The Application of Deep Learning in Qur'anic Tafsir Retrieval Using SBERT, FAISS and BERT-QA

Asti Herliana[1], Ina Najiyah[2,*], Sari Susanti[3], Lutfhi Muayyad Billah[4]

[1,2,3,4] *Adhirajasa Reswara Sanjaya University, Jl. Sekolah Internasional No 1-2, Bandung 40287, Indonesia*

(Received: June 5, 2025; Revised: August 1, 2025; Accepted: November 5, 2025; Available online: December 19, 2025)

Abstract

Accurate understanding of the Qur'an requires access to reliable tafsir, yet many classical tafsir resources remain non-digital, making search and retrieval time-consuming. This study presents a semantic-based retrieval system for Tafsir Ibn Kathir, covering 114 entries and 6,236 *Verse*s, using SBERT embeddings and FAISS indexing. The system enables users to perform semantic queries, retrieving relevant passages in response to their questions. Evaluation was conducted using 50 representative queries spanning diverse topics, including Fiqh, Aqidah, History, and Spirituality. Relevance judgments were independently provided by three Qur'anic studies experts and reconciled through discussion, with inter-annotator agreement indicating substantial consistency. Each query included 20 non-relevant passages as negative samples to increase evaluation difficulty. Two approaches were tested: retrieval-only and retrieval combined with a zero-shot QA module for span extraction. Retrieval-only achieved slightly higher top-1 accuracy (0.72), but retrieval + QA improved ranking-oriented metrics, including Accuracy@5 (0.88), Mean Reciprocal Rank (MRR = 0.76), and normalized Discounted Cumulative Gain at 5 (nDCG@5 = 0.82), with the increase in Accuracy@5 statistically significant (p = 0.01). The zero-shot QA module enabled the system to extract more precise and contextually relevant information, enhancing overall retrieval quality and robustness. These results indicate that the proposed system effectively retrieves relevant tafsir passages and provides accurate, context-specific answers. The study demonstrates the potential and limitations of zero-shot QA for domain-specific religious texts and supports the development of web-based applications or Islamic chatbots, facilitating easier access to shahih tafsir knowledge for scholars and the broader Muslim community.

Keywords: Tafsir Ibnu Katsir Digital, Retrieval Tafsir with AI, SBERT Cases Tafsir, FAISS Indexing Cases Tafsir, Zero-Shot QA

1. Introduction

The Qur'an is the foremost source of guidance for Muslims, encompassing a wide range of matters including general issues, daily activities, and rules of worship [1]. Revealed by Allah, the Qur'an consists of 114 chapters (*Chapter*s) and 6,236 verses (*Verses*) [2]. Fundamentally, the Qur'an is written in Arabic and has been translated into various languages to facilitate understanding—such as the Indonesian translation for the Muslim community in Indonesia. However, many verses in the Qur'an contain words and expressions with deep and complex meanings that require further interpretation (tafsir) to fully grasp the intended message of Allah [3]. For example, in *Chapter* Al-Ma'un, verse 4, the translation reads: "So woe to those who pray." Without proper interpretation, such a verse may be misunderstood and lead to misperceptions that deviate from the actual meaning intended by Allah [4]. The process of interpreting the Qur'an requires profound knowledge in several disciplines, including the science of Hadith, Usul al-Fiqh (principles of Islamic jurisprudence), and the Arabic language [5]. Consequently, not everyone is qualified to interpret the Qur'an. Scholars who have studied the Qur'an and Hadith extensively and authored works of Qur'anic exegesis (tafsir) have existed since the time of the Prophet's companions (Sahabah), the generation of Tabi'in, and the later renowned Islamic scholars [6]. Among these are prominent figures such as Al-Ṭabari [7], Fakhr al-Dīn al-Rāzī [8], Al-Qurṭubī [9], and Ibn Kathir—whose name is widely recognized in Indonesia as Ibnu Katsir [10].

Tafsir Ibn Kathir is one of the most renowned and frequently referenced works of Qur'anic exegesis, particularly in Indonesia [11], especially among scholars of the Ahlus Sunnah wal Jama'ah tradition. Authored by Ismail bin Umar bin Kathir al-Qurashi al-Dimashqi, Tafsir Ibn Kathir adopts a methodological approach based on interpreting the Qur'an through the Qur'an itself, through authentic Hadith of the Prophet, and through the transmitted opinions of the

*Corresponding author: Ina Najiyah (inajiyah@ars.ac.id)

DOI: https://doi.org/10.47738/jads.v7i1.1000

Companions (Sahabah) and their successors (Tabi'in) [12]. This tafsir is widely regarded as a credible and authoritative source due to its solid methodological foundations, including the use of scriptural evidence rather than personal opinion [13], strong reliance on sound hadith and athar, rejection of weak or fabricated reports, minimal use of excessive philosophical or speculative reasoning, and avoidance of inauthentic narratives from Jews-Christian traditions [14]. In Muslim-majority countries such as Indonesia[15], Qur'anic interpretation holds significant importance and urgency, particularly given the existence of various Islamic schools of thought (madzhab) and sects that may differ in their understanding of religious practices and interpretations of certain verses [16]. To prevent misinterpretation of the meanings and intent of Qur'anic verses, scholars must carefully examine the tafsir of each verse before disseminating religious teachings [17]. Although Tafsir Ibn Kathir is available in print and in some desktop applications, searching for specific interpretations within the book can be time-consuming and labor-intensive. Therefore, digitization is essential to facilitate easier and faster retrieval of Qur'anic tafsir, leveraging advancements in modern technology. The process of tafsir digitization can utilize technologies such as Deep Learning, specifically Neural Network models[18], to assist in understanding, learning, and retrieving interpretation of Qur'anic verses from large datasets [19]. These models can enable the development of user-friendly systems where users can search for interpretations based on specific keywords, thus significantly enhancing accessibility to Islamic knowledge [20].

Although classical Qur'anic exegesis works such as Tafsir Ibn Kathir serve as essential references for Muslims in understanding the meaning of the sacred verses, the manual process of searching for tafsir remains challenging. The complexity of the Arabic language particularly without diacritical marks (harakat), the lengthy and sequential structure of the text by verse, and limited digital accessibility often hinder comprehension, especially for laypeople. Moreover, traditional keyword-based search methods are often inadequate in capturing the semantic context of user queries. This creates a gap between the public's need for easy and contextually relevant access to tafsir and the capabilities of existing technological tools. Therefore, the core problem addressed in this research is how to develop a Deep Learning based Qur'anic tafsir search system that can understand the meaning and context of user queries and deliver relevant, accurate, and easily accessible interpretation results through digital platforms.

