An Ensemble Semantic Text Representation with Ontology and Query Expansion for Enhanced Indonesian Quranic Information Retrieval

Liza Trisnawati[1], Noor Azah Binti Samsudin[2], Shamsul Kamal Bin Ahmad Khalid[3], Ezak Fadzrin Bin Ahmad Shaubari[4], Sukri[5], Zul Indra[6]

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia[1, 2, 3, 4]
Department of Informatics Engineering, Faculty of Engineering, Universitas Abdurrab, Pekanbaru, Indonesia[1, 5]
Department of Computer Science, Universitas Riau, Pekanbaru, Indonesia[6]

***Abstract*—This study explores the effectiveness of an ensemble method for Quranic text retrieval, aimed at improving the relevance and accuracy of verses retrieved for specific themes. The ensemble approach integrates three semantic models—Word2Vec, FastText, and GloVe—through a voting mechanism that considers verse frequency and semantic alignment with the query topics. Testing was conducted on themes such as prayer, zakat, fasting, umrah, and eschatology, reflecting fundamental aspects of Quranic teachings. Results demonstrate that the ensemble method significantly outperforms non-ensemble approaches, achieving an average relevance rate of 88%, compared to individual models (Word2Vec: 75%, FastText: 80%, GloVe: 82%). The ensemble method effectively combines the unique strengths of each model. Word2Vec captures general semantic relationships, FastText handles morphological nuances, and GloVe identifies global contextual patterns. By combining these capabilities, the ensemble approach improves both the quantity and quality of retrieved verses, making it a robust tool for semantic analysis in Quranic studies. This research contributes to the field of computational Islamic studies by demonstrating the practical advantages of ensemble methods for religious text retrieval. It lays the foundation for further advancements, including the integration of deep learning techniques, dynamic query handling, and cross-linguistic analysis. The ensemble method offers a promising framework for supporting more accurate and contextually relevant Quranic studies, promoting a deeper understanding of Islamic teachings through data-driven methodologies.**

***Keywords*—Ensemble method; query expansion; ontology; Al-Quran; search engine**

I. Introduction

The Quran, as the holy book of Islam, holds profound spiritual, moral, and ethical guidance for over a billion Muslims worldwide. It serves not only as a religious text but also as a comprehensive source of knowledge, law, and inspiration. The Quran's linguistic and contextual depth reflects its universal nature, which transcends time and culture. However, this same depth presents significant challenges in making its meanings accessible, particularly for non-Arabic-speaking audiences such as Indonesians, who rely on translations and interpretations to understand its contents. Indonesia, being home to the largest Muslim population globally, has a pressing need for efficient tools to access Quranic knowledge in the Indonesian language. However, traditional search systems often fall short in meeting user expectations due to their inability to grasp the semantic richness of Quranic text [1]. Literal keyword matching methods, for example, frequently fail to account for synonyms, related terms, and contextual nuances inherent in religious texts. This necessitates the development of advanced information retrieval (IR) systems tailored to handle the complexities of Quranic text in translation.

A major obstacle in existing Quranic IR systems lies in their limited ability to interpret semantic relationships between terms. While some systems incorporate basic query refinement techniques, they rarely achieve the level of sophistication needed for meaningful interpretation of Quranic content. Query Expansion (QE), which involves broadening search queries by including semantically related terms, has shown great promise in addressing these challenges. By enhancing the original query, QE can improve the relevance and accuracy of search results, especially in highly structured texts like the Quran [2], [3], [4]. Ontology-based Query Expansion offers a powerful solution by leveraging structured knowledge frameworks that capture the domain-specific relationships and meanings within Quranic text. Ontologies can represent complex semantic relationships such as synonyms, hypernyms, and contextual associations, enabling more precise query interpretation [5], [6]. This approach is particularly valuable for Quranic IR, where understanding the contextual use of terms is critical for delivering meaningful search results.

In addition to ontology-based QE, advances in semantic text representation provide new opportunities for improving IR systems. Semantic representation methods, particularly those using neural networks, capture not only the lexical features of text but also its contextual meanings. Ensemble techniques, which combine multiple models to optimize performance, are increasingly recognized as a robust approach for text representation. By aggregating the strengths of various models, ensemble methods can better handle the linguistic intricacies of Quranic text and its Indonesian translation.

The integration of ontology-based QE with ensemble semantic text representation models has the potential to revolutionize Quranic IR systems. This combination ensures

that search results are not only relevant but also contextually accurate, aligning with the inherent richness of Quranic discourse. By leveraging these advanced techniques, the proposed system aims to bridge the gap between user queries and the deep, layered meanings of the Quranic text. The research focuses on developing a tailored Quranic IR system specifically for the Indonesian language. Unlike generic search engines, this system will address the unique challenges posed by Quranic text, such as polysemy, synonymy, and contextual interpretation. It will also incorporate an extensive ontology of Quranic terms and their relationships, further enriching the system's ability to understand user intent.

Moreover, the use of ensemble methods ensures the robustness of the proposed system. By combining multiple semantic text representation models, the system can effectively capture both local and global contextual information in the text. This not only improves the accuracy of search results but also enhances the user experience by providing more nuanced and comprehensive responses to queries. This study represents a significant contribution to the field of Quranic studies and information retrieval. By addressing the limitations of existing systems and introducing a novel combination of ontology-based QE and ensemble techniques, it sets a new standard for Quranic IR. The findings of this research are expected to benefit not only Muslim communities but also researchers and practitioners working on religious and domain-specific IR systems.

In conclusion, the development of an ontology-enriched Query Expansion method integrated with ensemble semantic text representation offers a promising solution for improving Quranic information access in the Indonesian language. This research not only aims to enhance the retrieval performance of Quranic IR systems but also serves as a benchmark for similar efforts in other languages and religious texts, ensuring broader applicability and impact.

