Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Comparison of Lexical and Semantic Approaches for Relevance Measurement in Quranic Verse Translation Retrieval

Abd. Charis Fauzan[1)], M. Abd. Rouf [2)*], Tito Prabowo[3)], Utrodus Said Al Baqi[4)]
[1,3,4)]Deparment of Computer Science, Universitas Nahdlatul Ulama Blitar, Indonesia
[2)]Deparment of Quranic Studies and Tafsir, Universitas Nahdlatul Ulama Blitar, Indonesia
[1)]abdcharis@unublitar.ac.id, [2)]abdrouf@unublitar.ac.id*, [3)]titoprabowo@unublitar.ac.id, [4)]contact.utrodus@gmail.com

ABSTRACT

This research explores the effectiveness of lexical and semantic approaches for relevance measurement in Quranic verse translation retrieval, focusing on Indonesian translations. Quranic verses encompass complex linguistic structures and diverse contexts, making precise retrieval challenging. Two retrieval methods were evaluated: lexical similarity, which focuses on exact word matches, and semantic similarity, which captures contextual meaning using word embeddings. The study utilized a dataset of Indonesian Quranic translations, preprocessed to normalize and tokenize text, with experimental queries derived from thematic exegesis on social responsibility. Evaluation was performed using precision, recall, and F1-score on top-5, top-10, and top-15 retrieved results. The lexical approach achieved perfect precision (100%) but exhibited lower recall (46%-58%), as it failed to retrieve relevant verses lacking exact matches. Conversely, the semantic approach demonstrated higher recall (56%-59%) and F1-scores (73%-74%) by identifying verses with contextual relevance, even in the absence of lexical similarity. The results reveal that while the lexical approach ensures precise matches, it overlooks semantic richness. The semantic approach, although computationally intensive, achieves greater contextual understanding. These findings highlight the potential for hybrid retrieval systems combining both approaches to enhance accuracy and relevance in Quranic information retrieval, supporting scholarly research and user engagement with Quranic content.

Keywords: lexical; semantic, relevance measurement, quranic verce translation

INTRODUCTION

The Quran, revered as the ultimate guidance for Muslims, spans 30 chapters, 114 surahs, and 6,326 verses, covering diverse topics such as worship, monotheism, faith, law, human behavior, and narratives of previous communities (Husni & Arifin, 2019). This extensive coverage makes the Quran a rich source of knowledge, but it also poses significant challenges for those seeking specific guidance. Locating verses relevant to particular topics manually is time-consuming and often impractical, especially for individuals with limited expertise in Quranic studies. This challenge has spurred the development of various digital Quran applications to facilitate verse retrieval (Mohamed & Shokry, 2022).

One such tool is the Qur'an Kemenag app, developed by Indonesia's Ministry of Religious Affairs. The app enables users to search Quranic verse translations using Indonesian keywords, providing easier access to Quranic content (Puspitasari, 2022). However, this and similar systems often rely on keyword-based approaches, which depend on lexical similarity to measure the relevance between user queries and Quranic texts. While keyword matching can efficiently handle straightforward queries, it frequently struggles with

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

capturing the contextual meanings of words, leading to irrelevant or imprecise results. These limitations highlight the need for a more nuanced comparison of different similarity measurement techniques for Quranic retrieval systems (Afzal & Mukhtar, 2019).

In text retrieval systems, two predominant approaches for measuring similarity are lexical similarity and semantic similarity. Lexical similarity focuses on the structural and syntactic properties of words, comparing their forms for exact matches or patterns (Kuzi et al., 2020). This approach is computationally efficient and suitable for basic queries, making it a popular choice in traditional information retrieval systems. However, its reliance on exact word forms renders it inadequate for handling complex linguistic nuances, such as synonyms, polysemy, or contextual variations (Huang et al., 2013). For example, in Quranic verses, words with the same root may convey different meanings depending on their context, which lexical similarity fails to address.

On the other hand, semantic similarity measures the meaning of words and their relationships within a broader context, regardless of their lexical forms (Darmalaksana et al., 2020). This approach has gained prominence with the advent of advanced natural language processing (NLP) techniques, such as word embeddings and large language models (LLMs), which capture deeper semantic relationships between words (Mohd et al., 2020). Semantic similarity enables retrieval systems to understand the intent behind user queries, improving the relevance and precision of search results. For instance, semantic techniques can link queries about "prayer" with Quranic verses that discuss the act of worship using various terminologies, even if the keywords do not directly match (Ahmad et al., 2021).

Although semantic approaches address the limitations of lexical similarity, they come with their own challenges, such as higher computational costs and the need for extensive training data to build effective models. Additionally, semantic similarity may occasionally overgeneralize relationships between words, leading to results that are contextually inappropriate or overly broad. These trade-offs make it essential to evaluate and compare the performance of both lexical and semantic approaches, particularly in domains like Quranic retrieval, where precision and contextual accuracy are crucial. Previous studies have explored the application of lexical and semantic similarity in Quranic information retrieval. For instance, (Mohamed & Shokry, 2022) examined various Quranic search techniques, emphasizing the potential of semantic methods to overcome the limitations of keyword-based systems. Similarly, Anggraini et al., (2022) demonstrated the effectiveness of semantic approaches in improving retrieval precision through WordNet-based relationships. These findings suggest that semantic similarity offers distinct advantages over lexical approaches, but the extent of its superiority and the contexts where it excels require further exploration.