2. The Proposed Method

After reviewing the existing problems and consulting several Islamic scholars regarding the expected outcomes, this study adopts an information retrieval approach combined with Deep Learning techniques. The first step involves examining the input dataset, in which the Tafsir Ibn Kathir systematically presents the interpretation of each Qur'anic verse sequentially from *Chapter* 1 to *Chapter* 114. In this tafsir, the Qur'anic verses are written with diacritical marks (harakat), whereas the accompanying Arabic commentary is written in unvocalized text without diacritical marks. Subsequently, the tafsir was translated into Indonesian to align with the research objective of providing accessibility for the Muslim community in Indonesia.

The primary dataset was obtained from Pondok Pesantren Persatuan Islam Bandung in the form of an SQL database containing the complete Tafsir Ibn Kathir, consisting of 6,236 rows with an original structure of nine columns, as shown in table 1. Tafsir Ibnu Katsir Datasets. The dataset is organized into several columns, each serving a specific purpose. The *Chapter* column refers to the name of the Qur'anic chapter where the verse appears, while the *Verse* column specifies the verse number within that *Chapter*. The Arabic column contains the original Arabic text of the Qur'anic verse, and the Translation column provides its Indonesian or English translation. To enrich the interpretation, additional columns are included: Tafsir Bil Qur'an, which explains the verse through cross-references to other Qur'anic passages; Tafsir Bil Hadith, which provides interpretations supported by authentic Prophetic traditions; Tafsir Sahabat, which records the interpretations transmitted from the Prophet's companions; and Tafsir Tabi'in, which presents explanations from the early successors who were students of the companions. Finally, the Fiqh/Khilāf column documents jurisprudential discussions along with differing scholarly opinions, thereby offering a comprehensive view of classical Islamic exegesis.

Table 1. Tafsir Ibnu Katsir Datasets

Element	**Al-Mā'idah (5:6)**	**An-Nisā' (4:43)**	**At-Tawbah (9:103)**	**An-Nūr (24:2)**
Chapter	Al-Mā'idah	An-Nisā'	At-Tawbah	An-Nūr

Verse	6	43	103	2
Arabic	يَا أَيُّهَا الَّذِينَ آمَنُوا إِذَا قُمْتُمْ إِلَى الصَّلَاةِ...	يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَقْرَبُوا الصَّلَاةَ وَأَنْتُمْ سُكَارَى...	خُذْ مِنْ أَمْوَالِهِمْ صَدَقَةً...	الزَّانِيَةُ وَالزَّانِي فَاجْلِدُوا كُلَّ وَاحِدٍ مِنْهُمَا مِائَةَ جَلْدَةٍ...
Translation	O you who have believed, when you intend to pray, wash your face...	O you who have believed, do not approach prayer while you are intoxicated...	Take from their wealth a charity...	The [unmarried] woman or [unmarried] man found guilty of sexual intercourse – lash each one of them with a hundred lashes...
Tafsir Bil Qur'an	Related to QS Al-Baqarah:222 about purification.	Related to the prohibition of *khamr* (QS Al-Baqarah:219).	Linked to QS Al-Baqarah:267 about charity from the best wealth.	Related to QS An-Nur:4 regarding accusations of adultery without proof.
Tafsir Bil Hadith	Hadith from Bukhari–Muslim about the ablution of the Prophet ﷺ.	Hadith stating that *khamr* is the mother of all evils.	"Take from the rich among them and give to the poor among them." (Bukhari)	"Take from me, Allah has prescribed the punishment for them..." (Muslim)
Tafsir of Sahabah (Companions)	Ibn Abbas explained 'مسح الرأس' means part of the head.	Ali ibn Abi Talib forbade praying while intoxicated or extremely drowsy.	Ibn Umar stated that zakat is obligatory upon every capable Muslim.	Umar ibn al-Khattab stated that lashing is the *hadd* punishment for unmarried adulterers.
Tafsir of Tabi'in (Successors)	Mujahid and Hasan al-Basri stated that feet must be washed.	Ikrimah interpreted this verse as a gradual prohibition.	Aṭā' ibn Abi Rabah stated that zakat is invalid if taken from unlawful wealth.	Al-Zuhri stated that *hadd* punishment is not lifted by repentance unless enforced by a judge.
Fiqh Differences (Jurisprudential Disputes)	Differences concerning wiping over *khuff* (leather socks) and order of ablution.	The Hanafi school allows prayer after sobriety even if previously intoxicated.	Disagreements regarding *nisab*, *haul*, and types of zakatable wealth.	Differences among scholars on proof of adultery: four witnesses, confession, or circumstantial evidence.

The textual dataset undergoes preprocessing steps such as tokenization and other procedures [21], followed by indexing of each word. Word embeddings are generated using the Sentence-BERT (SBERT) method, which, according to previous studies, demonstrates superior performance compared to other methods like Word2Vec [22]. After embedding generation, the Deep Learning model is trained to recognize and learn the semantic meaning and contextual sequence of words within the tafsir texts using a Neural Network architecture. Subsequently, the model is trained to understand user queries and provide relevant and accurate answers based on the input questions. Upon achieving optimal accuracy, the proposed model is implemented in a mobile application to facilitate ease of use. The methodology of the proposed approach is illustrated in figure 1.

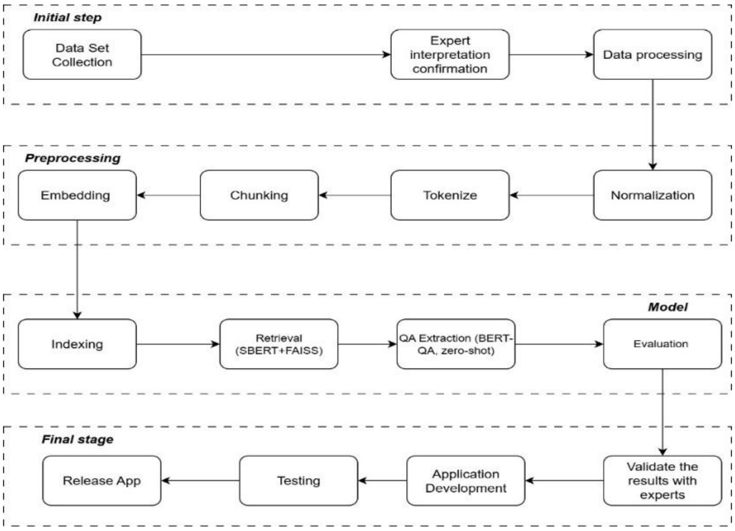


Figure 1. Methodology

2.1. Initial Step

The initial phase of this study involves dataset collection. Valid and reliable data is crucial to ensure accurate and trustworthy results. The data collection process includes gathering Qur'anic verses in both Arabic and Indonesian translation, with the primary focus on the Indonesian translations, hence Arabic text is excluded to reduce complexity and scope limitation. Additionally, data from Tafsir Ibn Kathir is collected. Since Tafsir Ibn Kathir is originally written in Arabic, referencing the corresponding *Chapter* and verse numbers is essential for accurate tafsir translation. The collected dataset is stored in an SQL format to facilitate easier data processing. Following data collection, analysis and preprocessing are conducted in collaboration with tafsir scholars and Arabic language translators to guarantee the validity of the data used. The dataset used in this study is primary data, sourced directly from the book of Ibn Kathir, and has been stored in both SQL and Microsoft Excel formats.