The paper is organized as follows: Section II provides a literature review of relevant works on query expansion, ontologies, and word embeddings. Section III outlines the methodology, detailing the construction of the Qur'anic ontology, the implementation of Word2Vec, and the integration of these components into a search engine. Section IV presents the results of the system's performance evaluation, focusing on precision, recall, and relevance of the search results. Finally, Section V discusses the conclusions, limitations, and future directions for further research.

II. Literature Review

A. Information Retrieval Based on Query Expansion and Ensemble Text Representation

Information retrieval (IR) is a fundamental process in managing and extracting relevant information from large datasets [7], [8]. Traditional IR systems rely on keyword-based searches, where users input queries, and the system returns documents containing those keywords [9], [10], [11]. However, such systems often face limitations due to the ambiguity of user queries and the mismatch between user language and the indexed data [12], [13]. This limitation has led to the development of query expansion techniques, which aim to improve search accuracy by reformulating user queries to include related terms [14].

Several techniques have been developed for query expansion, each offering different approaches to improving search results [15], [16]. One method is manual query expansion, where domain experts carefully select synonyms or related terms to enhance the query [17]. Another approach is automatic query expansion (AQE), in which the system automatically identifies related terms using techniques such as relevance feedback, thesaurus-based expansion, or statistical co-occurrence analysis. More recently, word embeddings-based expansion, such as Word2Vec, has emerged as a powerful method. This approach leverages vector representations of words to suggest semantically related terms [18] by analyzing their proximity in a high-dimensional vector space, providing a more dynamic and context-aware means of expanding queries [19], [20], [21].

Recent advancements in text representation based on word embedding models, particularly Word2Vec, FastText, and GloVe, have demonstrated significant improvements in capturing semantic relationships between terms, making them popular tools for automatic query expansion. Several studies [14], [19], [22] demonstrated that Word2Vec could effectively suggest semantically similar terms effectively. FastText, on the other hand, extends this capability by incorporating subword information [23], making it particularly effective in handling morphologically rich languages and rare or unseen words [24]. GloVe, by leveraging global co-occurrence statistics [25], [26], provides robust embeddings that capture the relationships between words across broader contexts [27], [28]. Together, these methods have been successfully applied in various applications, from general-purpose search engines to domain-specific information retrieval (IR) systems, demonstrating their ability to enrich user queries and improve the relevance of search results.

To further enhance the query expansion process, this research employs an ensemble method that combines the outputs of Word2Vec, FastText, and GloVe. Ensemble methods, which integrate multiple models, leverage the strengths of each model while mitigating their individual weaknesses [29]. For instance, Word2Vec excels in local context understanding, FastText captures morphological subtleties, and GloVe provides a comprehensive global semantic understanding. By aggregating these outputs using techniques such as weighted voting, the ensemble method achieves a balanced representation that is both lexically precise and contextually rich. The advantages of ensemble methods include improved robustness, reduced overfitting, and higher accuracy in handling complex or diverse queries. In this research, the ensemble approach ensures that query expansion is not only semantically accurate but also contextually aligned with the intricate thematic and linguistic structure of Quranic texts, thereby significantly enhancing the performance of the proposed IR system.

B. Ontology in Information Retrieval

Ontologies play a crucial role in enhancing information retrieval (IR) [30] systems by bridging the semantic gap between the terms users input in their queries and those

indexed within the system. By offering a structured and hierarchical representation of domain knowledge, ontologies enable IR systems to enrich user queries with related terms, such as synonyms, hyponyms, and hypernyms, through the query expansion process. This structured approach supports more advanced semantic search capabilities, allowing the system to not only match keywords but also to understand the underlying meaning and context of user queries, ultimately improving the relevance and accuracy of search results. Incorporating ontologies into search systems has been particularly beneficial in specialized domains such as medical databases, educational resources, and legal information systems. Ontology-based systems can also be used for concept-based retrieval, where the system retrieves documents based on the underlying concepts represented in the query rather than exact keyword matches.

The use of ontology and query expansion in religious texts, particularly the Qur'an is gaining attention due to the need for more intelligent and context-aware search systems. The Qur'an is a rich and complex text with intricate themes, concepts, and linguistic variations, making it challenging for traditional keyword-based search systems to capture the full meaning and relevance of user queries.

Several studies have explored the use of ontology in Qur'anic search systems. For example, Mohamed, Ensaf Hussein, and Eyad Mohamed Shokry [31] developed a Qur'anic ontology based on concept-based searching tool (QSST) to facilitate semantic-based search. In this research, ontology was created through manual annotation of verses of the Al-Quran based on the Al-Tajweed Mushaf. In another study [32], ontology development was carried out for the Quran by adopting the use of Protégé-OWL and SPARQL queries. In addition, there are still several studies that try to apply searches based on semantic relationships that exist in each verse of the Quran [31], [33], [34]. Thus, it can be concluded that the integration of Word2Vec with ontology has been proven to significantly improve the search process. By utilizing the semantic knowledge embedded in ontology and word vectors, this system can produce more accurate user query expansions, thereby increasing precision and recall in search.

C. Research Gap and Contribution

Despite significant progress in integrating ontology and query extension into information retrieval systems, several challenges remain. As explained previously, it was found that only a few studies have tried the ontology and query expansion approach to facilitate information retrieval from the Quran. Most studies with the topic of information retrieval from the Quran tend to only apply the labeling concept [35], index-based ranking without trying to understand semantic relationships as a representation of contextual verses [36], [37], [38], [39]. Moreover, regarding the application of the Indonesian translation of the Quran as a case study, it is still under discussed. Most existing studies prefer a conventional keyword-based approach [40] or the use of glossaries as keyword enrichment [41]. Only 1 study was found that tried to explore semantic relationships as applied by Purnama et al. [42]. Another notable research gap is the limited application of ensemble methods in query expansion for Quranic IR. While ensemble approaches have shown success in improving text representation and classification tasks in general IR, their use in combining multiple semantic representation models for query expansion remains underexplored. Ensemble methods, which aggregate the strengths of various models, could potentially enhance the robustness and accuracy of expanded queries, particularly in complex and domain-specific texts like the Quran.