This study aims to conduct a comprehensive comparison between lexical and semantic similarity approaches in the context of Quranic verse translation retrieval. By evaluating their effectiveness, strengths, and limitations, the research seeks to provide insights into which approach is more suitable for specific types of queries and contexts. Unlike previous works that focus predominantly on the benefits of semantic methods, this study places equal emphasis on both techniques, highlighting their respective capabilities and shortcomings. The results of this comparison will contribute to the broader understanding of how different similarity measures impact retrieval performance. It will also provide valuable recommendations for developers and researchers in optimizing Quranic retrieval systems, ensuring users receive accurate and contextually relevant results. This focus on comparative analysis will not only address existing gaps in the literature but also support the development of more effective tools for accessing Quranic knowledge in diverse linguistic and cultural settings.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

LITERATURE REVIEW

Lexical Similarity in Information Retrieval

Lexical similarity focuses on the exact matching of words and syntactic patterns in text, making it a straightforward approach for text retrieval systems. It evaluates the relevance of text by comparing word forms, ensuring high precision for direct keyword-based queries (Kuzi et al., 2020). This method has been widely used in traditional search systems, including Quranic retrieval tools like the Qur'an Kemenag app, which relies on matching user-provided keywords with verse translations. However, lexical similarity has notable limitations. It fails to capture the contextual meaning of words, leading to suboptimal results when dealing with synonyms, polysemous terms, or phrases with complex linguistic structures. For example, in Quranic text retrieval, different root words or forms may convey similar meanings, yet a purely lexical approach would treat them as unrelated. Recent studies also highlight how lexical techniques struggle with languages like Arabic, which involve rich morphological variations and complex syntax (Rouf Abd. M. et al., 2023).

Semantic Similarity in Information Retrieval

Semantic similarity evaluates the meaning and contextual relationships between words, making it a more advanced alternative to lexical approaches. Using techniques like word embeddings and large language models (LLMs), semantic similarity captures deeper connections between terms, improving relevance in complex queries (Goddard & Schalley, 2010). For instance, semantic embeddings, such as BERT or FastText, model word relationships within broader contexts, allowing retrieval systems to understand synonyms and polysemous words (Abriani & Yaqin, 2019). In Quranic information retrieval, semantic approaches have shown promise in overcoming the limitations of keyword-based systems. Afzal & Mukhtar, (2019) implemented WordNet-based semantic search to improve the precision of Quranic verse retrieval, demonstrating how contextual understanding enhances relevance. Similarly, Ahmad et al., (2021) reviewed Quranic search systems and ontologies, emphasizing the importance of semantic techniques for accurate verse retrieval. Despite these advantages, semantic similarity poses its own challenges. Computationally, semantic models require extensive training data and significant resources, which may limit their scalability in resource-constrained environments (Darmalaksana et al., 2020). Furthermore, semantic systems occasionally overgeneralize word relationships, resulting in contextually inappropriate results (Gunay & Yolum, 2007).

Research Gap

Existing literature has extensively explored the applications of lexical and semantic similarity in text retrieval systems, but direct comparisons between these approaches in the context of Quranic verse retrieval remain limited. While semantic methods show clear advantages in capturing contextual meaning, their scalability and efficiency in practical applications need further investigation. This study addresses this gap by evaluating and comparing lexical and semantic similarity approaches for Quranic verse translation retrieval, providing insights into their relative strengths and weaknesses.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

METHOD

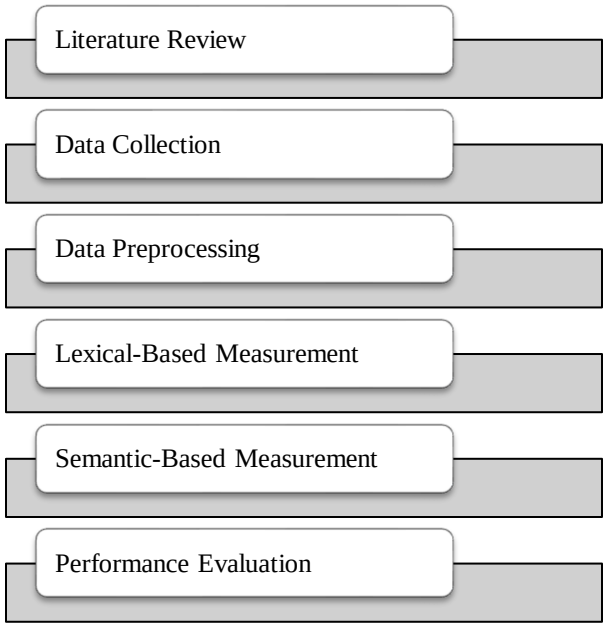


Figure 1. Research Method

The research methodology in Figure 1 is designed to systematically address the comparison of lexical and semantic approaches in measuring relevance for Quranic verse translation retrieval. The process begins with literature review, where prior research on lexical and semantic similarity, as well as their applications in Quranic retrieval systems, is explored to provide theoretical foundations and identify research gaps. Following this, data collection is conducted, focusing on acquiring a dataset of Quranic translations in Indonesian. This dataset is carefully curated to include various topics to test the effectiveness of both approaches comprehensively. Next, the implementation phase involves the application of lexical and semantic similarity measurement techniques. The lexical similarity is computed using word-based matching and structural comparisons, while semantic similarity utilizes advanced methods such as word embeddings or contextual models like BERT to evaluate meaning and context. These implementations are tested on user queries against the Quranic verse dataset to determine the relevance of retrieval results. The final phase is the evaluation of results, where the performance of both approaches is analyzed using metrics like precision, recall, and F1-score. The comparison provides insights into the strengths and weaknesses of each method, highlighting their applicability in Quranic retrieval systems. This structured methodology ensures a thorough and objective evaluation of both approaches.

Literature Review

In the literature review process for this research, the researchers gathered references from various sources such as articles, journals, and e-books. The literature review encompassed several topics relevant to the study, including Information Retrieval, Data Preprocessing Techniques, Lexical Similarity Measurement, Semantic Similarity Measurement, and the evaluation of retrieval results using metrics such as Recall, Precision, and F-Measure.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Data Collection

The data collection process begins with identifying reliable references for datasets containing the Quran and its translations in Indonesian. This step involves searching various sources, including existing databases, digital libraries, and repositories, with a primary focus on the official website of the Ministry of Religious Affairs of the Republic of Indonesia as the primary source. The website provides a trusted and authoritative database of Quranic translations in Indonesian, ensuring high accuracy and reliability. Once the database is identified, the next step is to thoroughly study and evaluate its content to ensure it meets the research criteria for completeness and relevance. Key components such as surah names, verse numbers, and translation texts are examined to verify their suitability for the study. After confirming the quality of the database, the process of data extraction begins. This involves systematically compiling Quranic translations from the Ministry's website and organizing them into a structured format. The final step involves refining the extracted data to create a well-organized dataset of Quranic verses and their Indonesian translations. This dataset is then prepared for the subsequent implementation and evaluation phases of the research. By utilizing an official and trusted source, the study ensures the reliability and validity of the dataset used for comparing lexical and semantic similarity approaches in Quranic verse retrieval.