The dataset was subsequently subjected to feature selection, resulting in the use of four columns for the tafsir retrieval process: *Chapter*, *Verse*, the translation column and the tafsir column. All selected columns were standardized into a single language, which in this study is Indonesian. The original verse content in Arabic was excluded from both the training and retrieval processes. however, in this paper, the content is presented in English. Furthermore, the columns Tafsir_Bil_Quran, Tafsir_Bil_Hadits, Tafsir_Sahabat, and Tafsir_Tabiin, as presented in table 1, were merged into a single column to simplify the model construction process. Table 2 illustrates the final dataset for the tafsir retrieval process after undergoing feature selection and feature transformation.

Table 2. Datasets After Feature Selection

Chapter	***Verse***	**Translation**	**Tafsir (Combine)**
al-ma'idah	6	O you who have believed, when you intend to pray, wash your face...	Tafsir bil Qur'an: Related to QS Al-Baqarah:222 on purification. Tafsir bil Hadith: Hadith from Bukhari-Muslim about ablution of the Prophet ﷺ. Tafsir Sahabah: Ibn Abbas explained 'مسح الرأس' means part of the head. Tafsir Tabi'in: Mujahid and Hasan Al-Bashri state feet must be washed. Fiqh Khilaf: Differences regarding wiping over khuf and ablution sequence.
an-nisa	43	O you who have believed, do not approach prayer while intoxicated...	Tafsir bil Qur'an: Related to prohibition of khamr (QS Al-Baqarah:219). Tafsir bil Hadith: Hadith stating khamr is the root of all evil. Tafsir Sahabah: Ali bin Abi Talib forbade praying while intoxicated or very sleepy. Tafsir Tabi'in: Ikrimah interprets gradual prohibition. Fiqh Khilaf: Hanafi school permits ablution for sober persons previously intoxicated after effects subside.
at-tawbah	103	Take from their wealth a charity...	Tafsir bil Qur'an: Linked to QS Al-Baqarah:267 on charity from best wealth. Tafsir bil Hadith: Hadith: "Take from the rich and give to the poor." (HR. Bukhari). Tafsir Sahabah: Ibn Umar: Zakat is obligatory on every capable Muslim. Tafsir Tabi'in: Atha' bin Abi Rabah: Zakat invalid if from unlawful wealth. Fiqh Khilaf: Differences in nisab, haul, and types of zakatable wealth.
an-nur	2	The [unmarried] woman or [unmarried] man guilty of fornication—lash each with a hundred lashes...	Tafsir bil Qur'an: Linked to QS An-Nur:4 on accusation of adultery without proof. Tafsir bil Hadith: Hadith: "Take from me, Allah has prescribed the punishment..." (HR. Muslim). Tafsir Sahabah: Umar bin Khattab: Lashing is hadd punishment for unmarried adulterers. Tafsir Tabi'in: Az-Zuhri: Hadd punishment not lifted by repentance unless enforced by a judge. Fiqh Khilaf: Differences on proof of adultery: four witnesses, confession, or circumstantial evidence (e.g., modern DNA).

2.2. Preprocessing

This process represents a critical stage that requires careful attention to maintain the integrity of the tafsir content and avoid omitting any essential parts of the verses. Consequently, the data will be stored in two distinct versions: the original data as shown in table 1, and the model data used for processing as illustrated in figure 1. The original data will later be used to present the output to users. Preprocessing of tafsir hadith data differs significantly from general text preprocessing because the content of the tafsir must remain unchanged. Therefore, this study applies preprocessing steps including normalization, tokenization, chunking, and indexing. Stopword removal and stemming are not used, as they may alter the meaning of the verses or tafsir. Normalization involves converting all text to lowercase, removing punctuation marks, and cleaning irrelevant special characters. Tokenization then breaks the text into the smallest units, typically words.

For example, the verse "*The woman who commits adultery and the man who commits adultery — flog each one of them a hundred times.*" would be tokenized into 12 tokens. The next step, chunking, is necessary for handling very long verses because BERT-based models, including SBERT, have a maximum input limit of 512 tokens [23]. Without chunking, any text exceeding this limit would be truncated automatically and thus not analyzed by the model [24] For instance, the 12 tokens could be divided into two chunks: tokens 0–6 as chunk 1 and tokens 7–12 as chunk 2. This ensures the entire content is processed, preserves the semantic context within each chunk, and improves embedding efficiency. The preprocessed data is then embedded using Sentence-BERT (SBERT) because this model can convert entire tafsir documents and user queries into vectors [25]. The cosine similarity between these vectors is calculated, and the result with the highest score is selected as the answer. All outputs from these stages are stored in JSON format and serve as inputs for the subsequent semantic-based search system.

2.3. Model

During the model implementation phase, document indexing is performed to assign unique identifiers to each document within the dataset. All chunked outputs from table 1 are converted into embedding vectors and stored in an index using FAISS, accompanied by metadata such as *Chapter*, verse number, source of tafsir, and the original text. These vectorized data are then used in a two-stage pipeline: (1) semantic retrieval using SBERT embeddings with FAISS similarity search, and (2) optional span extraction using a pre-trained BERT-based Question Answering (QA) model without additional fine-tuning. The evaluation phase assesses the effectiveness of this pipeline in retrieving and extracting relevant tafsir passages. The performance is reported using information retrieval metrics such as Accuracy@k, Mean Reciprocal Rank (MRR), and nDCG@k, as well as qualitative expert validation to ensure interpretability and relevance.

2.4. Final Stage

After obtaining the performance results, the application development phase commenced with the creation of system mockups and validation of the tafsir outputs by tafsir experts and Arabic language specialists. The final product is a website application. Prior to its deployment, usability testing was conducted involving 50 students and instructors at Pesantren Persatuan Islam to ensure that the developed system is fit for practical use.

3. Method

The process and architecture of the SBERT model used are illustrated in figure 2. The Sentence-BERT (SBERT) process begins by receiving input text, which may consist of a sentence or a user query. In this study, tokenization is carried out using the XLM-RoBERTa tokenizer, which employs the SentencePiece algorithm rather than the WordPiece approach.

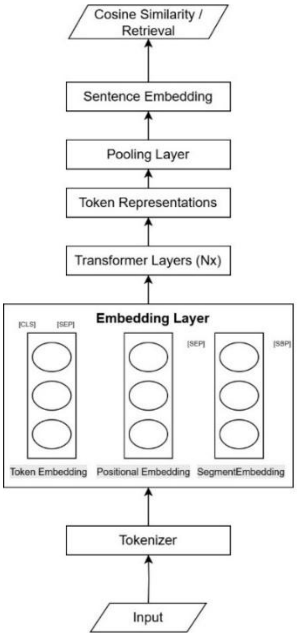


Figure 2. SBERT Process

This method segments the input into subword units without relying on explicit prefix markers such as “##”, thereby producing cleaner and more consistent tokens across languages. Special tokens <s> at the beginning and </s> at the end are automatically added to indicate input boundaries [26]. Importantly, XLM-RoBERTa is pre-trained on a large multilingual corpus covering more than 100 languages, including Arabic [27]. This enables the tokenizer to effectively process morphologically rich Arabic text alongside Indonesian and English translations, ensuring robust handling of the mixed-script dataset used in this research.