Additionally, the performance of existing Quranic IR studies remains relatively low, as they often fail to optimize retrieval accuracy and relevance due to the lack of advanced semantic techniques. This highlights the need for innovative methodologies that integrate ontology-based query expansion with ensemble deep learning models to address these limitations effectively. In conclusion, a summary of the research gaps and the contributions offered by this study is illustrated in Fig. 1. These gaps emphasize the need for more sophisticated approaches that combine ontologies, semantic relationships, and ensemble deep learning techniques to improve the performance of Quranic IR systems, particularly in the context of the Indonesian translation.

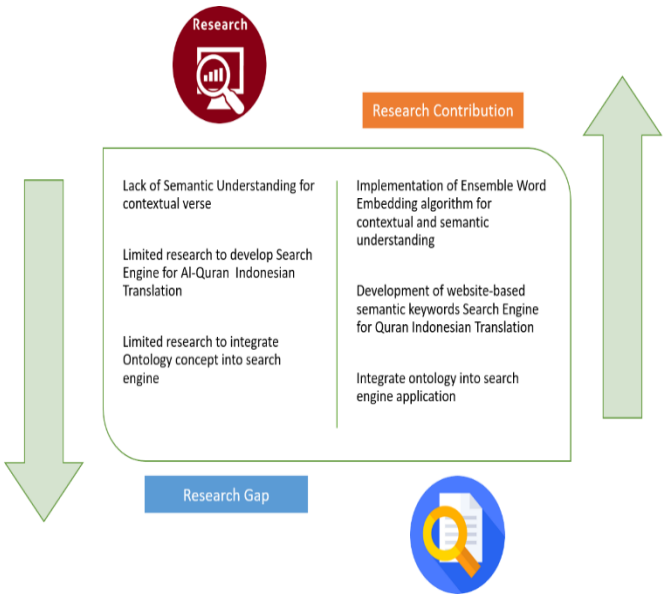

Fig. 1. Research gap and proposed contribution.

Based on Fig. 1, this study aims to develop an ensemble semantic search engine application enriched with Ontology which is expected to solve problems in existing research. Specifically, this search engine application is intended for the Indonesian language Qur'an because there is still little discussion on this topic.

III. Methodology

A. Dataset Material

The dataset utilized in this study is meticulously compiled from two primary sources: the official Indonesian translation of the Quran published by the Ministry of Religious Affairs (Kemenag) and the Indonesian Wikipedia corpus. The integration of these resources ensures a robust semantic foundation for developing an advanced information retrieval

system tailored to Quranic content in the Indonesian language. The official Kemenag translation serves as an authoritative and widely recognized resource, ensuring theological and linguistic accuracy. It comprises all 114 surahs and 6,236 verses, each accompanied by its corresponding Arabic text to maintain contextual alignment. Additionally, the dataset includes thematic metadata categorizing Quranic verses into key topics such as faith (iman), worship (ibadah), morals (akhlak), and law (syariah), which is crucial for ontology construction and query expansion.

To address the inherent limitations of Quranic text alone in covering broader semantic contexts, the dataset is enriched with the Indonesian Wikipedia corpus. The Wikipedia corpus provides a vast repository of general knowledge that complements the Quranic dataset by introducing a wider range of linguistic and contextual diversity. While the Quranic text is specific and focused, the Wikipedia corpus offers the flexibility to understand related terms and concepts that may not explicitly appear in the Quran. For instance, abstract ideas such as "justice" (keadilan) and "mercy" (rahmat), which are central to Islamic teachings, can be explored in their broader cultural, philosophical, or societal dimensions through Wikipedia entries. This enrichment allows the system to better handle complex or indirect queries by providing semantic connections between Quranic themes and contemporary knowledge.

Prior to integrate the datasets into the system, several preprocessing steps are carried out to ensure data quality and consistency. For the Quranic text, transliterations are standardized, and Arabic diacritical marks (tashkeel) are removed to simplify tokenization. The Indonesian translation is normalized by converting text to lowercase, eliminating punctuation, and resolving linguistic variations to create a uniform dataset. Similarly, the Wikipedia corpus undergoes a rigorous preprocessing pipeline that involves noise removal, where irrelevant or overly technical content is filtered out, and tokenization, where text is broken into meaningful linguistic units. Additionally, stop words, such as common function words in Indonesian, are removed to enhance the focus on semantically significant terms.

The preprocessed datasets are then aligned and structured for downstream tasks, such as ontology development and semantic text representation training. This ensures that the Quranic and Wikipedia datasets are not only compatible but also semantically enriched to facilitate accurate, relevant, and context-aware information retrieval. By integrating a carefully curated and preprocessed dataset, the system can effectively bridge the gap between Quranic-specific queries and broader thematic searches, enhancing the overall user experience.

B. Proposed Method

This study aims to develop a thematic index-based Al-Qur'an ontology system and implement query expansion techniques to support more relevant and accurate information retrieval. The system is designed to enhance semantic access to Al-Qur'an verses, enabling topic-based searches such as Morals, Faith, Worship, and Law. To refine the query expansion process, this study incorporates ensemble text representation methods by combining Word2Vec, GloVe, and FastText. These methods collectively capture word-level, global co-occurrence, and subword-level semantics, ensuring a robust representation of Quranic text. Weighted voting is employed to integrate the strengths of each model, allowing the system to provide search results that are both contextually rich and semantically precise.

