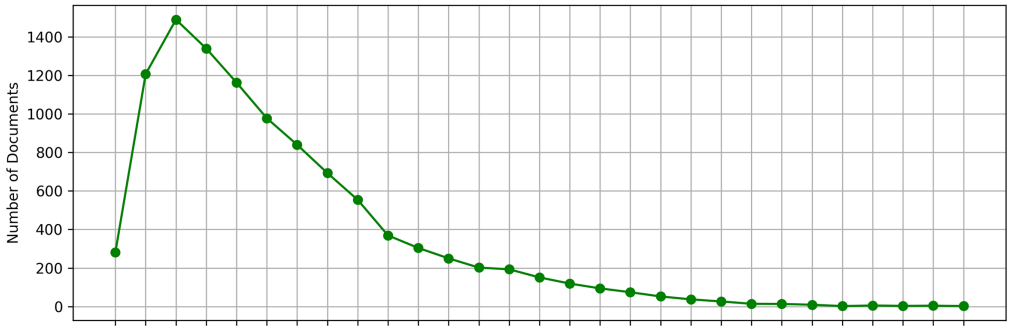


Figure 7
WordCloud ICD-10 Bahasa Indonesia

Table 3
Performance comparison of the LaBSE model before and after transfer learning

Metric	LaBSE (MIRACL-ID)	LaBSE (Master ICD-10)
MRR	0.782	0.935
Recall@5	91.4%	95.73%

Figure 8
Top-5 ICD-10 retrieval results with similarity scores and bilingual labels

Results Top-5 Related Diagnosis Codes:

Rank 1: Score = 0.8066
ICD Code: A01.0
Name (EN): Typhoid fever
Name (ID): Demam tifoid

Rank 2: Score = 0.7580
ICD Code: A75.9
Name (EN): Typhus fever, unspecified
Name (ID): Demam tifus , tidak spesifik

Rank 3: Score = 0.7269
ICD Code: A01.1
Name (EN): Paratyphoid fever A
Name (ID): Demam paratifoid

Rank 4: Score = 0.7151
ICD Code: A01.4
Name (EN): Paratyphoid fever, unspecified
Name (ID): Demam paratifoid, tidak spesifik

Rank 5: Score = 0.7059
ICD Code: A79.0
Name (EN): Trench fever
Name (ID): Demam parit

specifically pretrained on Indonesian text and widely adopted for Indonesian NLP tasks. MedCPT represents the state of the art in biomedical IR, pretrained on large-scale PubMed search logs and demonstrating superior performance in medical domain tasks. These baselines collectively enable evaluation against both language-specific optimization (IndoBERT) and domain-specific pretraining (MedCPT), providing a comprehensive assessment of our cross-lingual approach.

A comprehensive comparative study across all three models on the same dataset is presented in Table 4. This unified evaluation ensures a fair and technically valid comparison, as all models were assessed under identical preprocessing, training, and evaluation conditions.

1) Comparative Evaluation of a Unified Dataset

To establish a fair and technically valid comparison, all models were evaluated on the same Master ICD-10 bilingual dataset under identical preprocessing, training, and evaluation conditions. Table 3 presents a direct quantitative comparison among IndoBERT, MedCPT, and the proposed LaBSE framework. IndoBERT serves as the native Indonesian transformer baseline, while MedCPT represents a strong biomedical IR model pretrained on large-scale PubMed logs. Our proposed LaBSE model, enhanced with an attention-based explainability layer and bilingual fine-tuning, consistently outperforms both baselines across all ranking metrics. Specifically, the proposed model achieves the highest MRR (MRR = 0.9352) and Recall@5 (95.73%), reflecting superior semantic retrieval precision and robustness in multilingual medical contexts. These results confirm that integrating multilingual sentence embeddings with token-level attention substantially improves both the accuracy and interpretability of retrieval in low-resource settings.