The tokenized output is then transformed into Token Embeddings, which are vector representations of each token. To enrich the positional and structural information of the sentence, SBERT incorporates two additional embeddings: Positional Embedding, which encodes the token's order in the sentence, and Segment Embedding, which differentiates between two sentence segments if a pair of sentences is input together. These three embeddings are summed to form the final input to the Transformer network.

The embedding vectors are processed through multiple Transformer encoder layers, which include multi-head self-attention mechanisms and feedforward neural networks. Each layer enables the model to capture comprehensive contextual relationships among tokens, both locally and globally.

After all tokens pass through the encoder layers, SBERT applies a pooling operation to generate a single sentence representation from the token embeddings. The commonly used pooling technique is mean pooling, which computes the average of all token vectors. The output is a fixed-dimensional sentence embedding that represents the overall semantic meaning of the input sentence. This sentence embedding can then be used to compute semantic similarity between the input query and documents in the database (such as Qur'anic tafsir), enabling the system to perform contextual and meaningful information retrieval.

To ensure reproducibility, the model configuration and experimental environment are explicitly reported. The XLM-RoBERTa base checkpoint (xlm-roberta-base, HuggingFace Hub) was used with its official SentencePiece tokenizer (vocabulary size = 250,000). Indexing was performed using FAISS with an IndexFlatIP configuration, cosine similarity as the distance metric, and nprobe = 10. Experiments were run on NVIDIA® GeForce® RTX™ 3050 Laptop GPU 4GB with 32GB RAM.

4. Results and Discussion

4.1. Preprocessing

All columns or fields in table 1 are utilized in the proposed model's processing. Subsequently, data preparation is carried out before the model implementation. Normalization is performed to enhance the model's performance. The results of the normalization process are presented in table 3.

Table 3. Normalization Result

Chapter	*Verse*	**Translation**	**Tafsir (Combine)**
al-ma'idah	6	o you who have believed, when you intend to pray, wash your face...	tafsir bil qur'an: related to qs al-baqarah:222 on purification. tafsir bil hadith: hadith from bukhari-muslim about ablution of the prophet ﷺ. tafsir sahabah: ibn abbas explained 'مسح الرأس' means part of the head. tafsir tabi'in: mujahid and hasan al-bashri state feet must be washed. fiqh khilaf: differences regarding wiping over khuf and ablution sequence.
an-nisa	43	o you who have believed, do not approach prayer while intoxicated...	tafsir bil qur'an: related to prohibition of khamr (qs al-baqarah:219). tafsir bil hadith: hadith stating khamr is the root of all evil. tafsir sahabah: ali bin abi talib forbade praying while intoxicated or very sleepy. tafsir tabi'in: ikrimah interprets gradual prohibition. fiqh khilaf: hanafi school permits ablution for sober persons previously intoxicated after effects subside.
at-tawbah	103	take from their wealth a charity...	tafsir bil qur'an: linked to qs al-baqarah:267 on charity from best wealth. tafsir bil hadith: hadith: "take from the rich and give to the poor." (hr. bukhari). tafsir sahabah: ibn umar: zakat is obligatory on every capable muslim. tafsir tabi'in: atha' bin abi rabah: zakat invalid if from unlawful wealth. fiqh khilaf: differences in nisab, haul, and types of zakatable wealth.
an-nur	2	the [unmarried] woman or [unmarried] man guilty of fornication—lash each with a hundred lashes...	tafsir bil qur'an: linked to qs an-nur:4 on accusation of adultery without proof. tafsir bil hadith: hadith: "take from me, allah has prescribed the punishment..." (hr. muslim). tafsir sahabah: umar bin khattab: lashing is hadd punishment for unmarried adulterers. tafsir tabi'in: az-zuhri: hadd punishment not lifted by repentance unless enforced by a judge. fiqh khilaf: differences on proof of adultery: four witnesses, confession, or circumstantial evidence (e.g., modern dna).

Considering the difference in the number of columns between table 1 and table 2, several columns have been merged into a single "tafsir" column (combining Tafsir_Bil_Quran, Tafsir_Bil_Hadits, Tafsir_Sahabah, Tafsir_Tabiin, and Fiqih_Khilaf). This approach is due to the way the SBERT and FAISS models operate on text units per document, where SBERT converts one entire text string into a single vector. FAISS then compares these vectors based on their overall semantic meaning, rather than on individual columns. Therefore, the more comprehensive the information contained in each vector, the better the retrieval performance. Moreover, users typically seek the tafsir of an entire verse, not segmented by tafsir dimension. For example, a user query such as "What is the tafsir of QS Al-Ma'idah verse 6 regarding ablution and fiqh?" requires an answer that integrates all perspectives (Qur'an, Hadith, Sahabah, Tabi'in, and Fiqh). If these columns are separated, SBERT/FAISS would only have fragmented pieces of information, resulting in disjointed and potentially irrelevant search results. Additionally, merging columns enhances indexing efficiency.

The SentencePiece tokenizer used in XLM-RoBERTa differs from standard tokenization, which typically splits words based on spaces. Instead of relying on whitespace, SentencePiece applies a data-driven subword segmentation approach that encodes text into smaller, statistically learned units (e.g., “bermain” → [“▁ber”, “main”]). The first step is to determine the number of tokens in each column. Table 3 presents the token counts and the counts after chunking based on table 2, while the tokenization results are shown in table 4 and the chunking results in table 5. This study focuses only on the “translation” and “tafsir” columns for tokenization, as the “*Verse*” and “*Chapter*” columns do not require tokenization.

Table 4. Number of Tokens in the Dataset

Chapter	*Verse*	Number of Token in The Translations Colum	Number of Token in The Tafsir Colum	Number of Chunks in the Translation Column	Number of Chunks in the Tafsir Column
al-ma'idah	6	28	124	1	1
an-nisa	43	26	133	1	1
at-tawbah	103	12	136	1	1
an-nur	2	23	153	1	1

To prepare the dataset for downstream semantic similarity and question answering tasks, tokenization was performed using the XLM-RoBERTa base tokenizer, which supports multilingual corpora including Arabic, Indonesian, and English. Each column (Arabic verse, translation, and tafsir) was segmented into subword units through SentencePiece. For every entry, three aspects were recorded: (i) the total number of tokens, (ii) the number of unknown tokens ([UNK]), and (iii) the first 15 tokens. The tokenization logs indicate that the [UNK] rate was negligible, confirming that the chosen tokenizer adequately handles the mixed Arabic–Latin script present in the dataset. The tokenized sequences were subsequently divided into chunks suitable for SBERT embedding. Since SBERT accepts a maximum of 512 tokens, the chunk length was set to 510 tokens to accommodate the special [CLS] and [SEP] tokens automatically added during embedding generation. As an example, in table 1, the ‘Translation’ and ‘Tafsir’ columns of the first row contained 28 tokens (Translation_wordpiece_token_count = 28). Because 28 is smaller than 510, the entire sequence fits into a single chunk, resulting in only one chunk for that entry. Thus, the number of chunks for the 'Translation' column is one.