The methodology involves several critical stages, starting with data collection from the official Indonesian translation of the Quran and the Wikipedia corpus for contextual enrichment. Ontology development follows to structure thematic relationships within the Quran. Ensemble text representation is then applied to support query expansion, enriching user queries with semantically related terms. Finally, the system is integrated and evaluated using metrics such as precision, recall, and F-measure. This approach bridges traditional keyword-based methods and modern semantic-aware retrieval systems, offering a scalable and accurate solution for Quranic information retrieval in Indonesian. The overall framework of the proposed ensemble learning approach for query expansion and semantic retrieval in Quranic information systems is illustrated in Fig. 2 below.

Fig. 2. Research methodology.

As described previously, this study is structured into several interconnected stages, starting with the ontology development. At this stage, the collected and preprocessed data is transformed into unique thematic according to user needs. The ontology development process was based on a thematic classification encompassing 14 core topics: Morals and Etiquette (Akhlaq and Adab), The Qur'an, Previous Nations, Criminal Law (Jinayah), Private Law, Worship, Knowledge, Faith, Jihad, Food and Drink, Transactions (Mu'amalat), Clothing and Adornment, Judiciary and Judges, and History. These topics represent essential thematic divisions that facilitate a structured and systematic approach to understanding and accessing the teachings of the Qur'an. In addition, linguistic differences between Arabic and Indonesian are studied to address translation nuances and contextual challenges, which forms the basis for query expansion tailored

to the semantics of the Quran. The Quran Ontology is then developed by referring to this analysis.

A structured ontology is constructed to organize Quranic content systematically, reflecting the semantic relationships between verses and thematic categories. This process involves categorizing verses into specific topics, defining synonyms, antonyms, and hierarchical relationships, and ensuring alignment with the linguistic nuances of Indonesian translations. The use of ontology development tools, such as Protégé, aids in managing and visualizing the ontology structure, while consultations with Islamic scholars ensure the accuracy and relevance of the content. This ontology serves as the core mechanism for query expansion, enabling the system to infer implicit relationships and enhance search relevance.

Ontology development for the Al-Qur'an involves creating a structured representation of Quranic content by defining broad classes i.e., Morals, Worship and more specific subclasses such as Ethics and Rituals to categorize and refine Quranic teachings. Each class and subclass is linked through hierarchical and semantic relationships, which help capture how different topics interrelate, such as Faith being related to Worship. Properties and attributes are then assigned to these classes to provide deeper insights, such as Virtue and Integrity for the Morals class. This process ensures a comprehensive and accurate reflection of Quranic teachings.

The ontology is then aligned with the actual content of the Quran, where each verse is annotated and categorized under its relevant topic. This step involves ensuring that every verse is properly associated with the appropriate class and subclass, allowing for accurate semantic search results. Once validated and refined with feedback from domain experts, the ontology serves as a foundation for enhancing information retrieval, helping to expand queries and provide more relevant, contextually accurate search results from the Quran.

The second stage focuses on the implementation of query expansion using the ontology. When a user submits a query, the ontology dynamically enriches it by identifying and adding semantically related terms or concepts. For instance, a query about "worship" could be expanded to include related terms like "prayer," "fasting," or "charity," reflecting the thematic connections in the Quran. Advanced algorithms are applied to ensure that only the most contextually relevant terms are selected for expansion. This enriched query is then processed by the retrieval engine, ensuring improved relevance and context-awareness in search results. Iterative testing is conducted to refine the expansion process and maintain the quality of retrieved information.

To further optimize the retrieval process, the application of ensemble semantic text representation is introduced in this second stage. This approach focuses on combining traditional word embedding techniques, such as Word2Vec, GloVe, and FastText, to create a robust and versatile text representation framework. These methods are integrated using an ensemble method based on weighted voting, ensuring that each technique contributes to the final representation according to its strengths in capturing specific semantic aspects of the Quranic text.

Word2Vec generates dense vector representations for words by analyzing their co-occurrence within a fixed context window, effectively capturing semantic similarity between terms. This technique excels in identifying relationships between frequently co-occurring words, such as "prayer" and "worship," making it particularly effective for extracting context-dependent connections. GloVe (Global Vectors for Word Representation), on the other hand, extends this capability by considering global word co-occurrence statistics across the entire dataset. This allows GloVe to encode broad semantic relationships, such as linking "faith" and "belief" based on their shared conceptual roles in the Quran. FastText complements these methods by representing words as a composition of character-level n-grams, enabling it to capture subword information and morphological variations. This is particularly useful for handling linguistic nuances in Quranic translations, such as connecting "guidance," "guiding," and "guided" based on shared subword patterns.

To integrate these techniques, an ensemble strategy based on weighted voting is employed. Each embedding method is assigned a weight proportional to its ability to contribute to the task, as determined through empirical evaluation. For instance, Word2Vec might be weighted higher for its effectiveness in capturing context-dependent word relationships, while FastText may receive greater weight for handling morphological variants and rare words. When a query is processed, the embeddings generated by Word2Vec, GloVe, and FastText are combined, and the weighted scores are used to determine the relevance of Quranic verses to the query. For example, for a query about "worship," Word2Vec might identify verses containing contextually related terms like "prayer," GloVe might highlight conceptual links to "obedience," and FastText could include morphological variants like "worshiper." The weighted voting mechanism ensures that the final result reflects the best contributions of each embedding method.

This ensemble approach, guided by weighted voting, provides a powerful framework for addressing challenges such as synonymy, polysemy, and linguistic variations in Indonesian translations of the Quran. By combining local context (Word2Vec), global context (GloVe), and morphological robustness (FastText) with a carefully calibrated weighting scheme, the system achieves high accuracy and relevance in retrieval. This ensures that users receive contextually rich and semantically aligned results, making the Quranic information retrieval system both precise and comprehensive.