Beyond numerical improvements, the experimental results indicate that domain-specific bilingual fine-tuning plays a crucial role in cross-lingual medical retrieval. The proposed framework consistently outperforms strong baselines, suggesting that multilingual sentence embeddings alone are insufficient without

Table 4
Comparative study on the same dataset

Model	Dataset	MRR	Recall@1	Recall@5	Recall@10	Remarks
IndoBERT (ICD-10 baseline)	Master ICD-10	0.8945	0.8838	0.9064	0.9139	Fine-tuned Indonesian transformer baseline
MedCPT (PubMed pretrained, adapted)	Master ICD-10	0.9127	0.9012	0.9408	0.9561	Strong biomedical IR baseline (English-centric)
LaBSE + Attention (proposed)	Master ICD-10	0.9352	0.9164	0.9573	0.9759	Domain-specific bilingual fine-tuning + explainability

adaptation to medical terminology and ICD-10 semantics. This performance gain highlights the importance of aligning bilingual representations within a specialized clinical domain, particularly in low-resource settings.

2) Comparison with Previous Work

Compared to prior studies on explainable medical NLP, which predominantly focus on supervised ICD coding or monolingual clinical text classification, the proposed framework addresses a different and underexplored problem: cross-lingual medical IR in low-resource settings. Existing explainable ICD coding approaches, such as attention-based classification models, primarily aim to predict codes from structured or unilingual clinical narratives, whereas our work focuses on retrieval-oriented ranking of bilingual ICD-10 descriptions. Moreover, while biomedical retrieval models like MedCPT demonstrate strong performance in English-centric corpora, their effectiveness in low-resource and bilingual contexts remains limited without domain-specific adaptation. The superior performance achieved by our LaBSE-based framework on a bilingual Indonesian–English ICD-10 dataset demonstrates that combining multilingual sentence embeddings with attention-based explainability offers a practical and effective alternative for low-resource clinical retrieval tasks.

3) Ablation and Interpretability Analysis

To further validate the impact of the attention-based explainability layer, an ablation experiment was conducted by removing this module while maintaining the same LaBSE encoder and training configuration. As summarized in Table 5, the model without attention achieved lower performance (MRR = 0.9017, Recall@5 = 92.84%) compared to the full proposed model (MRR = 0.9352, Recall@5 = 95.73%). This demonstrates that the attention mechanism not only enhances interpretability but also contributes to improved ranking precision by focusing on clinically relevant terms across languages.

Beyond quantitative improvements, the attention-based visualization significantly strengthens interpretability by providing insight into how individual query and document tokens contribute to the final similarity score. For example, in the bilingual query "demam tifoid," the attention heatmap highlights "demam" (fever) and "tifoid" (typhoid) as dominant terms, precisely aligning with "fever" and "typhoid" in the retrieved ICD-10 codes. Such token-level alignment illustrates that the explainability mechanism supports semantic transparency and clinical trust. This dual improvement (quantitative accuracy and qualitative transparency) addresses one of the most critical limitations of neural retrieval systems: their "black-box" nature.

While attention-based visualization enhances transparency and helps users identify influential query and document tokens, it should not be interpreted as a fully faithful causal explanation of model decisions. In this study, attention is employed as a transparency-enhancing mechanism rather than a definitive explanation. This design choice balances interpretability and performance, addressing the limitations of black-box neural retrieval systems highlighted in prior work. Although clinician-centered user studies represent the gold standard for evaluating explainability in medical systems, such evaluations require ethical approval and clinical collaboration beyond the scope of this work. Human-centered validation of explainability is therefore identified as an important direction for future research.

Collectively, the comparative and ablation results demonstrate that the proposed LaBSE-based framework offers both high retrieval performance and practical explainability, establishing a balanced foundation for trustworthy CLIR in multilingual healthcare environments.

Qualitative inspection of retrieval outcomes suggests that queries explicitly referring to disease entities tend to achieve higher accuracy than symptom-based queries, which are inherently more ambiguous and context-dependent. Due to the scale and structured nature of the evaluation dataset, error analysis in this study is conducted qualitatively. A systematic categorization of query types and a quantitative breakdown of error sources would require more diverse and extensive annotated queries and are therefore left for future work.

4) Explainability Effectiveness

The explainability features are centered on visualizing the importance of tokens in both the query and the retrieved

Table 5
Ablation study on attention-based explainability layer

Model configuration	Attention mechanism	MRR	Recall@5 (%)	Observation
LaBSE (baseline)	✗ Removed	0.9017	92.84	Reduced interpretability and ranking precision
LaBSE + Attention (proposed)	✓ Applied	0.9352	95.73	Improved semantic focus and user explainability

documents. On the query side, attention heatmaps are used to highlight the significance of individual terms, as shown in Figure 9. This heatmap visually represents the token-level attention weights, where warmer colors indicate higher importance, helping users identify which query terms are most salient. The token "demam" receives the highest attention weight, confirming its central role in influencing the retrieval results. This is also reinforced by the query-side word cloud, as shown in Figure 10, where token prominence is visually scaled according to its computed contribution, offering an immediate, intuitive understanding of key terms.