Table 5. Tokenization and Chunking Results

Tokenization Results for the Translation Column	Tokenization and Chunking Results for the Tafsir Column
o,you,who,have,believed,,,when,you,intend,to,perform,prayer,,,then,wash	tafsir,bil-qur'an,:,linked,to,qs,al,-,baqarah,:,222,regarding,purification,.,tafsir,bil-hadith,:,hadith,bukhari,-,muslim,regarding,ablution,of,the,prophet,.,tafsir,sahabah,:,ibn,abbas,explained,“,مسح,الرأس,”,means,part,of,the,head,.,tafsir,tabi'in,:,mujahid,and,hasan,al,-,bashri,mentioned,that,the,feet,must,be,washed,.,fiqh,khilaf,:,differences,regarding,wiping,over,khuf,and,the,sequence,of,ablution
o,you,who,have,believed,,,do,not,approach,prayer,while,intoxicated	tafsir,bil-qur'an,:,linked,to,qs,al,-,baqarah,:,222,regarding,purification,.,tafsir,bil-hadith,:,hadith,bukhari,-,muslim,regarding,ablution,of,the,prophet,.,tafsir,sahabah,:,ibn,abbas,explained,“,مسح,الرأس,”,means,part,of,the,head,.,tafsir,tabi'in,:,mujahid,and,hasan,al,-,bashri,mentioned,that,the,feet,must,be,washed,.,fiqh,khilaf,:,differences,regarding,wiping,over,khuf,and,the,sequence,of,ablution

take,zakat,from,a,portion,of,their,wealth	tafsir,bil-qur'an,:,linked,to,qs,al,-,baqarah,:,267,regarding,charity,from,the,best,wealth,.,tafsir,bil-hadith,:,hadith,:,take,from,the,wealthy,and,distribute,to,the,poor,.,tafsir,sahabah,:,ibn,umar,:,zakat,is,obligatory,on,every,muslim,.,tafsir,tabi'in,:,atha',bin,abi,rabah,:,zakat,is,invalid,from,unlawful,wealth,.,fiqh,khilaf,:,differences,in,nisab,,,haul,,,and,types,of,wealth,that,are,subject,to,zakat
women,who,commit,adultery,and,men,who,commit,adultery,,,flog,each,of,them,a,hundred,times	tafsir,bil-qur'an,:,linked,to,qs,an,-,nur,:,4,regarding,accusation,of,adultery,without,proof,.,tafsir,bil-hadith,:,hadith,:,allah,has,prescribed,punishment,for,them,.,tafsir,sahabah,:,umar,bin,khattab,:,flogging,is,the,hadd,punishment,for,unmarried,adulterers,.,tafsir,tabi'in,:,az,-,zuhri,:,hadd,is,not,lifted,by,repentance,except,enforced,by,a,judge,.,fiqh,khilaf,:,differences,among,scholars,regarding,proof,of,adultery,:,four,witnesses,or,confession,,,or,with,circumstantial,evidence,(modern,dna,methods)

The embedding process using SBERT utilizes the paraphrase-multilingual-MiniLM-L12-v2 model implemented in Python technology[28]. This model is specifically designed to convert sentences or paragraphs into numerical vectors (embeddings) that capture the semantic meaning of the text. Being "multilingual" means it can work with multiple languages, including Indonesian. A function was created to process the list of token chunks, which accepts a list of token chunks, the SBERT model, and the previously used XLM-RoBERTa base tokenizer.

Since the SentenceTransformer (SBERT) model generally accepts input in the form of text strings (not token lists), the function first takes each list of tokens within a chunk and converts it back into a text string by merging the split tokens (##) into their original words. Once the chunk tokens are converted back into text strings, the model is used to generate embeddings for each of those text strings. The encode method takes the text string and outputs it as a numerical vector (in this model's case, the vector has 384 dimensions). Each text chunk will have one embedding vector. The embedding results for each chunk (in NumPy array form) are then converted into Python lists (list(embeddings)) to be more easily stored in a single cell of a DataFrame

4.2. Model Implementation

The indexing process using FAISS is performed by adding an ID column to each row in the dataset, followed by computing the similarity scores across the entire dataset. Table 6 is an example of the indexing result.

Table 6. Indexing Result

Rank	Chapter	Verse	Score
1	Al-Ma'idah	6	0.9124
2	An-Nisa	43	0.8968
3	At-Tawbah	103	0.8731
4	An-Nur	2	0.858
5	Al-Baqarah	222	0.8275

The process requires the entire dataset for computing similarity scores, as FAISS indexing can only be performed when there is more than one embedding [29]. A single vector is insufficient to build an index or conduct semantic (similarity) search. The indexing results can be seen in the 'Score' column, with the following steps involved: (1) store the embedding vectors in an indexing structure such as FAISS or Annoy. (2) Add an ID or metadata to identify which document each vector corresponds to. (3) When a search is performed, the query is also converted into a vector, and the most similar vectors in the index are retrieved.

The QA component in this study is executed in a zero-shot setting (no fine-tuning on tafsir QA data) due to the lack of publicly available, well-annotated QA datasets specifically tailored to Qur'anic tafsir in Arabic/Indonesian. The zero-shot approach is adopted to evaluate cross-lingual and cross-domain generalization of pre-trained multilingual models [30]. To rigorously assess the reliability of zero-shot QA, an ablation comparison is performed between (1) retrieval-only (selecting answers from top-k retrieved passages without span extraction) and (2) retrieval + QA (span extraction using the zero-shot QA model). This comparison determines whether the QA module provides incremental benefit over retrieval alone and highlights limitations of zero-shot QA on domain-specific religious texts. This process allows the system not only to retrieve the most relevant documents, but also to extract precise and specific information, thereby enhancing the accuracy of the semantic-based question answering system. Table 7 illustrates the performance using retrieval alone and adding QA+Retrieval

Table 7. Index Retrieval metrics

Condition	**Accuracy@1**	**Accuracy@5**	**MRR**	**nDCG@5**	**95% CI (Accuracy@5)**
Retrieval-only	0.72	0.84	0.74	0.80	[0.81–0.87]
Retrieval + QA	0.69	0.88	0.76	0.82	[0.85–0.90]
Difference (R+QA − R)	−0.03	+0.04	+0.02	+0.02	p = 0.01 (paired bootstrap)

The ablation study reveals that while the retrieval-only condition achieved slightly higher top-1 accuracy (0.72 vs. 0.69), the integration of a QA stage yielded consistent improvements on ranking-oriented metrics, including Accuracy@5 (+0.04), MRR (+0.02), and nDCG@5 (+0.02). Notably, the increase in Accuracy@5 was statistically significant (p = 0.01, paired bootstrap), with the confidence interval shifting from [0.81–0.87] to [0.85–0.90]. This suggests that although QA may not always enhance the first retrieved item, it contributes to better overall ranking quality and robustness when multiple candidates are considered. Consequently, the retrieval + QA pipeline provides more reliable performance in realistic settings where top-k results are inspected, whereas retrieval-only remains competitive for strict top-1 use cases.