Preprocessing Data



Figure 2. Steps of Preprocessing

Figure 2 represents the preprocessing steps used in this research. Preprocessing involves a sequence of operations that prepare textual data for further analysis or processing. The steps are as follows:

a. Case Folding

This initial step converts all letters in the text to lowercase and removes punctuation marks. The purpose of this step is to reduce complexity and eliminate duplicates caused by differences in uppercase and lowercase letters. By standardizing the text, case folding ensures consistency across the dataset and simplifies subsequent processing.

b. Tokenization

In this step, the text is broken down into its smallest units, called tokens. This process involves removing special characters and separating words based on spaces. For example, a sentence is split into individual words or tokens, making it easier to analyze each component separately. Python's Natural Language Toolkit (NLTK) library is used to implement tokenization effectively and efficiently. This step is critical for preparing the text for further analysis, as it isolates meaningful components of the text.

c. Stopword Removal

This stage removes frequently occurring words that do not contribute meaningfully to the context, such as conjunctions, prepositions, and common adjectives. These words, known as stopwords, are filtered out to reduce the complexity of the text and speed up processing. For this implementation, the Sastrawi library in Python is utilized, ensuring efficient removal of these unnecessary words from the dataset.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

d. Stemming

The final step transforms words into their root forms according to proper Indonesian morphological structures. By using the Sastrawi library, stemming removes affixes such as prefixes, suffixes, infixes, and confixes. This ensures that words with the same root are treated uniformly, simplifying the text for further analysis. For example, "berlari," "lari," and "berlarian" would all be reduced to the root word "lari." This step is essential for reducing redundancy and improving text analysis.

These preprocessing steps collectively ensure that the text is cleaned, standardized, and simplified, creating a robust foundation for subsequent stages of text processing and analysis.

Lexical-Based Measurement

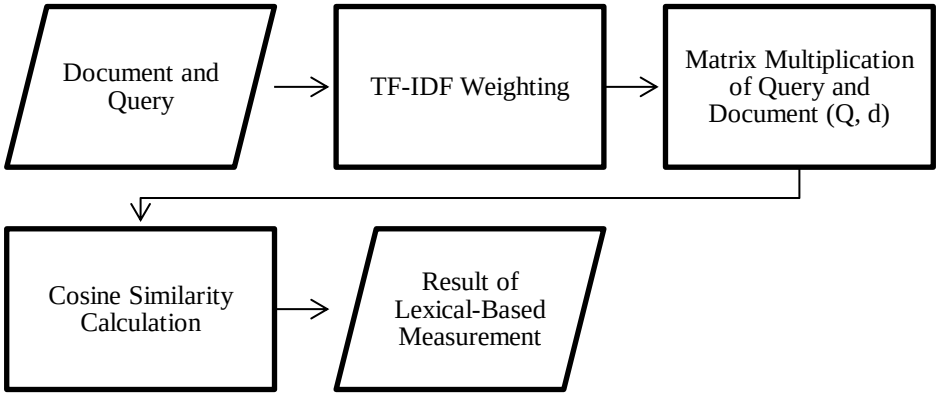


Figure 3. Steps of Lexical Based Measurement

The process of calculating lexical similarity follows a structured based on Figure 3 to evaluate the relevance between a query and a document based on word usage and structure. The steps are outlined below:

a. TF-IDF Weighting

The first step involves assigning weights to words in both the query and documents using the Term Frequency-Inverse Document Frequency (TF-IDF) method (Rofiqi et al., 2019). This approach emphasizes the importance of words that appear frequently in a document but are less common across the entire dataset, highlighting unique terms that contribute significantly to the content. TF-IDF Wighting is calculated using Equation (1).

$$TF.IDF(w_i, d_j) = W_{tf}(w_i, d_j) * W_{idf}(w_i) \dots\dots\dots (1)$$

b. Matrix Multiplication of Query and Document (Q, d)

After assigning TF-IDF weights, a mathematical operation multiplies the weighted vectors of the query and each document. This step quantifies the contribution of each term in the query to the terms in the document, preparing the data for similarity computation.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

c. Cosine Similarity Calculation

The cosine similarity formula is applied to measure the angle between the vector representations of the query and the document (Kesuma, 2016). This calculation determines the degree of alignment or similarity between the two, with a value closer to 1 indicating high similarity and a value closer to 0 indicating low similarity.

d. Lexical Similarity Scores

The final step involves generating lexical similarity scores for each query-document pair. The scores are ranked to identify the most relevant documents based on their similarity to the query. This calculated using Equation (2).

$$LexSim\ (Q, d_j) = \frac{\sum(TF.IDF(w_i, Q)).\left(TF.IDF\left(w_i, d_j\right)\right)}{\sqrt{\sum |TF.IDF_Q|^2}.\sqrt{\sum |TF.IDF_{d_j}|^2}} \dots\dots\dots (2)$$

This process ensures that lexical matching is accurately quantified, focusing on the structural alignment of terms in the query and the document while leveraging TF-IDF to prioritize meaningful terms. The use of cosine similarity further refines the evaluation by normalizing the comparison to account for differences in text length.

Semantic-Based Measurement

The process of semantic similarity calculation focuses on understanding the contextual meaning of the text in the Quranic verse translations and queries. It involves several steps to achieve a meaningful comparison between the query and documents based on Figure 4.