Figure 9
Query token importance heatmap

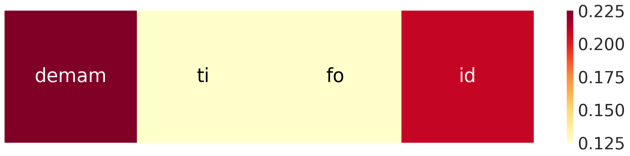


Figure 10
Query token word cloud

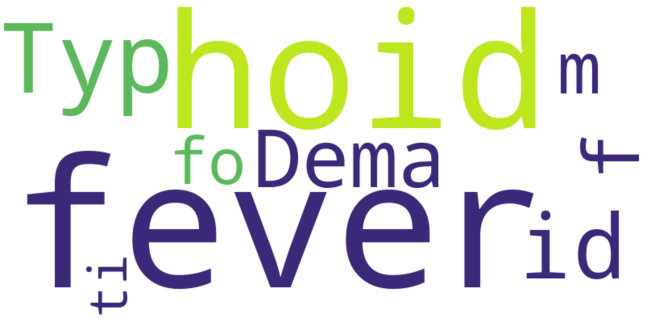
To validate the statistical significance of performance improvements, we conducted paired *t*-tests across the 5-fold cross-validation results. The proposed LaBSE + Attention model significantly outperformed both IndoBERT ($p < 0.001$, 95% CI [0.0408, 0.0552] for MRR difference) and MedCPT ($p < 0.01$, 95% CI [0.0255, 0.0369] for MRR difference). The confidence intervals for our model's performance metrics are MRR = 0.9352 ± 0.0041 and Recall@5 = 95.73% ± 1.24%, indicating consistent and reliable retrieval performance across different data partitions.

In the medical domain, terms often considered stopwords in general text, such as "dengan" (with) or "yang" (which), are critical for precise diagnosis and differentiation of ICD-10 codes. For instance, the presence or absence of "dengan" in medical descriptions can significantly alter the meaning and lead to distinct ICD-10 classifications. Therefore, the attention mechanism's focus on these seemingly common words is not misleading but rather crucial for capturing the nuanced semantic relationships essential for accurate medical code retrieval.

To complement this, document-side attention visualizations provide insight into the semantic alignment between the query and each retrieved ICD code description. Figure 11 presents the document token heatmap, which visualizes the attention intensities on tokens within the retrieved document, revealing which terms received the highest focus from the model during relevance computation. Similarly, Figure 12 offers a document word cloud that further illustrates the dominant biomedical tokens, providing an intuitive overview of the document's relevance context through word size.

Figure 11
Document token importance heatmap

Figure 12
Document token word cloud



These visual tools collectively serve to enhance the interpretability of the retrieval process. Users are able to verify why certain ICD codes are retrieved by tracing the attention flow from query to document. This also supports error analysis, where users can identify which tokens may have caused misalignment or reduced relevance in certain cases. Furthermore, this attention-based transparency directly addresses the common critique that neural models act as black boxes by making visible the inner workings of the matching process at the token level. A comprehensive overview of the retrieval outputs and explainability components is provided in Table 6. This table summarizes the main visual and performance components, detailing their purpose in enhancing both functional robustness and user-centered design.

Through the combination of high retrieval accuracy and interpretable outputs, the proposed system demonstrates both functional robustness and user-centered design. This approach is particularly valuable in clinical environments where trust and clarity are essential for decision support systems.