The evaluation phase was designed to systematically measure the effectiveness of the tafsir retrieval and QA system. Given that the system leverages a semantic similarity approach based on SBERT embeddings and FAISS indexing, the assessment emphasized the system's ability to return relevant and contextually appropriate passages in response to user queries. The evaluation employed standard information retrieval ranking metrics, specifically Accuracy@k, Mean Reciprocal Rank (MRR), and normalized Discounted Cumulative Gain (nDCG@k). These metrics provide a comprehensive view of both the correctness of retrieved results and the overall ranking quality.

To ensure reliability and reproducibility, a total of 50 representative queries was constructed, covering a diverse range of topics in Qur'anic exegesis. Relevance judgments (gold labels) were provided independently by three subject matter experts in Qur'anic studies. Disagreements were resolved through discussion, and inter-annotator agreement was quantified using Cohen's Kappa (κ), indicating substantial agreement. For illustration purposes, table 8 presents 5 sample queries, while the complete set of 50 queries was used during the evaluation process.

Table 8. Index Retrieval metrics

No	**Query (User Question)**	**Category/Theme**
1	What is the ruling on touching the mushaf without ablution (wudhu)?	Fiqh (Purification)
2	How is the verse on tayammum interpreted in the context of illness?	Fiqh (Tayammum)
3	Who is meant by "ahl al-kitab" in QS Al-Baqarah:62?	Aqidah / Interfaith
4	Why did the qibla direction change from Bayt al-Maqdis to the Ka'bah?	History (Qibla Change)
5	What is the meaning of "qalbun salim" in QS Ash-Shu'ara:89 according to classical tafsir?	Tasawwuf / Spirituality

For each query, a total of 20 non-relevant passages were included as negative samples, randomly drawn from the same *Chapter* to increase difficulty. For illustration purposes, table 9 presents only 2 examples of such negative samples, while the complete set of 20 was used in the actual evaluation process. The final evaluation dataset thus consisted of curated user queries with corresponding relevant and non-relevant passages. The system's retrieval outputs were compared against the gold labels to generate the evaluation scores.

Table 9. Index Retrieval metrics

Query	Relevant Passage (Gold)	Non-Relevant Samples (Randomly Selected from the Same Chapter)
What are the requirements for a valid ablution (wudhu) according to the Qur’an?	Tafsir Al-Ma’idah:6 on wudhu & tayammum	- Tafsir Al-Ma’idah:3 (law on lawful and unlawful food) - Tafsir Al-Ma’idah:38 (punishment of hand-cutting for thieves) - Tafsir Al-Ma’idah:45 (law of retribution/qisas)
What are the scholarly differences of opinion regarding washing the feet in wudhu?	Tafsir Al-Ma’idah:6 (juristic disagreement on washing/wiping the feet)	- Tafsir Al-Ma’idah:2 (mutual assistance in righteousness) - Tafsir Al-Ma’idah:33 (punishment for rebels) - Tafsir Al-Ma’idah:51 (prohibition of taking Jews/Christians as allies)

4.3. Final Stage

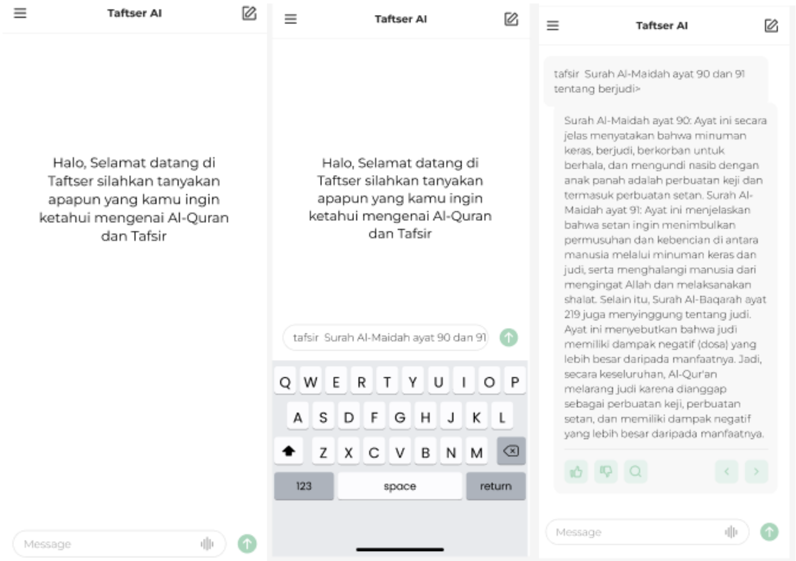
Overall, the evaluation results indicate that the integration of SBERT and FAISS demonstrates strong performance in semantic-based tafsir retrieval tasks and holds substantial potential for deployment in both mobile and web-based application environments. To further validate the usability aspect, a user study was conducted with 50 participants. Each participant was asked to complete a set of representative search tasks, such as “retrieve the tafsir on tayammum in the Qur’an” or “find the meaning of qalbun salim in QS Asy-Syu’ara:89.”

The quantitative results are presented in table 10, which reports usability metrics including the SUS score, task success rate, average completion time, error rate, and qualitative findings. The obtained SUS score was 78.5 (SD = 6.2), indicating good usability as it exceeds the standard threshold of 68. The task success rate reached 92%, with an average completion time of 42 seconds per query.

Table 10. Usability Metrics

Metric	Result (Mean ± SD)	Notes
SUS Score	78.5 ± 6.2	Acceptable usability (above 68)
Task Completion Rate	92%	46/50 participants completed all tasks
Average Task Time	42 seconds per query	Measured across 5 representative tasks
Error Rate	8%	Mostly due to ambiguous queries
Qualitative Feedback Themes	+ UI simple & fast - Needs better error handling	From post-test interviews

The system was built using a PostgreSQL database and implemented with FastAPI in Python. The user interface is shown in figure 3.