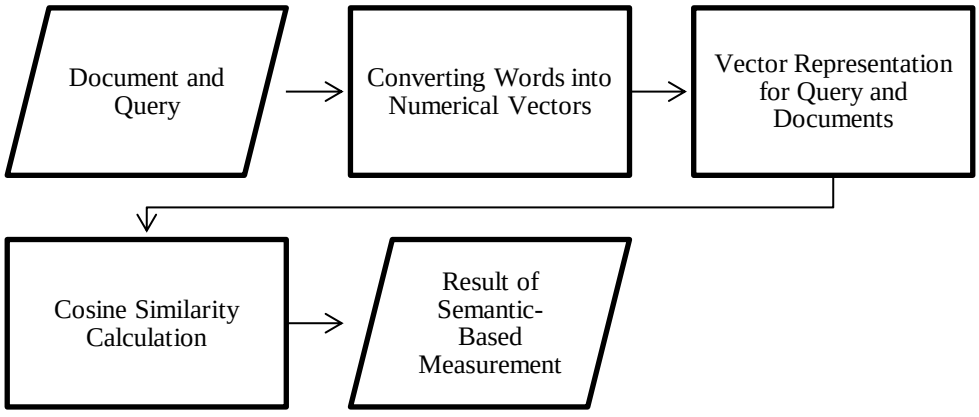


Figure 4. Steps of Semantic Based Measurement

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

a. Converting Words into Numerical Vectors

The first step involves transforming each word in the preprocessed text and query into numerical vectors (Naf'an et al., 2021). This is achieved using a pretrained word embedding model, such as FastText, which assigns dense vector representations to words based on their contextual usage in large datasets. These embeddings capture semantic relationships between words, enabling a deeper understanding beyond mere lexical matches.

b. Vector Representation for Query and Documents

Each document and query are represented as collections of numerical vectors. These vectors encode the semantic meaning of words, allowing for nuanced comparisons between the query and the documents.

c. Cosine Similarity Calculation

Using the vector representations, the semantic similarity is calculated by measuring the cosine similarity between the query vectors and the document vectors (Amrizal, 2018). This step evaluates the alignment of semantic meaning, with higher cosine similarity values indicating closer contextual relevance between the query and the document. This is calculated using Equation (3).

$$SemSim\left(Q, d_j\right) = \frac{\sum\left(vecw_Q \,.\, vecw_{id_j}\right)}{\sum |Q|_{term}{}^2 + \sum |d_j|_{term}{}^2} \ldots\ldots\ldots (3)$$

d. Ranking Semantic Similarity Scores:

The final step involves ranking the semantic similarity scores for each query-document pair. The highest scores indicate the documents most semantically relevant to the query.

This approach leverages the power of pretrained word embeddings to capture the deeper meanings of words and their contextual relationships, resulting in a sophisticated mechanism for retrieving relevant Quranic verse translations based on the query's intent. The use of cosine similarity ensures a consistent and normalized comparison, facilitating accurate ranking of results.

Performance Evaluation

The evaluation of search performance will be conducted based on the top 5, 10, and 15 retrieved results using the metrics of recall, precision, and F-measure. These metrics are defined as follows:

a. Recall

Recall measures the proportion of relevant verses that were successfully retrieved by the search system out of all relevant verses present in the document (Mhawi et al., 2022). It evaluates the completeness of the retrieval process, emphasizing how well the system identifies all relevant results. The calculation of recall is represented by Equation 4, where the numerator captures the relevant verses retrieved, and the denominator represents the total number of relevant verses available in the dataset.

$$R = \frac{TP}{(TP + FN)} \ldots\ldots\ldots (4)$$

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

b. Precision

Precision assesses the proportion of retrieved verses that are actually relevant to the search topic (Pardede & Pamungkas, 2023). It focuses on the accuracy of the results by determining the ratio of relevant documents to all documents retrieved by the system. Precision ensures that the system does not return irrelevant results, enhancing the relevance of the search output. The calculation of precision is represented by Equation 5.

$$P = \frac{TP}{(TP + FP)} \dots\dots\dots (5)$$

c. F-Measure:

The F-Measure combines precision and recall into a single metric, providing a balanced view of the retrieval system's performance (Anggraini et al., 2022). It is calculated as the harmonic mean of precision and recall, ensuring equal weighting of both metrics to account for trade-offs between them. The formula for F-Measure is shown in Equation 6.

$$F - measure = \frac{2 \,.R\,.\, P}{(R + P)} \dots\dots\dots (6)$$

By analyzing the top results (5, 10, and 15), these metrics enable a comprehensive evaluation of the search system's effectiveness in retrieving relevant Quranic verses while minimizing irrelevant results. This approach ensures a robust assessment of the performance of lexical and semantic similarity-based retrieval methods.

RESULT

Result of Data Collection

The data collection process involved downloading datasets from the publicly available repository at https://github.com/renomureza/quran-api-id. These datasets were obtained through a crawling process from the official website of the Quran maintained by Indonesia's Ministry of Religious Affairs. Crawling is an automated technique used to collect information from sources such as web pages by following links and extracting relevant data. In this context, the dataset was crawled directly from the Ministry's official Quran website, ensuring that the data reflects the content provided on the website. The dataset is presented in JSON format, containing structured array data that includes all 114 surahs and a total of 6,236 ayahs (verses). Structured array data refers to an organized representation of information stored systematically in computer memory. Each data element (e.g., surah and ayah) is arranged in a well-defined position, enabling efficient and systematic access to individual elements within the dataset. Each surah is supplemented with detailed metadata, including the surah number, name, category (Meccan or Medinan), ayah number in the Quran, ayah number within the surah, Arabic text of the ayah, Indonesian translation, brief interpretation (tafsir) from the Ministry of Religious Affairs, tafsir Jalalain, tafsir Al-Misbah (by Quraish Shihab), audio, and an Arabic text image. This comprehensive dataset provides a rich source of Quranic information for further analysis. The data is subsequently processed during the preprocessing stage and stored in a SQL database to facilitate efficient querying and retrieval. This ensures that the dataset is ready for use in lexical and semantic similarity measurements, enabling accurate and meaningful retrieval of Quranic translations.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Result of Data Preprocessing