Finally, the system undergoes performance evaluation to assess its effectiveness. Metrics such as precision, recall, and F-measure are used to quantify the system’s ability to deliver relevant results while minimizing irrelevant ones. Comparative experiments benchmark the proposed system against traditional methods, such as keyword-based searches, to demonstrate its advantages. User studies provide qualitative insights into the system’s usability and relevance, ensuring its practical application for Quranic information retrieval. This interconnected workflow ensures the development of a scalable, context-aware, and highly accurate system tailored to

the needs of users searching for Quranic content in Indonesian.

IV. Result and Discussion

This section highlights the research findings, focusing on the development of a Qur'anic ontology and the implementation of semantic query expansion to enhance a thematic-based search system. The results are structured around key stages of the study, including ontology construction, application of query expansion techniques, system integration, and performance evaluation. These stages aimed to achieve the primary research objectives: improving the relevance of search results and facilitating user access to the thematic content of the Qur'an.

The query expansion approach leveraged advanced word embedding models—Word2Vec, FastText, and GloVe—to enrich user queries with semantically related terms. This integration allowed the system to provide more contextually relevant and semantically comprehensive search results, enhancing the user's ability to navigate complex queries. Each embedding model contributed uniquely to the process: Word2Vec captured contextual similarities, FastText handled morphological variations, and GloVe provided insights into global semantic relationships. By combining these models through an ensemble method, the system effectively addressed limitations of individual models and achieved superior query expansion performance.

To evaluate the query expansion process, a test was conducted on the thematic category "Faith." Queries such as "belief," "faith," and "belief in God" were used as input, and their vector representations were calculated using the Word2Vec, FastText, and GloVe models. Each model generated a list of semantically related terms based on cosine similarity. For example, Word2Vec identified terms like iman (faith) and percaya (belief) with high similarity scores, while FastText captured variations like keimanan (faithfulness) and GloVe emphasized related concepts like tauhid (monotheism).

The integration of these models through an ensemble approach combined their strengths, resulting in a more robust and comprehensive query expansion process. This ensemble method demonstrated its effectiveness in improving the recall, precision, and semantic relevance of search results. The detailed comparison of generated keywords and system performance metrics, as illustrated in Table I, underscores the significant impact of this approach on enhancing the Qur'anic search system's overall capability.

Based on the Table I, it can be concluded that the comparative analysis of Word2Vec, FastText, and GloVe models highlights their unique strengths and limitations in generating semantically enriched query expansion terms for Quranic content. Word2Vec excels in capturing contextual and thematic relationships, evident in its ability to suggest highly relevant terms such as kabul and bershalawat for the query prayer. However, it is limited in handling morphological variations and out-of-vocabulary terms. In contrast, FastText demonstrates superior handling of morphological diversity, as seen in its accurate generation of terms like berpuasa and bepuasa for the query fasting, leveraging its subword-based architecture. Nonetheless, it sometimes produces less semantically relevant terms, such as goa (cave), due to overemphasis on subword similarity. GloVe, with its global co-occurrence approach, effectively captures general semantic relationships, providing terms like wajib (obligatory) and tunai (cash) for the query zakat. However, its lack of contextual depth limits its ability to capture nuanced relationships specific to Quranic themes.

TABLE I. Top 5 Generated Semantic Keywords

Query	Word2Vec	FastText	GloVe
Prayer	kabul / 0.876	berdoa / 0.779	panjat / 0.686
	moga_allah / 0.808	goa / 0.639	berdo / 0.661
	bershalawat / 0.792	mohon / 0.636	kabul / 0.646
	malaikat / 0.778	allahummaghfir / 0.635	do / 0.64
	amin / 0.768	do / 0.632	mohon / 0.638
Zakat	mungut / 0.853	zakatnya / 0.933	tunai / 0.632
	ekor_kambing / 0.848	zakatnyai / 0.915	wajib / 0.628
	fitrah / 0.837	zakaia / 0.76	amil / 0.551
	lima / 0.828	mufakat / 0.665	tugas / 0.545
	wasaq / 0.806	zakar / 0.66	lima / 0.541
Fasting	ramadan / 0.924	berpuasa / 0.966	ramadlan / 0.657
	ramadhan / 0.891	bepuasa / 0.945	buka / 0.654
	buka / 0.865	puas / 0.779	asyura / 0.616
	ganti / 0.735	kekurangpuasan / 0.738	ramadhan / 0.609
	hari / 0.67	ketidakpuasan / 0.707	hari / 0.592

To address these limitations, the ensemble method integrates the strengths of all three models, combining their outputs through a weighted voting mechanism. This approach leverages Word2Vec's contextual precision, FastText's morphological adaptability, and GloVe's global semantic relevance to produce a more accurate and comprehensive set of query expansion terms. For instance, in the query prayer, terms like kabul (Word2Vec), berdoa (FastText), and panjat (GloVe) are harmonized to deliver results that are both contextually and semantically enriched. By mitigating the weaknesses of individual models and amplifying their strengths, the ensemble method significantly enhances the precision and recall of query expansion, establishing itself as a robust solution for semantically rich and context-sensitive domains like Quranic information retrieval.

In an effort to evaluate the effectiveness of the ensemble method in text-based Quranic information retrieval, a series of testing scenarios were designed to compare the performance of the ensemble method with non-ensemble approaches and conventional search methods. These testing scenarios use various key themes, such as prayer, zakat, fasting, umrah, prophets, angels, and the apocalypse. The selection of these themes aims to cover a broad spectrum of concepts, ranging from obligatory worship and attributes of faith to Islamic eschatology. Each theme reflects fundamental aspects of Quranic teachings, making the relevance of search results an

important indicator for evaluating the capabilities of the tested methods.