Halo, Selamat datang di Taftser silahkan tanyakan apapun yang kamu ingin ketahui mengenai Al-Quran dan Tafsir

Halo, Selamat datang di Taftser silahkan tanyakan apapun yang kamu ingin ketahui mengenai Al-Quran dan Tafsir

tafsir Surah Al-Maidah ayat 90 dan 91

tafsir Surah Al-Maidah ayat 90 dan 91 tentang berjudi>

Surah Al-Maidah ayat 90: Ayat ini secara jelas menyatakan bahwa minuman keras, berjudi, berkorban untuk berhala, dan mengundi nasib dengan anak panah adalah perbuatan keji dan termasuk perbuatan setan. Surah Al-Maidah ayat 91: Ayat ini menjelaskan bahwa setan ingin menimbulkan permusuhan dan kebencian di antara manusia melalui minuman keras dan judi, serta menghalangi manusia dari mengingat Allah dan melaksanakan shalat. Selain itu, Surah Al-Baqarah ayat 219 juga menyinggung tentang judi. Ayat ini menyebutkan bahwa judi memiliki dampak negatif (dosa) yang lebih besar daripada manfaatnya. Jadi, secara keseluruhan, Al-Qur'an melarang judi karena dianggap sebagai perbuatan keji, perbuatan setan, dan memiliki dampak negatif yang lebih besar daripada manfaatnya.

Figure 3. User Interface

5. Conclusion

This study presents a semantic-based Qur'anic tafsir retrieval and question-answering system leveraging SBERT embeddings and FAISS indexing. The system efficiently indexes tafsir passages and retrieves relevant content in response to user queries, enabling precise information extraction even in a zero-shot QA setting. The evaluation demonstrates that while retrieval-only achieves slightly higher top-1 accuracy, the integration of a QA component improves overall ranking metrics such as Accuracy@5, MRR, and nDCG@5, with the increase in Accuracy@5 being statistically significant.

These results indicate that the retrieval + QA pipeline offers more robust performance when multiple candidate passages are considered, supporting realistic user scenarios. User studies further confirm the system's usability, with a SUS score of 78.5, a 92% task completion rate, and an average query completion time of 42 seconds. Participants highlighted the system's simplicity and responsiveness, though some improvements in error handling are suggested. Overall, the proposed framework demonstrates strong potential for deployment in mobile and web applications, providing reliable access to Qur'anic exegesis in both Arabic and Indonesian. Future work may explore fine-tuning QA models on domain-specific datasets and expanding coverage to additional tafsir sources to enhance retrieval precision and answer accuracy.

6. Declarations

6.1. Author Contributions

Conceptualization, A.H., I.N., and S.S.; Methodology, A.H. and L.M.B.; Software, S.S. and I.N.; Validation, I.N. and L.M.B.; Formal Analysis, A.H.; Investigation, S.S. and I.N.; Resources, I.N. and L.M.B.; Data Curation, S.S.; Writing Original Draft Preparation, A.H.; Writing Review and Editing, L.M.B. and I.N.; Visualization, I.N. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This research is a mandatory outcome of the basic research grant program under the fundamental research scheme for funding in 2025. The study received full funding from the Directorate of Research, Technology, and Community Service (DRTPM) of the Ministry of Higher Education, Science, and Technology (KEMENDIKTISAINTEK) of the Republic of Indonesia. The authors also express their sincere gratitude to the Persatuan Islam Union in Bandung for their assistance in the dataset collection process.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. M. Abdulrahman and A. H. M. Walusimbi, "The methodology of Al-Sabuni in interpreting legal Quranic verses: A critical examination of Rawai' Al-Bayan," *Burhān Journal of Qurʾān and Sunnah Studies*, vol. 9, no. 2, pp. 46–68, Aug. 2025, doi: 10.31436/alburhn.v9i2.373.

[2] H. A. D. Khazaleh, A. A. Sapar, and J. Mohd Jan, "A pragmatic analysis of the speech act of supplication in the Holy Quran," *Al-Dad Journal*, vol. 7, no. 1, pp. 40–53, Jul. 2023, doi: 10.22452/aldad.vol7no1.3.

[3] A. K. Qaramaleki, "Applications of intellect as a source in the interpretation of the Qur'an with emphasis on its scope," *Journal of Contemporary Islamic Studies (JCIS)*, vol. 6, no. 1, pp. 73–83, 2024, doi: 10.22059/jcis.2023.354380.1315.

[4] S. N. 'A. A. Wahid and W. S. W. Abdullah, "On some misconceptions concerning the meaning and nature of ʿUlamāʾ," *Journal of Usuluddin*, vol. 50, no. 1, pp. 145–175, Jun. 2022, doi: 10.22452/usuluddin.vol50no1.7.

[5] H. F. S. Faizi and H. S. Ali, "The core principles of Islamic jurisprudence within legal theory: A comprehensive analysis," *Online Journal of Research in Islamic Studies*, vol. 11, no. 2, pp. 57–72, 2024, doi: 10.22452/ris.vol11no2.4.

[6] R. Alizadeh, "Reviewing and criticizing Averroes's viewpoint on granting the people of reasoning (Ahl al-Burhān) the right to interpret the Quran esoterically," *Religious Inquiries*, vol. 13, no. 1, pp. 23–40, Jun. 2024, doi: 10.22034/ri.2024.360400.1643.

[7] A. Jalil, "Naqd al-Qirāʾāt ʿinda al-Mufassirin: Dirāsat Muqāranat linaqd al-Ṭabari wa al-Ṭūsi liriwāyati Ḥafṣ ʿan ʿĀṣim," *Jurnal Studi Ilmu-ilmu Al-Qur'an dan Hadis*, vol. 23, no. 1, pp. 19–48, Jan. 2022, doi: 10.14421/qh.2022.2301-02.

[8] M. Lutfianto, "Male leadership (husband) in the household in Surah An-Nisa, verse 34 (comparative study of Tafsir Mafatih Al-Ghayb by Fakhr Al-Din Al-Razi and Tafsīr Al-Manār by Muhammad Abduh)," *Journal International Dakwah and Communication,* vol. 1, no. 1, pp. 1-33, 2021, doi: 10.55849/judastaipa.v1i1.71.

[9] A. Rohman, B. Mubaroka, and Q. Butlam, "Methodology of Tafseer Al-Qurtubi: Sources, styles and manhaj," *QiST: Journal of Quran and Tafseer Studies*, vol. 2, no. 2, pp. 180–202, Mar. 2023, doi: 10.23917/qist.v2i2.1451.

[10] B. Al Ghoni, A. Azizah, A. Nurrohim, Y. Dahliana, and A. Nirwana, "Comparative study of the criteria of 'Ibadurrahman Tafsir Ibn Katsir and Tafsir Al-Misbah," *Jurnal Asy-Syukriyyah*, vol. 25, no. 2, pp. 227–243, Dec. 2024, doi: 10.36769/asy.v25i2.616.

