Searching Sahih Hadiths Based on Queries using Neural Models and FastText

Sari Susanti[1], Ina Najiyah[2,*], Yudi Ramdhani[3], Asti Herliana[4],
Masaldi Kharisma Muckti[5], Fani Rahma Oktaviani[6]

[1,2,4,5,6]*Information Systems Study Program, Faculty of Information Technology, Adhirajasa Reswara Sanjaya University, Bandung, Indonesia*

[3]*Department of Information Systems at the Faculty of Creative Technology, Satu University, Bandung 40253, Indonesia*

(Received: June 13, 2024; Revised: September 22, 2024; Accepted: October 12, 2024; Available online: December 29, 2024)

Abstract

Hadith is the second source of Islamic law after the Qur'an, and the availability of accurate and easily accessible information about hadith is crucial, as it directly affects a person's belief (aqidah). This highlights the importance of having hadith collections as essential guidance in everyday life. Today, digital versions of hadiths are available in various applications, e-books, and websites. However, users often complain that these sources are incomplete and do not contain the entire collection of the Prophet's hadiths from al-Kutub as-Sittah. Additionally, the complex presentation of these digital resources makes it difficult to find relevant hadiths efficiently. This study aims to improve access to accurate and relevant hadith information, focusing specifically on al-Kutub as-Sittah, using Information Retrieval systems that search for hadiths based on keywords. IR is employed because it has proven effective in retrieving precise documents according to the search terms. A Neural Network is used to match user queries with the document collection, while FastText word embedding is implemented for text representation. FastText is particularly useful for detecting similar meanings across different words, which is essential when interpreting Indonesian-translated hadiths that require nuanced understanding. The dataset used in this study consists of 31,275 Indonesian-translated hadiths from al-Kutub as-Sittah. In this study, it was found that many hadith translations have ancient language so that query reformulation is needed to get the right hadith because users often enter commands with currently trending words. In this study, it was also found that word2vec has less performance than FastText in weighting words in hadith translations. The results indicate that the neural network performs well in retrieving relevant hadith content according to the user's commands or keywords. With a training data proportion of 70% and a testing data proportion of 30%, the Recall value was 0.7721 and the Precision value was 0.75112.

Keywords: Information Retrieval, Hadith, Neural Network' Kutubus Sittah, Text Mining, Word Embedding

1. Introduction

During the time of the Prophet Muhammad, hadiths were not compiled into books; the Prophet's companions relied on memorization and simple writings when seeking a ruling in Islamic law [1]. After the Prophet Muhammad's passing, the compilation of hadiths began due to fears among the caliphs and the tabi'in (followers of the companions) that the hadiths would eventually be lost and forgotten. This led to the compilation of hadiths during the caliphate of Umar ibn Abdul Aziz [2] . Kutub As-Sittah refers to the six books written by six Hadith Imams: Imam Bukhari (194-252 AH), Imam Muslim (204-261 AH), Abu Dawud (202-275 AH), al-Nasa'i (215-303 AH), al-Tirmidhi (200-279 AH), and ibn Majah (207-273 AH) [3]. After the end of the caliphate era, many weak (da'if) hadiths began to appear, whose Sahihity was questionable.

The emergence of these weak hadiths could lead to a decline in faith [4], religious deviations, misguidance in religion [5] and misunderstandings in religious practice [6] This situation arises because Muslims only accept hadiths from lectures or words from people around them without knowing and studying the hadith directly in its original form. In determining weak hadiths, it is also necessary to study in depth who the narrator is, and what the meaning of the hadith is. This is also because Muslims today often cannot identify the authenticity of hadiths because of the many authentic

*Corresponding author: Ina Najiyah (inajiyah@ars.ac.id)
DOI: https://doi.org/10.47738/jads.v6i1.567

hadiths and the limited sources that can be easily accessed to search for hadiths. Therefore, there is a need for authentic hadiths to be presented in an easily accessible manner so that Muslims can easily find them. Sources providing hadiths in digital or physical form can be found by Muslims in the form of applications, websites, and books.

However, these sources often present hadiths in a manner that is too broad and overwhelming. This research will apply the field of Text Mining to assist in the retrieval of hadith information using a dataset from the Kitab As-sittah. Text mining employs information extraction techniques to retrieve information from existing data. It also utilizes Natural Language Processing (NLP) techniques to automatically process data based on the information provided by the text. NLP is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language, even when dealing with unstructured data, such as hadiths, which often contain unstructured language [7], [8], utilizing Indonesian translations of the hadiths [9]. The information retrieval process involves finding documents that match a given command or query, commonly referred to as Information Retrieval [10] (IR) IR can be used in the processes of data collection and document retrieval [11] ensuring that the information sought aligns with the user's needs or keywords. The retrieval process of hadiths from Indonesian text poses its own challenges, as different words may have the same meaning. Therefore, the field of Tafsir (interpretation) contributes to the text weighting or representation process [12].

The performance of information retrieval can be significantly influenced by the text representation process. The better the system understands the text being searched for based on the command, the better its performance, and the more accurate the document retrieval will be [13]. This research employs the FastText Word Embedding model for the representation and vector search process, as FastText performs better than the Word2Vec model when dealing with text data that has the same meaning despite using different words [14]. Furthermore, a Neural Network model is used for matching queries with hadith documents in the dataset, as this model has proven successful in various previous research cases [15] and performs well in ranking [16] such as in research that retrieves documents based on keywords in legal books [17] Some Neural models, like the bag-of-words model, have shown a 10% better performance compared to other models [18].