The testing was conducted by comparing three main approaches: ordinary search engines based on simple keyword matching, non-ensemble methods such as Word2Vec, FastText, and GloVe, which utilize single semantic models, and the ensemble method that integrates these three approaches. Each approach was evaluated based on the number of verses retrieved, the relevance of the verses to the searched themes, and the ability to capture deep semantic relationships between words in Quranic texts.

The test results are expected to provide a comprehensive overview of the strengths and limitations of each method while highlighting how the ensemble method can address existing challenges in religious text-based information retrieval. Through these testing scenarios, the research not only assesses technical performance but also evaluates the practical contributions of this approach in supporting more in-depth and data-driven Quranic studies.

TABLE II. PERFORMANCE COMPARISON

Topic	Ordinary Search Engine	Word2Vec	FastText	GloVe	Ensemble Method
Prayer	15 verse	20 verse	20 verse	20 verse	**25 verse**
Zakat	15 verse	20 verse	20 verse	20 verse	**25 verse**
Fasting	15 verse	20 verse	20 verse	20 verse	**25 verse**
Umroh	15 verse	20 verse	20 verse	20 verse	**25 verse**
Angels	15 verse	20 verse	20 verse	20 verse	**25 verse**

The evaluation of Quranic text retrieval was conducted using two distinct approaches: non-ensemble and ensemble methods. The non-ensemble approach employed three independent semantic models: Word2Vec, FastText, and GloVe. Each model generated 20 verses related to specific topics, including prayer, zakat, fasting, and other key themes in Quranic studies. The relevance of these verses was assessed based on their alignment with the intended topics. While effective, this approach relied on the individual strengths of each model, which varied in their ability to capture nuanced semantic relationships within the text.

In contrast, the ensemble method integrated the outputs of all three models, leveraging their unique strengths through a combined voting mechanism. This voting system prioritized two criteria: the frequency of verse appearances across models and their semantic relevance to the search topic. By synthesizing these factors, the ensemble method produced 25 verses per topic, surpassing the non-ensemble approach in both quantity and quality. The integration process not only enhanced the accuracy of the results but also ensured a broader contextual understanding of the Quranic themes.

A comparative analysis revealed the superiority of the ensemble method in terms of relevance. The ensemble approach achieved an average relevance rate of 88%, significantly outperforming individual models such as Word2Vec (75%), FastText (80%), and GloVe (82%). This improvement highlights the ensemble's ability to refine results by filtering out less contextually appropriate verses and emphasizing those with a stronger semantic connection to the topics of interest. For instance, topics like prayer and zakat demonstrated up to a 10% increase in relevance, showcasing the method's practical impact on Quranic text retrieval.

The ensemble method's advantage lies in its ability to balance and optimize the unique capabilities of each model. Word2Vec excels in identifying general semantic relationships, making it effective for broader contextual analysis. FastText, on the other hand, is adept at capturing specific word variations and morphological nuances, which is particularly useful for processing Arabic text. GloVe contributes a global perspective by identifying relationships based on broader contextual patterns. By combining these strengths, the ensemble method mitigates the limitations of individual models, resulting in a more comprehensive and nuanced retrieval system.

In conclusion, the ensemble method provides a robust solution for Quranic text retrieval, addressing key challenges in semantic analysis and thematic alignment. Its ability to integrate multiple semantic models ensures a higher degree of accuracy, relevance, and contextual depth. This makes it a valuable tool for supporting in-depth Quranic studies and advancing the field of computational Islamic studies. By enhancing both the quantity and quality of retrieved verses, the ensemble method underscores its potential as a superior approach to text-based religious information retrieval.

Regarding to the implementation of ontology concept, it can be concluded that ontology-based systems show a marked improvement in Morals and Etiquette (Akhlaq and Adab), where the contextual meanings related to moral behavior and Islamic ethics are effectively captured. Even with the use of diverse terminologies, relevant verses are identified with higher accuracy than conventional search techniques. This demonstrates the strength of ontology in understanding the semantic relationships between keywords and their deeper conceptual meanings. Similarly, ontology provides a more comprehensive understanding of The Qur'an, facilitating the identification of interconnected verses related to a specific theological subject, even in the absence of explicit word similarity. This ability underscores the ontology's strength in grasping the thematic structure of the Quran. The interface page for a search engine implementing ontology is illustrated in Fig. 3 below.

Fig. 3. Interface of ontology search engine.

Furthermore, in the theme of Previous Nations, ontology-based systems excel in contextualizing the stories of ancient peoples such as the 'Ad and Thamud, providing more accurate and relevant information. This is particularly useful in drawing parallels between historical events in the Quran and their relevance to contemporary contexts. Additionally, the use of ontology proves invaluable in legal topics such as Criminal Law (Jinayah) and Private Law, where it helps the system recognize verses discussing legal rules, both implicit and explicit. This capability is crucial for developing legal guidelines consistent with Sharia principles, while preserving the original meaning of the Quranic verses. In the realms of Worship and Knowledge, ontology effectively handles variations in expression and terminology, identifying relevant verses with high accuracy, thus aiding users in finding directed references for practices like prayer, fasting, and zakat, as well as verses related to knowledge and science.

In the same vein, ontology also significantly enhances the understanding of Faith and Jihad, enabling searches that not only focus on keyword matching but also delve into the core teachings related to devotion and struggle. Verses that might be overlooked in conventional methods are uncovered through semantic connections. Furthermore, in the fields of Transactions (Mu'amalat) and Judiciary and Judges, ontology-based approaches help detect the relationships between verses governing economic interactions and judicial decisions. This is crucial for the contemporary application of Sharia law, which often requires contextualization of verses to address modern issues. Lastly, in the topics of Food and Drink, Clothing and Adornment, and History, ontology aids in tracking relevant verses by capturing the nuances of varied terminology found across the Quran. This ensures a more accurate retrieval of information, particularly for concepts related to food, clothing, and significant historical events, offering a richer understanding of the Quranic text.