[11] M. Nasrullah, F. Indarti, and A. Ghufrani, "Family resilience strategies in the contemporary era: A comparative analysis of Al-Azhar and Ibn Katsir's Tafsir on Surah At-Tahrim verse 6," *Al-Karim: International Journal of Quranic and Islamic* Studies, vol. 2, no. 2, pp. 116–137, Aug. 2024, doi: 10.33367/al-karim.v2i2.6032.

[12] R. N. Ashfiy, "New direction of the Qur'an interpretation in Indonesia: A study of Nadirsyah Hosen's interpretation on social media," *Al-Karim: International Journal of Quranic and Islamic Studies*, vol. 2, no. 2, pp. 165–184, Aug. 2024, doi: 10.33367/al-karim.v2i2.5254.

[13] A. Z. A. Ridho, "From tafsir to tadabbur: A preliminary survey of Saudi scholars' trends in understanding the Qur'an," *Tanzil: Jurnal Studi Al-Quran*, vol. 7, no. 1, pp. 103–120, Oct. 2024, doi: 10.20871/tjsq.v7i1.370.

[14] H. Ruhul Jihad, M. Shobahiya, M. Nur, and R. Maksum, "Verses of prohibition for children: Perspectives from Tafsir Ibn Kathir and Tafsir Al-Azhar and their relevance to educational psychology," *in Proceeding ISETH (International Summit on Science, Technology, and Humanity), Surakarta: Universitas Muhammadiyah Surakarta*, vol. 2023, no. Jan., pp. 2935–2942, 2023, doi: 10.23917/iseth.5449.

[15] S. S. Karimullah, "The implications of Islamic law on the rights of religious minorities in Muslim-majority countries," *MILRev: Metro Islamic Law Review*, vol. 2, no. 2, pp. 90–114, Nov. 2023, doi: 10.32332/milrev.v2i2.7847.

[16] S. A. Munandar and S. Amin, “Contemporary interpretation of religious moderation in the Qur’an: Thought analysis of Quraish Shihab and its relevance in the Indonesian context,” *QiST: Journal of Quran and Tafseer Studies*, vol. 2, no. 3, pp. 290–309, Aug. 2023, doi: 10.23917/qist.v2i3.1448.

[17] D. Rahmawati, M. S. Abouzeid, A. Nirwana, S. Hidayat, and A. Rhain, “Investigating online Quran interpretation: Methods and sources on Muslimah.or.id and its contribution to online Islamic discourse,” *QiST: Journal of Quran and Tafseer Studies*, vol. 4, no. 1, pp. 75–90, Jan. 2025, doi: 10.23917/qist.v4i1.7322.

[18] S. Susanti, I. Najiyah, Y. Ramdhani, A. Herliana, M. K. Muckti, and F. R. Oktaviani, “Searching sahih hadiths based on queries using neural models and FastText,” *Journal of Applied Data Sciences,* vol. 6, no. 1, pp. 272–285, 2025, doi: 10.47738/jads.v6i1.567.

[19] M. F. Fasyani, A. Muqsid, and A. W. Mujtaba, “Artificial intelligence and digital tafsīr: Assessing the interpretive accuracy of ChatGPT’s engagement with Tafsīr al-Qurṭūbī,” *Journal of Ushuluddin and Islamic Thought,* vol. 2, no. 1, pp. 86–118, Jun. 2024, doi: 10.15642/juit.2024.2.1.86-118.

[20] M. N. Omar and A. A. Sa’ad, “Exploring the potential of artificial neural network in Sharīʿah decision-making for digital banking: A literature review,” *TAFHIM: IKIM Journal of Islam and the Contemporary World*, vol. 17, no. 2, pp. 105–130, Nov. 2024, doi: 10.56389/tafhim.vol17no2.5.

[21] S. Gupta et al., “Aspect based feature extraction in sentiment analysis using Bi-GRU-LSTM model,” *Journal of Mobile Multimedia*, vol. 20, no. 4, pp. 935–960, Oct. 2024, doi: 10.13052/jmm1550-4646.2048.

[22] A. Riaz, O. Abdulkader, M. J. Ikram, and S. Jan, “Exploring topic modelling: A comparative analysis of traditional and transformer-based approaches with emphasis on coherence and diversity,” *International Journal of Electrical and Computer Engineering (IJECE),* vol. 15, no. 2, pp. 1933–1948, Apr. 2025, doi: 10.11591/ijece.v15i2.pp1933-1948.

[23] D. Zaikis and I. Vlahavas, “From pre-training to meta-learning: A journey in low-resource-language representation learning,” IEEE Access, vol. 11, pp. 115951–115967, 2023, doi: 10.1109/ACCESS.2023.3326337.

[24] Y. Wu, W. Qiu, M. Zeng, X. Chen, M. Li, and H. Zhu, “GoM-ICD: Automatic ICD coding with gap schemes and mixture of experts,” *Big Data Mining and Analytics*, vol. 8, no. 6, pp. 1211–1224, Dec. 2025, doi: 10.26599/BDMA.2025.9020019.

[25] T. Afzal, S. A. Rauf, and Q. Majid, “Semantic similarity of the Holy Quran translations with Sentence-BERT,” *in Proc. Int. Bhurban Conf. on Applied Sciences and Technology (IBCAST), Murree: IEEE*, vol. 2023, no. Aug., pp. 1211–1220, 2023, doi: 10.1109/IBCAST59916.2023.10712955.

[26] A. Erkan and T. Gungor, “Analysis of deep learning model combinations and tokenization approaches in sentiment classification,” *IEEE Access*, vol. 11, no. 1, pp. 134951–134968, 2023, doi: 10.1109/ACCESS.2023.3337354.

[27] Md. A. Mia, F. S. Tamim, Z. S. Taheri, and Md. A. Talukder, “Enhanced semantic relatedness prediction using XLM-RoBERTa and CNNs with K-fold cross-validation,” *The Journal of Engineering*, vol. 2025, no. 1, pp. 1-12, Jan. 2025, doi: 10.1049/tje2.70106.

[28] Md. S. Uddin, R. H. Rifat, M. Kamal, K. D. Gupta, R. George, and M. A. Haque, “Bangla SBERT – Sentence embedding using multilingual knowledge distillation,” *in Proc. Annu. Ubiquitous Computing, Electronics & Mobile Communication Conf. (UEMCON), Yorktown Heights: IEEE*, vol. 2024, no. Nov., pp. 112–121, 2024, doi: 10.1109/UEMCON62879.2024.10754765.

[29] G. P. Rusum and S. Anasuri, “Vector databases in modern applications: Real-time search, recommendations, and retrieval-augmented generation (RAG),” *BigData, Computational and Management Studies*, vol. 5, no. 4, pp. 124–136, 2024, doi: 10.63282/3050-9416.IJAIBDCMS-V5I4P113.

[30] M. Hardalov, A. Arora, P. Nakov, and I. Augenstein, “Few-shot cross-lingual stance detection with sentiment-based pre-training,” 2022. [Online]. Available: http://github.com/meetDeveloper/freeDictionaryAPI