The implementation of document preprocessing is based on the previously described stages of query and document preprocessing, including tokenization, stopword removal, and stemming. Document preprocessing is performed only once, processing the JSON-formatted dataset collected earlier. The results of the preprocessing are stored in a SQL database using the sqlite3 library in the Python programming language. Tokenization involves splitting the text into smaller units, or tokens, which facilitates subsequent processing steps. Stopword removal eliminates commonly used words that do not contribute significant meaning to the context, such as conjunctions and prepositions, thereby simplifying the dataset. Stemming reduces words to their base or root forms, ensuring that words with similar meanings are treated uniformly during further analysis. Once these preprocessing steps are completed, the cleaned and structured data is stored in the SQL database. This storage method enables efficient data retrieval and management, ensuring that the processed dataset is ready for further analysis or application in similarity measurements. An example of the program's output when the preprocessing steps are executed is illustrated in Table 1, showcasing the effectiveness of the preprocessing pipeline in transforming raw textual data into a structured and analyzable format.

Table 1. Sampel of Preprocessing Result

Query	**Before Preprocessing**	**After Preprocessing**
Short query	*Hati bersih*	*['hati', 'bersih']*
	Doa istiqamah	*['doa', 'istiqamah']*
	Sabar musibah	*['sabar', 'musibah']*
	Kebaikan tetangga	*['baik', 'tetangga']*
	Tawakal Allah	*['tawakal', 'allah']*
Long query	*Ayat Al-Qur'an tentang bagaimana bersyukur kepada Allah SWT*	*['ayat', 'alquran', 'syukur', 'allah', 'swt']*
	Kisah Nabi Yusuf dan istri potifar dalam Al-Qur'an	*['kisah', 'nabi', 'yusuf', 'istri', 'potifar', 'alquran']*
	Ajaran Al-Qur'an mengenai kesabaran dalam menghadapi cobaan	*['ajar', 'alquran', 'sabar', 'hadap', 'coba']*
	Penjelasan Al-Qur'an mengenai hukum riba dan perjudian	*['jelas', 'alquran', 'hukum', 'riba', 'judi']*
	Referensi Al-Qur'an tentang akhlak mulia dan tata krama dalam kehidupan sehari-hari	*['referensi', 'alquran', 'akhlak', 'mulia', 'tata', 'krama', 'hidup', 'hari', 'hari']*

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

The results of document preprocessing are organized and stored in a structured SQL database with two interconnected tables: "ayahs" and "surah." This database design ensures systematic data storage and facilitates efficient retrieval for analysis. The "ayahs" table is dedicated to detailed information about individual Quranic verses. It includes eight columns: an id serving as a unique identifier for each ayah, a surah_id linking each ayah to its corresponding surah in the "surah" table, and numberInQuran and numberInSurah indicating the position of the ayah in the entire Quran and within its specific surah, respectively. The arabic column stores the original Arabic text of the verse, while the preprocessed column contains the processed text of the Indonesian translation after applying techniques such as tokenization, stopword removal, and stemming. Additionally, the translation column provides the full Indonesian translation of the verse, and the tafsir column includes a concise interpretation from the Ministry of Religious Affairs. The "surah" table summarizes metadata for each chapter of the Quran. Represented in Figure 5. It consists of five columns: id, which links to the "ayahs" table via the surah_id, the name of the surah, its translation (the meaning of the surah's name), revelation category (Meccan or Medinan), and numberOfAyahs, indicating the total number of verses in the surah. This relational structure enables seamless linkage between the detailed verse-level data in the "ayahs" table and the higher-level chapter information in the "surah" table. The preprocessed Indonesian translations stored in the "preprocessed" column are critical for subsequent similarity computations, supporting lexical and semantic analyses. Meanwhile, the organized structure ensures that both verse-specific details and chapter-level summaries can be efficiently accessed, enhancing the database's utility for research and application purposes.

	id	name	translation	revelation	numberOfAyahs
	Filter	Filter	Filter	Filter	Filter
1	1	Al-Fatihah	Pembukaan	Makkiyah	7
2	2	Al-Baqarah	Sapi	Madaniyah	286
3	3	Ali 'Imran	Keluarga Imran	Madaniyah	200
4	4	An-Nisa'	Wanita	Madaniyah	176
5	5	Al-Ma'idah	Hidangan	Madaniyah	120
6	6	Al-An'am	Binatang Ternak	Makkiyah	165
7	7	Al-A'raf	Tempat Tertinggi	Makkiyah	206
8	8	Al-Anfal	Rampasan Perang	Madaniyah	75
9	9	At-Taubah	Pengampunan	Madaniyah	129
10	10	Yunus	Yunus	Makkiyah	109
11	11	Hud	Hud	Makkiyah	123
12	12	Yusuf	Yusuf	Makkiyah	111
13	13	Ar-Ra'd	Guruh	Madaniyah	43
14	14	Ibrahim	Ibrahim	Makkiyah	52
15	15	Al-Hijr	Hijr	Makkiyah	99
16	16	An-Nahl	Lebah	Makkiyah	128
17	17	Al-Isra'	Memperjalankan Malam Hari	Makkiyah	111

1 - 18 of 114 Go to: 1

Figure 5. Result of Preprocessing Data in Surah Table

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Result of Lexical Based Searching



Figure 6. Interface of Lexical Based Searching

The results of lexical-based searching display information about Quranic verse retrieval based on a method that focuses on the similarity of word forms and structures in the text. In this process, after a user inputs a query or specific keywords, the digital Quran application searches for verses with lexical similarities to the query. The relevant verses are presented in a list format. Each entry in the list includes the surah name, the verse number within the surah, the Arabic text of the verse, and its translation in Indonesian. Additionally, the results feature a similarity score for each verse. This score quantifies the degree of lexical similarity between the verses and the user's query. A higher similarity score indicates that the verse is more closely aligned with the entered query. This provides users with a clear metric to evaluate the relevance of the search results. The user-friendly presentation of this information ensures that users can easily identify and assess the verses that match their queries. The layout of these results, along with the imilarity scores, is illustrated in Figure 6, highlighting the system's capability to deliver structured and detailed feedback for lexical-based searches.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Result of Lexical Based Searching