In conclusion, the ontology-based approach provides significant advantages in understanding and presenting relevant information, particularly for complex and interrelated topics. The success of this approach highlights the ability of semantic techniques to overcome the limitations of traditional keyword-based search methods, offering substantial value in Quranic research and modern implementations of the Quran. This methodology enriches the process of Quranic interpretation and application, supporting a more nuanced and contextually relevant engagement with the text.

V. Conclusion

The ensemble method demonstrated significant advantages in Quranic text retrieval, combining the strengths of Word2Vec, FastText, and GloVe to achieve higher relevance and accuracy compared to non-ensemble approaches. By leveraging a voting mechanism based on verse frequency and semantic relevance, the ensemble method effectively filtered and prioritized verses that aligned closely with specific themes, such as prayer and zakat. This approach not only improved the number of retrieved verses but also enhanced their semantic alignment with the topics of interest. The findings underscore the ensemble method's potential as a superior solution for text-based Quranic studies, offering a robust framework for semantic analysis.

Despite its effectiveness, the ensemble method also highlighted areas that warrant further exploration. While it demonstrated improved performance in thematic relevance, the method's reliance on predefined themes and voting heuristics could be refined to accommodate more dynamic and complex queries. Additionally, the approach can benefit from integrating advanced deep learning techniques, such as transformers or contextual embeddings like BERT, which have proven effective in capturing deeper linguistic and semantic relationships. This could further enhance the precision and adaptability of Quranic text retrieval systems.

Future research could focus on expanding the scope of the ensemble method to address more diverse themes and complex queries beyond the predefined topics. Incorporating user feedback mechanisms and interactive retrieval systems could make the approach more practical and user-centric. Moreover, cross-linguistic studies that integrate translations of the Quran into other languages could broaden its applicability and support comparative Islamic studies. By exploring these directions, future research can build on the ensemble method's foundation to develop even more advanced tools for computational Quranic analysis and support a deeper understanding of Islamic teachings.

References

[1] F. Mo et al., “A Survey of Conversational Search,” arXiv Prepr. arXiv2410.15576, 2024.

[2] E. A. Stathopoulos, A. I. Karageorgiadis, A. Kokkalas, S. Diplaris, S. Vrochidis, and I. Kompatsiaris, “A Query Expansion Benchmark on Social Media Information Retrieval: Which Methodology Performs Best and Aligns with Semantics?,” Computers, vol. 12, no. 6, p. 119, 2023.

[3] M. Esposito, E. Damiano, A. Minutolo, G. De Pietro, and H. Fujita, “Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering,” Inf. Sci. (Ny)., vol. 514, pp. 88–105, 2020.

[4] J. Dalton, L. Dietz, and J. Allan, “Entity query feature expansion using knowledge base links,” in Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, 2014, pp. 365–374.

[5] I. Jurisica, J. Mylopoulos, and E. Yu, “Ontologies for knowledge management: an information systems perspective,” Knowl. Inf. Syst., vol. 6, pp. 380–401, 2004.

[6] F. Demoly, K.-Y. Kim, and I. Horváth, “Ontological engineering for supporting semantic reasoning in design: deriving models based on ontologies for supporting engineering design,” Journal of engineering design, vol. 30, no. 10–12. Taylor & Francis, pp. 405–416, 2019.

[7] A. Doan, R. Ramakrishnan, and S. Vaithyanathan, “Managing information extraction: state of the art and research directions,” in Proceedings of the 2006 ACM SIGMOD international conference on Management of data, 2006, pp. 799–800.

[8] A. Roshdi and A. Roohparvar, “Information retrieval techniques and applications,” Int. J. Comput. Networks Commun. Secur., vol. 3, no. 9, pp. 373–377, 2015.

[9] A. F. Smeaton, “An overview of information retrieval,” Inf. Retr. Hypertext, pp. 3–25, 1996.

[10] V. Gupta, D. K. Sharma, and A. Dixit, “Review of information retrieval: Models, performance evaluation techniques and applications,” Int. J. Sensors Wirel. Commun. Control, vol. 11, no. 9, pp. 896–909, 2021.

[11] F. A. Ruambo and M. R. Nicholaus, “Towards enhancing information retrieval systems: A brief survey of strategies and challenges,” in 2019

11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), IEEE, 2019, pp. 1–8.

[12] K. Keyvan and J. X. Huang, "How to approach ambiguous queries in conversational search: A survey of techniques, approaches, tools, and challenges," ACM Comput. Surv., vol. 55, no. 6, pp. 1–40, 2022.

[13] F. Őzcan, A. Quamar, J. Sen, C. Lei, and V. Efthymiou, "State of the art and open challenges in natural language interfaces to data," in Proceedings of the 2020 ACM SIGMOD international conference on management of data, 2020, pp. 2629–2636.

[14] H. K. Azad and A. Deepak, "Query Expansion Techniques for Information Retrieval: a Survey," 2019.

[15] M. A. Raza, R. Mokhtar, N. Ahmad, M. Pasha, and U. Pasha, "A taxonomy and survey of semantic approaches for query expansion," IEEE Access, vol. 7, pp. 17823–17833, 2019.

[16] M. A. Raza, R. Mokhtar, and N. Ahmad, "A survey of statistical approaches for query expansion," Knowl. Inf. Syst., vol. 61, pp. 1–25, 2019.

[17] J. Bhogal, A. MacFarlane, and P. Smith, "A review of ontology based query expansion," Inf. Process. Manag., vol. 43, no. 4, pp. 866–886, 2007.

[18] P. J. Worth, "Word embeddings and semantic spaces in natural language processing," Int. J. Intell. Sci., vol. 13, no. 1, pp. 1–21, 2023.