4.2. Error analysis and limitations

Error analysis indicates three main sources of retrieval failure. First, lexical ambiguity in generic symptom queries such as "demam" (fever) or "demam berkepanjangan" (prolonged fever) often yields overly broad results across ICD-10 chapters. Clinically, "demam berkepanjangan" corresponds to R50.9 (fever), unspecified when the cause is unknown (Chapter XVIII), but to specific disease codes such as A01.0 (typhoid fever), A90 (dengue fever), B05.9 (measles), or B01.9 (varicella) when an etiology is identified. The model sometimes retrieves both symptom-level (R50.9) and disease-level codes simultaneously, reflecting difficulty in distinguishing contextual intent. This issue arises from high lexical overlap among short ICD-10 descriptions and the absence of hierarchical awareness during ranking. Second,

Table 6
Summary of retrieval outputs and explainability components

Component	Description
Retrieval performance	MRR and Recall@k scores demonstrate superior accuracy on the Master ICD-10 tasks
Top-5 ICD-10 results	Retrieved bilingual ICD-10 codes with similarity scores based on the input query, enabling user verification
Query token heatmap	Visualizes token-level attention weights from user query, highlighting influential terms
Query word cloud	Highlights the most influential query terms via attention scoring for intuitive understanding
Document token heatmap	Displays attention on document tokens most related to the query, showing semantic alignment
Document word cloud	Shows dominant biomedical terms in the retrieved document for an intuitive overview of relevance context

code-switching within queries (mixing Indonesian and English terms) reduces precision because the preprocessing pipeline treats each language separately, causing inconsistent tokenization and normalization. Third, rare or low-frequency medical entities achieved a lower Recall@5 of 87.3% compared with the overall 95.7%, indicating reduced representation in the bilingual corpus and embeddings.

Finally, these results emphasize the necessity for targeted cross-lingual retrieval as we move forward. Future iterations could include ICD hierarchy-aware reranking and contextual disambiguation to avoid inter-chapter mismatches, token-level language behavior modeling with synthetic code-switched augmentation for multilingual coverage robustness, few-shot fine-tuning, or hybrid dense-lexical retrieval for rare entity coverage. Nonetheless, the LaBSE + Attention framework we proposed remains robust for common clinical queries (error rate ≈ 4.3%) and exhibits strong generalizability and interpretability across the ICD-10 domain based on the limitations mentioned above.

5. Conclusion

This study introduced a CLIR framework with attention-based, explainable features for retrieving bilingual ICD-10 codes in a low-resourced medical field scenario. With native or experimental results showing that the LaBSE-based model has strong retrieval performance, an MRR of 0.9352 and Recall@5 of 95.73% achieved after domain-specific bilingual fine-tuning. In addition to yielding improvements, the built-in attention structure offered token-level explanations that were immediate as well as enhanced ranking effectiveness. These findings demonstrate the effectiveness of integrated multilingual embeddings and explainability in developing a trustworthy and effective clinical decision support system for multilingual healthcare settings.

Acknowledgment

During the preparation of this work, the authors used OpenAI ChatGPT to improve the work's readability and language. After using this tool/service, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in GitHub at https://github.com/fendis0709/icd-10, and in Hugging Face at https://huggingface.co/datasets/miracl/miracl.

Author Contribution Statement

Nasir Hamzah: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Zico Pratama Putra:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

References

[1] Pachpande, S. D., & Bhalchandra, P. U. (2020). Cross language information retrieval (CLIR): A survey of approaches for exploring web across languages. *International Journal of Innovative Technology and Exploring Engineering*, *10*(1), 326–332. https://doi.org/10.35940/ijitee.K7833.1110120

[2] Zhang, L., & Zhao, X. (2020). An overview of cross-language information retrieval. In *International Conference on Artificial Intelligence and Security*, 26–37. https://doi.org/10.1007/978-3-030-57884-8_3

[3] Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., …, & Ruder, S. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, *1*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

[4] Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., …, & Ruder, S. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 815–834. https://doi.org/10.18653/v1/2023.eacl-main.57

[5] Chang, C. H., & Yang, C. C. (2023). On bridging consumer health search across languages using cross-lingual word space. *Electronic Commerce Research and Applications*, *59*, 101254. https://doi.org/10.1016/j.elerap.2023.101254

[6] Saleh, S., & Pecina, P. (2020). Document translation vs query translation for cross-lingual information retrieval in the medical domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6849–6860. https://doi.org/10.18653/v1/2020.acl-main.613

[7] Buonocore, T. M., Crema, C., Redolfi, A., Bellazzi, R., & Parimbelli, E. (2023). Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, *144*, 104431. https://doi.org/10.1016/j.jbi.2023.104431

[8] Shaitarova, A., Zaghir, J., Lavelli, A., Krauthammer, M., & Rinaldi, F. (2023). Exploring the latest highlights in medical natural language processing across multiple languages: A survey. *Yearbook of Medical Informatics*, *32*(01), 230–243. https://doi.org/10.1055/s-0043-1768726