Figure 7. Interface of Semantic Based Searching

The results of semantic-based searching provide users with information about Quranic verse retrieval using a method focused on the similarity of meaning within the verse text. Relevant verses identified through this approach are presented in a list format. Each entry includes the surah name, the verse number within the surah, the Arabic text of the verse, and its translation in Indonesian. Additionally, the results are accompanied by a similarity score. This score indicates the degree of semantic similarity between the verses and the user's query. A higher similarity score reflects a closer match in meaning between the verses and the input query, allowing users to better understand the relevance of the results to their search intent. This approach emphasizes the importance of capturing the context and deeper meaning of the text, which goes beyond surface-level word matching. The presentation of semantic-based search results, along with their corresponding similarity scores, ensures a comprehensive and meaningful experience for the user. Figure 7 illustrates the layout of these results, demonstrating the system's ability to prioritize semantic understanding in Quranic verse retrieval.

Performace Evaluation

The evaluation phase of this study involved testing the implementation of lexical and semantic similarity measurements using the confusion matrix method to calculate the accuracy of text relevance retrieval in Quranic verse translations. The evaluation aimed to assess the system's effectiveness in retrieving relevant documents. The experiment utilized a dataset of 6,236 Quranic verses. The queries used for testing were derived from the thematic interpretation chapter titled "Social Responsibility" by Pustaka Lajnah, Ministry of Religious Affairs. A total of 20 queries were designed, as outlined in Table 2, based on thematic insights from the interpretation chapter. These queries were utilized to compare the retrieval results

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

generated by the application of lexical similarity, semantic similarity, and a combination of both methods. The purpose of this evaluation was to analyze how well each method performs in identifying relevant verses in the context of individual social responsibility, as highlighted in the thematic interpretation. By focusing on this thematic context, the evaluation provides insights into the relative strengths and weaknesses of lexical and semantic approaches in Quranic verse retrieval. This comparison further illustrates how these methods address nuanced interpretations of social responsibility within the verses.

Table 2. Queries Used for Quranic Verse Translation Retrieval

Searching	Daftar *query yang digunakan*
Lexical-Based	*1. umat islam, 2. bukti-bukti yang nyata, dapat berlaku adil. 3. Bertanggung Jawab, Apa yang telah dilakukannya, 4. Kebajikan, Kejahatan, 5. dapat memikul dosa, 6. orang-orang yang zalim, bahwa allah sangat keras siksanya, 7. Yang maha pengasih, pada hari kiamat, 8. Setiap umat berlutut, buku catatan amalnya, 9. orang-orang yang beriman, sembahlah tuhanmu, berbuatlah kebaikan. 10. memohon ampunan, 11. melaksanakan salat, memberi makan orang miskin, 12. melaksanakan salat, 13. biarlah dia kafir, 14. mengerjakan kebajikan, mengerjakan kejahatan, 16. api neraka, 17. pelindung bagi perempuan, khawatirkan akan nusyuz, 18. perhiasan kehidupan dunia, 19. Sembahlah allah, 20. berbuat-baiklah kepada kedua orang tua.*
Semantic-Based	*3. Apa yang telah dilakukannya, 4. Kebajikan, Kejahatan, 6. peliharalah dirimu dari siksaan yang tidak hanya menimpa, bahwa allah sangat keras siksanya, 7. menghitung mereka dengan hitungan yang teliti.8. Setiap umat berlutut, buku catatan amalnya, 9 rukuklah. 10. mereka sedikit sekali tidur pada waktu malam, harta benda mereka ada hak untuk orang miskin, orang miskin yang tidak meminta, 11. untuk tujuan yang batil 12. mereka yang tetap setia melaksanakan salatnya, hartanya disiapkan bagian tertentu, 14. mengerjakan kejahatan, 15. Kami telah memuliakan anak cucu Adam*

The experiment also evaluated the results for the top 5, 10, and 15 retrieved documents to provide a more granular analysis of the system's performance. In addition to testing across all queries, the evaluation considered the word frequency of queries by dividing them into two categories: short queries (queries containing two or fewer words) and long queries (queries containing more than two words). This categorization aimed to highlight the importance of query structure in document retrieval performance. By analyzing both short and long queries, the study assessed the system's ability to handle variations in query complexity and length. This approach provides deeper insights into how different query formats influence the relevance of retrieved documents. The retrieved document evaluation results for these scenarios are summarized in Table 3, illustrating the effectiveness of the lexical and semantic similarity measurements across diverse query types and document ranking criteria.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Table 3. Evaluation Metrics for Lexical and Semantic Appraches

Searching	*Confusion Matrix*	*Top 5*	*Top 10*	*Top 15*
Lexical-Based	*Precision*	100%	100%	100%
	Recall	46%	50%	58%
	F-1 Score	63%	66%	73%
Semantic-Based	*Precision*	100%	100%	100%
	Recall	58%	56%	59%
	F-1 Score	73%	71%	74%

Table 3 presents the evaluation metrics of lexical-based and semantic-based similarity measurements in the retrieval of Quranic verse translations. The evaluation is conducted for the top 5, 10, and 15 retrieved results using three metrics: Precision, Recall, and F1-Score. These metrics offer insight into the accuracy and relevance of the retrieval system for both approaches. The lexical-based approach achieves 100% precision across all top results (5, 10, and 15). This indicates that every retrieved document in the top results is relevant to the user query. However, the recall scores are comparatively lower, starting at 46% for the top 5 results, increasing to 50% for the top 10, and reaching 58% for the top 15. This reflects that while the retrieved results are accurate, the system fails to retrieve a significant portion of the relevant documents. Consequently, the F1-Score improves with more retrieved results, rising from 63% for the top 5 to 73% for the top 15. Similarly, the semantic-based approach achieves 100% precision across all top results, showcasing its ability to ensure relevance in retrieved documents. However, the recall scores are slightly higher than the lexical-based approach, with 58% for the top 5, 56% for the top 10, and 59% for the top 15. This improvement in recall demonstrates the semantic approach's ability to identify more relevant documents compared to the lexical method. The F1-Score also follows this trend, with scores ranging from 73% for the top 5 to 74% for the top 15.