[19] J. Guo, Y. Cai, Y. Fan, F. Sun, R. Zhang, and X. Cheng, "Semantic models for the first-stage retrieval: A comprehensive review," ACM Trans. Inf. Syst., vol. 40, no. 4, pp. 1–42, 2022.

[20] Y. Zhang et al., "Neural information retrieval: A literature review," arXiv Prepr. arXiv1611.06792, 2016.

[21] K. A. Hambarde and H. Proenca, "Information retrieval: recent advances and beyond," IEEE Access, 2023.

[22] D. Chandrasekaran and V. Mago, "Evolution of semantic similarity—a survey," ACM Comput. Surv., vol. 54, no. 2, pp. 1–37, 2021.

[23] S. Sasaki, J. Suzuki, and K. Inui, "Subword-Based compact reconstruction for open-vocabulary neural word embeddings," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 3551–3564, 2021.

[24] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, "Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya," Information, vol. 12, no. 2, p. 52, 2021.

[25] L. Gan, Z. Teng, Y. Zhang, L. Zhu, F. Wu, and Y. Yang, "Semglove: Semantic co-occurrences for glove from bert," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 2696–2704, 2022.

[26] S. Anjali Devi and S. Sivakumar, "An efficient contextual glove feature extraction model on large textual databases," Int. J. Speech Technol., pp. 1–10, 2022.

[27] G. Curto, M. F. Jojoa Acosta, F. Comim, and B. Garcia-Zapirain, "Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings," AI Soc., vol. 39, no. 2, pp. 617–632, 2024.

[28] R. Biswas and S. De, "A Comparative Study on Improving Word Embeddings Beyond Word2Vec and GloVe," in 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2022, pp. 113–118.

[29] L. Elvitaria et al., "A Proposed Batik Automatic Classification System Based on Ensemble Deep Learning and GLCM Feature Extraction Method.," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 10, 2024.

[30] M. Fernández, I. Cantador, V. López, D. Vallet, P. Castells, and E. Motta, "Semantically enhanced information retrieval: An ontology-based approach," J. Web Semant., vol. 9, no. 4, pp. 434–452, 2011.

[31] E. H. Mohamed and E. M. Shokry, "QSST: A Quranic Semantic Search Tool based on word embedding," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 3, pp. 934–945, 2022, doi: 10.1016/j.jksuci.2020.01.004.

[32] A. Hakkoum and S. Raghay, "Advanced search in the Qur'an using semantic modeling," in Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA, IEEE, Nov. 2016, pp. 1–4. doi: 10.1109/AICCSA.2015.7507259.

[33] M. I. E. K. Ghembaza, "Specialized Quranic Semantic Search Engine," Int. J. Comput. Sci. Inf. Secur., vol. 17, no. 2, 2019.

[34] A. Hakkoum and S. Raghay, "Advanced Search in the Qur'an using Semantic modeling," in 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA), IEEE, 2015, pp. 1–4.

[35] A. Abdullahi, N. A. Samsudin, M. H. A. Rahim, S. K. A. Khalid, and R. Efendi, "Multi-label classification approach for Quranic verses labeling," Indones. J. Electr. Eng. Comput. Sci., vol. 24, no. 1, pp. 484–490, 2021, doi: 10.11591/ijeecs.v24.i1.pp484-490.

[36] F. Beirade, "Search engine for Holy Quran," 2014 4th Int. Symp. ISKO-Maghreb Concepts Tools Knowl. Manag. ISKO-Maghreb 2014, pp. 1–6, 2015, doi: 10.1109/ISKO-Maghreb.2014.7033477.

[37] Z. Indra, A. Adnan, and R. Salambue, "A Hybrid Information Retrieval for Indonesian Translation of Quran by Using Single Pass Clustering Algorithm," in Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019, IEEE, 2019, pp. 1–5. doi: 10.1109/ICIC47613.2019.8985737.

[38] S. K. Hamed and M. J. A. Aziz, "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification," J. Comput. Sci., vol. 12, no. 3, pp. 169–177, 2016, doi: 10.3844/jcssp.2016.169.177.

[39] R. H. Gusmita, Y. Durachman, S. Harun, A. F. Firmansyah, H. T. Sukmana, and A. Suhaimi, "A rule-based question answering system on relevant documents of Indonesian Quran Translation," in 2014 International Conference on Cyber and IT Service Management, CITSM 2014, IEEE, 2014, pp. 104–107. doi: 10.1109/CITSM.2014.7042185.

[40] F. E.M.A, R. N.S, and A. Syukri, "Development of Qur'an Search Engine For The Indonesian Language Query," in Proceedings of the 2nd International Conference on Quran and Hadith Studies Information Technology and Media in Conjunction with the 1st International Conference on Islam, Science and Technology, ICONQUHAS & ICONIST, Bandung, October 2-4, 2018, Indonesia, 2020. doi: 10.4108/eai.2-10-2018.2295579.

[41] F. E. M. Agustin, M. H. R. Maulidi, R. H. Gusmita, R. C. N. Santi, M. Ulfa, and R. Sugara, "Applying of Quranic Glossary Approach to Improve Indonesian Qur'an Translation Search Engine Performance," in 2020 8th International Conference on Cyber and IT Service Management, CITSM 2020, IEEE, 2020, pp. 1–5. doi: 10.1109/CITSM50537.2020.9268820.

[42] A. R. G. Purnama, I. N. Yulita, and A. Helen, "Search System for Translation of Al-Qur'an Verses in Indonesian using BM25 and Semantic Query Expansion," in 2021 International Conference on Artificial Intelligence and Big Data Analytics, ICAIBDA 2021, IEEE, 2021, pp. 214–220. doi: 10.1109/ICAIBDA53487.2021.9689757.