[9] Abka, A. F., Azizah, K., & Jatmiko, W. (2022). Transformer-based cross-lingual summarization using multilingual word embeddings for English-Bahasa Indonesia. *International Journal of Advanced Computer Science and Applications*, *13*(12), 636–645. https://doi.org/10.14569/IJACSA.2022.0131276

[10] Mustafa, A. M., Nakhleh, S., Irsheidat, R., & Alruosan, R. (2024). Interpreting Arabic transformer models: A study on XAI interpretability for Qur'anic semantic-search models. *Jordanian Journal of Computers and Information Technology*, *10*(4), 350–366. https://doi.org/10.5455/jjcit.71-1704878720

[11] Teng, Q., Liu, Z., Song, Y., Han, K., & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, *28*(6), 2335–2355. https://doi.org/10.1007/s00530-022-00960-4

[12] Yada, S., Nakamura, Y., Wakamiya, S., & Aramaki, E. (2024). Cross-lingual natural language processing on limited annotated case/radiology reports in English and Japanese: Insights from the Real-MedNLP workshop. *Methods of Information in Medicine*, *63*(05/06), 145–163. https://doi.org/10.1055/a-2405-2489

[13] Yang, C., He, B., Li, C., & Xu, J. (2017). A feedback-based approach to utilizing embeddings for clinical decision support. *Data Science and Engineering*, *2*(4), 316–327. https://doi.org/10.1007/s41019-017-0052-2

[14] Hu, S., Teng, F., Huang, L., Yan, J., & Zhang, H. (2021). An explainable CNN approach for medical codes prediction from clinical text. *BMC Medical Informatics and Decision Making*, *21*(Suppl 9), 256. https://doi.org/10.1186/s12911-021-01615-6

[15] Liu, L., Perez-Concha, O., Nguyen, A., Bennett, V., & Jorm, L. (2022). Hierarchical label-wise attention transformer model for explainable ICD coding. *Journal of Biomedical Informatics*, *133*, 104161. https://doi.org/10.1016/j.jbi.2022.104161

[16] Hou, W. H., Wang, X. K., Wang, Y. N., Wang, J. Q., & Xiao, F. (2024). Modelling long medical documents and code associations for explainable automatic ICD coding. *Expert Systems with Applications*, *249*, 123519. https://doi.org/10.1016/j.eswa.2024.123519

[17] Zhao, Q., Kang, Y., Li, J., & Wang, D. (2018). Exploiting the semantic graph for the representation and retrieval of medical documents. *Computers in Biology and Medicine*, *101*, 39–50. https://doi.org/10.1016/j.compbiomed.2018.08.009

[18] Chraibi, A., Delerue, D., Taillard, J., Draa, I. C., Beuscart, R., & Hansske, A. (2021). A deep learning framework for automated ICD-10 coding. In J. Mantas, L. Stoicu-Tivadar, C. Chronaki, A. Hasman, P. Weber, P. Gallos, …, & O. S. Chirila (Eds.), *Public health and informatics* (pp. 347–351). IOS Press. https://doi.org/10.3233/shti210178

[19] Kane, M. J., King, C., Esserman, D., Latham, N. K., Greene, E. J., & Ganz, D. A. (2023). A compressed large language model embedding dataset of ICD 10 CM descriptions. *BMC Bioinformatics*, *24*(1), 482. https://doi.org/10.1186/s12859-023-05597-2

[20] Maree, M., Noor, I., Rabayah, K. S., Belkhatir, M., & Alhashmi, S. M. (2020). Head concepts selection for verbose medical queries expansion. *IEEE Access*, *8*, 93987–93999. https://doi.org/10.1109/ACCESS.2020.2987568

[21] Matthies, F., Beger, C., Schäfermeier, R., Höffner, K., & Uciteli, A. (2024). Extending the TOP framework with an ontology-based text search component. In R. Röhrig, N. Grabe, U. H. Hübner, K. Jung, U. Sax, C. O. Schmidt, ..., & A. Zapf (Eds.), *German medical data sciences 2024*, (pp. 180–189). IOS Press, https://doi.org/10.3233/shti240854