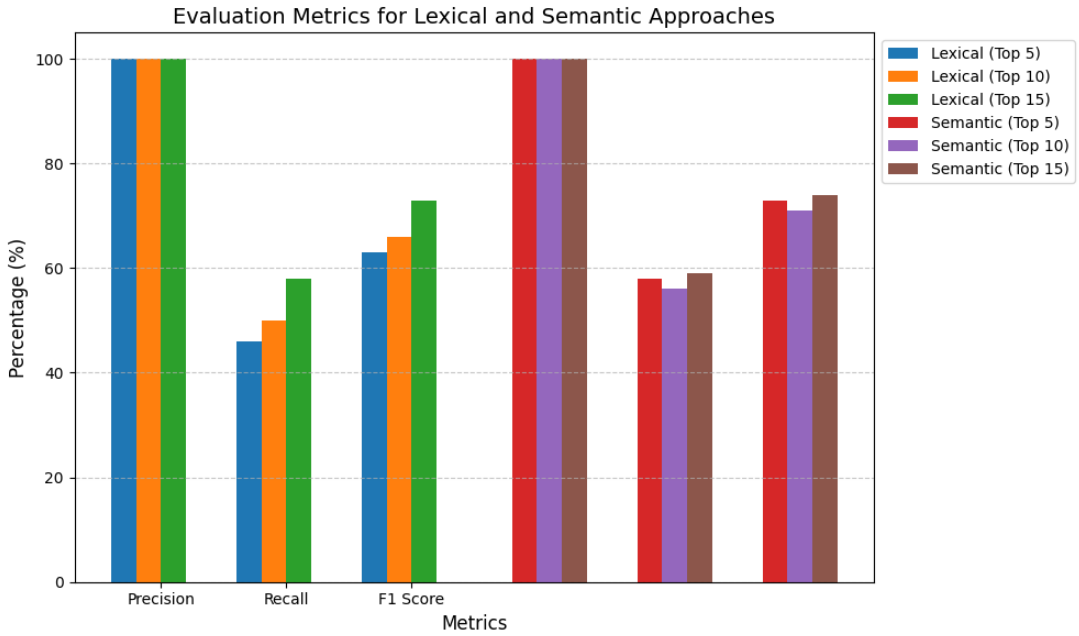


Figure 8. Graphic of Evaluation Metrics for Lexical and Semantic Appraches

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Based on Figure 8, while both approaches deliver perfect precision, the semantic-based approach consistently outperforms the lexical-based method in recall and F1-Score. This indicates that the semantic approach retrieves a greater proportion of relevant documents while maintaining accuracy. The higher F1-Score for semantic-based measurements reflects its balanced performance in ensuring both precision and recall, especially for queries requiring an understanding of contextual meaning. The semantic-based approach demonstrates a superior ability to capture nuanced meanings in queries and documents, making it a more effective method for Quranic verse retrieval. However, both methods exhibit strengths that can be complementary, emphasizing the potential of combining these approaches for optimal retrieval performance.

DISCUSSIONS

In this The results of the evaluation provide an insightful comparison of the effectiveness of lexical and semantic approaches in retrieving relevant Quranic verses based on their translations. The discussion focuses on the performance metrics of precision, recall, and F1-score for the top 5, 10, and 15 retrieved results, as presented in Table 3 and its corresponding graphical representation. The lexical-based approach demonstrates perfect precision (100%) across all tested scenarios (Top 5, 10, and 15). This result indicates that every retrieved verse in these cases was relevant to the user query. However, recall scores are notably lower, starting at 46% for the top 5 results and gradually improving to 58% for the top 15. The relatively low recall suggests that while the method retrieves only relevant verses, it misses a substantial portion of the relevant dataset. Consequently, the F1-score, which balances precision and recall, ranges from 63% to 73%, showing modest improvements as more results are retrieved. This indicates the lexical approach's limited ability to generalize beyond exact word matches, restricting its effectiveness in more nuanced queries. In contrast, the semantic-based approach shows equally high precision (100%) but outperforms the lexical approach in recall, starting at 58% for the top 5 results and slightly fluctuating around 56%-59% for the top 10 and 15 results. The higher recall demonstrates the semantic approach's ability to identify relevant verses even when they do not share exact lexical matches with the query. This advantage is reflected in the F1-score, which is consistently higher than that of the lexical approach, ranging from 73% to 74%. These results highlight the semantic approach's strength in understanding context and meaning, enabling it to retrieve verses with greater interpretative relevance. The findings emphasize the importance of semantic similarity in addressing the limitations of lexical matching, particularly in contexts requiring deeper comprehension of textual meanings. While semantic approaches provide a more robust retrieval framework, their computational complexity must be balanced against the requirements of real-time applications. Future research could explore hybrid methods to leverage the strengths of both approaches, optimizing precision and recall simultaneously.

CONCLUSION

This research highlights the strengths and limitations of lexical and semantic similarity approaches in identifying relevant Quranic verses based on user queries. The study provides valuable insights into the effectiveness of both methods through the analysis of performance metrics such as precision, recall, and F1-score for various query scenarios. The research concludes that while the lexical approach is effective for retrieving exact matches and ensuring high precision, it struggles with recall due to its inability to interpret semantic variations. On the other hand, the semantic approach offers better balance, capturing deeper contextual relevance and retrieving a broader set of relevant verses. The findings underscore the need for a more nuanced retrieval system that can combine the precision of lexical methods with the contextual understanding of semantic approaches. This hybrid methodology could provide a more comprehensive solution for Quranic verse retrieval, enhancing the accessibility and usability of Quranic

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

content for diverse user queries. Future studies should focus on integrating these approaches and addressing computational challenges for real-world implementation.