[22] Chen, P. F., He, T. L., Lin, S. C., Chu, Y. C., Kuo, C. T., Lai, F., …, & Yang, C. Y. (2022). Training a deep contextualized language model for international classification of diseases, 10th revision classification via federated learning: Model development and validation study. *JMIR Medical Informatics*, *10*(11), e41342. https://doi.org/10.2196/41342

[23] Nawab, K., Fernbach, M., Atreya, S., Asfandiyar, S., Khan, G., Arora, R., …, & Schreiber, R. (2024). Fine-tuning for accuracy: Evaluation of Generative Pretrained Transformer (GPT) for automatic assignment of International Classification of Disease (ICD) codes to clinical documentation. *Journal of Medical Artificial Intelligence*, *7*, 8. https://doi.org/10.21037/jmai-24-60

[24] Krishna, A., Riedel, S., & Vlachos, A. (2022). Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, *10*, 1013–1030. https://doi.org/10.1162/tacl_a_00503

[25] Yadav, R. K., & Nicolae, D. C. (2022). Enhancing attention's explanation using interpretable tsetlin machine. *Algorithms*, *15*(5), 143. https://doi.org/10.3390/a15050143

[26] Chen, H., Gomez, C., Huang, C. M., & Unberath, M. (2022). Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. *NPJ Digital Medicine*, *5*(1), 156. https://doi.org/10.1038/s41746-022-00699-2

[27] Tucci, V., Saary, J., & Doyle, T. E. (2022). Factors influencing trust in medical artificial intelligence for healthcare professionals: A narrative review. *Journal of Medical Artificial Intelligence*, *5*, 4. https://doi.org/10.21037/jmai-21-25

[28] Chang, Y. H., Guo, Y. T., Fu, L. C., Chiu, M. J., Chiu, H. M., & Lin, H. J. (2023). Interactive healthcare robot using attention-based question-answer retrieval and medical entity extraction models. *IEEE Journal of Biomedical and Health Informatics*, *27*(12), 6039–6050. https://doi.org/10.1109/JBHI.2023.3320939

[29] Teng, F., Ma, Z., Chen, J., Xiao, M., & Huang, L. (2020). Automatic medical code assignment via deep learning approach for intelligent healthcare. *IEEE Journal of Biomedical and Health Informatics*, *24*(9), 2506–2515. https://doi.org/10.1109/JBHI.2020.2996937

[30] Ghasemi, R., & Momtazi, S. (2023). How a deep contextualized representation and attention mechanism justifies explainable cross-lingual sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(11), 1–15. https://doi.org/10.1145/3626094

[31] Upadhyay, R., Knoth, P., Pasi, G., & Viviani, M. (2023). Explainable online health information truthfulness in consumer health search. *Frontiers in Artificial Intelligence*, *6*, 1184851. https://doi.org/10.3389/frai.2023.1184851

[32] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., . . . , & Poon, H. (2025). A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm AI*, *2*(1). https://doi.org/10.1056/AIoa2400640

[33] Yuan, H., Yuan, Z., & Yu, S. (2022). Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4038–4048. https://doi.org/10.18653/v1/2022.naacl-main.296

[34] Rusadi, P., Putra, Z. P., Setyawan, A., Romli, M., Yusuf, M., Pratama, H. A., & Sumarbagiono, R. (2025). Synthetic data for radioactive waste management: A comparative study for disused sealed radioactive sources in Indonesia. *Nuclear Engineering and Technology*, *57*(7), 103524. https://doi.org/10.1016/j.net.2025.103524

[35] Rahman, I. A., Putra, Z. P., Rusadi, P., Irmanti, K. S. D., Setyawan, A., Romli, M., . . . , & Yusuf, M. (2025). Integration of machine learning models for enhancing radioactive waste management of disused sealed radioactive sources. *Nuclear Engineering and Design*, *442*, 114272. https://doi.org/10.1016/j.nucengdes.2025.114272

[36] Pamungkas, N. S., Putra, Z. P., Pratama, H. A., & Yusuf, M. (2025). Supervised machine learning-based categorization and prediction of uranium adsorption capacity on various process parameters. *Journal of Hazardous Materials Advances*, *17*, 100523. https://doi.org/10.1016/j.hazadv.2024.100523

How to Cite: Hamzah, N., & Putra, Z. P. (2026). Attention-Based Explainability for Cross-Lingual Neural Information Retrieval in Low-Resource Languages: A LaBSE Framework with a Bilingual Medical Context Case Study. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62027530