ACKNOWLEDGMENT

The researchers express their deepest gratitude to the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for their financial support of this research. This funding has been instrumental in enabling this study to contribute to advancements in Quranic information retrieval systems. We also extend our heartfelt thanks to Universitas Nahdlatul Ulama Blitar for their unwavering support throughout the research process.

REFERENCES

Abriani, G. U., & Yaqin, M. A. (2019). Analisis Implementasi Metode Semantic Similarity untuk Pengukuran Kemiripan Makna antar Kalimat. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *1*(2), 47–57. https://doi.org/10.28926/ilkomnika.v1i2.15

Afzal, H., & Mukhtar, T. (2019). Semantically Enhanced Concept Search of the Holy Quran: Qur'anic English WordNet. *Arabian Journal for Science and Engineering*, *44*(4), 3953–3966. https://doi.org/10.1007/s13369-018-03709-2

Ahmad, R., Khan, F. Z., & Khan, M. A. (2021). Ontology Based Knowledge Retrieval and Semantic Modelling of Qur'an with Contextual Information. *International Journal on Islamic Applications in Computer Science And Technology*, *9*(1), 10–25.

Amrizal, V. (2018). Penerapan Metode Term Frequency Inverse Document Frequency (Tf-Idf) Dan Cosine Similarity Pada Sistem Temu Kembali Informasi Untuk Mengetahui Syarah Hadits Berbasis Web (Studi Kasus: Hadits Shahih Bukhari-Muslim). *Jurnal Teknik Informatika*, *11*(2), 149–164. https://doi.org/10.15408/jti.v11i2.8623

Anggraini, S., Purwitasari, D., & Arifin, A. Z. (2022). Pengukuran Kemiripan berbasis Leksikal dan Semantik untuk Perangkingan Dokumen Berbahasa Arab. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *4*(2), 134–145. https://doi.org/10.28926/ilkomnika.v4i2.495

Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah, & Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1), 217–227. https://doi.org/10.12928/TELKOMNIKA.V18I1.14874

Goddard, C., & Schalley, A. C. (2010). Semantic Analysis. In *Handbook of Natural Language Processing* (2nd Editio). CRC Press.

Gunay, A., & Yolum, P. (2007). Structural and Semantic Similarity Metrics for Web Service Matchmaking. *Lecture Notes in Computer Science (LNCS)*, 129–138.

Husni, H., & Arifin, B. (2019). Rancangan Dan Implementasi Aplikasi Pencarian Teks Terjemahan Al Qur'an Berbasis Model Ruang Vektor. *Jurnal Simantec*, *7*(2), 90–96. https://doi.org/10.21107/simantec.v7i2.7224

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

Journal of Computer Networks, Architecture and High Performance Computing
Volume 7, Number 1, Januari 2025
https://doi.org/10.47709/cnahpc.v7i1.5194

Submitted : Dec 29, 2024
Accepted : Jan 14, 2025
Published : Jan 17, 2025

Kesuma, H. W. A. (2016). *Penerapan Metode TF-IDF dan Cosine Similarity dalam Aplikasi Kitab Undang-Undang Hukum Dagang*.

Kuzi, S., Zhang, M., Li, C., Bendersky, M., & Najork, M. (2020). Leveraging Semantic and Lexical Matching to Improve the Recall of Document Retrieval Systems: A Hybrid Approach. In *Proceedings of ACM Conference (Conference'17)* (Vol. 1, Issue 1). Association for Computing Machinery.

Mhawi, D. N., Oleiwi, H. W., Saeed, N. H., & Al-Taie, H. L. (2022). An Efficient Information Retrieval System Using Evolutionary Algorithms. *Network*, *2*(4), 583–605. https://doi.org/10.3390/network2040034

Mohamed, E. H., & Shokry, E. M. (2022). QSST: A Quranic Semantic Search Tool based on word embedding. *Journal of King Saud University - Computer and Information Sciences*, *34*(3), 934–945. https://doi.org/10.1016/j.jksuci.2020.01.004

Mohd, M., Jan, R., & Shah, M. (2020). Text document summarization using word embedding. *Expert Systems with Applications*, *143*. https://doi.org/10.1016/j.eswa.2019.112958

Naf'an, M. Z., Sari, Y., & Suyanto, Y. (2021). Word Embeddings Evaluation on Indonesian Translation of AI-Quran and Hadiths. *IOP Conference Series: Materials Science and Engineering*, *1077*(1), 012025. https://doi.org/10.1088/1757-899x/1077/1/012025

Pardede, J., & Pamungkas, D. P. (2023). The Impact of Balanced Data Techniques on Classification Model Performance. *Scientific Journal of Informatics*, *11*(2), 401–412. https://doi.org/10.15294/sji.v11i2.3649

Puspitasari, D. A. (2022). Kebijakan Pentashihan Aplikasi Al-Qur'an Digital di Indonesia: Studi Perkembangan Aplikasi "Al-Quran Kementerian Agama" dan Permasalahannya. *J-PAI: Jurnal Pendidikan Agama Islam*, *8*(1), 12–22. https://doi.org/10.18860/jpai.v8i1.13425

Rofiqi, M. A., Fauzan, A. C., Agustin, A. P., & Saputra, A. A. (2019). Implementasi Term-Frequency Inverse Document Frequency (TF- IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query. *Jurnal Computer Science and Applied Informatics*, *1*(2), 58–64.

Rouf Abd. M., & Fauzan Charis Abd. (2023). Relevansi Ayat al-Quran Secara Tematik Menggunakan Pendekatan Graph-Based Knowledgedan Lexical-Search. *ILKOMNIKA: Journal of Computer Science and Applied Informatics*, *5*(1), 96–104.

* Corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).