Based on the previous explanations, this research aims to develop a hadith search system that responds to user conversations or commands. The system will first undergo training and model design using Neural Network and FastText algorithms to achieve optimal performance. The model design will focus on extracting training data based on keywords, specifically words or phrases found within the hadiths. Once the relevant hadith keywords are identified, a ranking process will follow to determine which hadiths are most relevant. Using the proposed model and available dataset, the results of this research show that the system performs well in retrieving relevant documents according to the user's commands or queries.

2. Literature Review

2.1. Hadith Search Engine

A hadith search engine using a dataset of 380 translated hadiths from Sahih Bukhari, utilizing the Vector Space Model (VSM) for matching hadith documents with queries, has been developed before. Additionally, in the representation process, this research employs Weighted Inverse Document Frequency (WIDF) to determine the frequency of each word in the documents. The proposed model yielded satisfactory performance as the text representation process successfully maintained the meaning of the text [19]. A summary of 48 journals on hadith classification and retrieval concluded that research trends on hadith mostly share the same language and corpus, with the most common topic being classification at 33.33%, the most used language being Arabic at 43.75%, and the most used algorithm being SVM at 12.5%. Furthermore, the most commonly used dataset is the public dataset by Al-Bukhari at 30.53% [20].

Another study successfully performed hadith retrieval based on queries using the Greedy Algorithm, utilizing a dataset of 7,500 hadiths from 97 books and conducting experiments with 4,000 hadiths on 3,825 queries. This research achieved good performance with the Greedy model compared to other discussed methods, reaching a precision rate of 83% [21]. There is research comparing several Neural Network models such as RNN and CNN using five different datasets to evaluate the performance of Neural Networks with various approaches. The concept used is Question and Answer, adopting the field of Information Retrieval. The results of this research showed that the Neural Network model concept is very effective, producing good performance despite different approaches compared to other methods [22]. In the document matching process, it is necessary to find words with similar meanings to improve performance. Therefore, this research uses the FastText Word Embedding, which performs better than Word2Vec, as shown in a study comparing these two models in information retrieval on Sirah Nabawiyyah. The study indicated that FastText had better performance and increased document relevance [23]. The novelty of our research, compared to previous studies, is clearly evident in several aspects, starting with the increased size of the dataset. Additionally, the proposed and implemented model in this study is more relevant and better suited to the dataset used.

3. The Proposed Method

This research relies on information retrieval technology and adopts Neural Networks for ranking and matching documents to the given queries, commonly known as the document matching process. The process is divided into two parts: the preparation of hadith documents and the query search process based on the user's desired information. The complete stages are shown in figure 1.

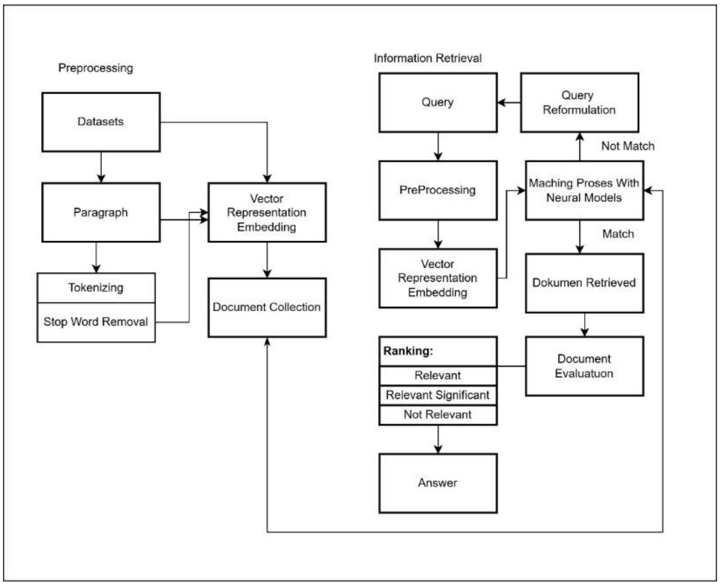


Figure 1. Research Method

Figure 1 illustrates all the stages conducted in this research, from model development to the ranking process. The stages are briefly summarized as follows: the dataset will be divided into two different parts to be processed either word by word or sentence by sentence. This is done because the model requires word weights from the hadith and document weights. Next, both documents and words undergo a preprocessing phase to obtain more prepared and refined "features." Subsequently, during IR process, document searches will be conducted based on the queries input by the user, along with a query reformulation process in case of null searches. A more complete explanation of the image in figure 1 is explained in the following sub-chapter:

3.1. Preprocessing

This stage is carried out before the information retrieval process. It involves discussing the dataset size, subsequent stages of the dataset, preparation processes such as tokenizing and stopword removal, and the process of weighting words into vectors using FastText word embedding. Subsequently, the dataset will be divided into two different columns, with the final result being the indexing or weighting of documents and words stored in the data collection.

3.2. Dataset Hadith

The data used consists of hadiths in both Indonesian and Arabic. For the information retrieval process, only the Indonesian translations of the hadiths are used, as the system is designed with a focus on Indonesia. However, the Arabic texts of the hadiths are stored in the documents to be displayed when users search for hadith information. The Kutub As-Sittah contains six hadith collections with a total of 9,241 hadiths in the Sahih Bukhari, 3,030 hadiths in

Sahih Muslim, 5,274 hadiths in Sunan Abu Dawud, 3,956 hadiths in Jami' at-Tirmidhi, 5,774 hadiths in Sunan an-Nasa'i, and 4,000 hadiths in Sunan Ibn Majah, amounting to a total dataset of 31,275 hadiths. Each book has discussion chapters such as prayer chapters, zakat chapters, shodaqoh and other chapters related to the rules set out. The detailed data used in this study is provided in table 1.

Table 1. Number of Datasets

No	Books	Number of Book	Number of Hadith	Number of Chapter
1	al-Jami’ al-Shahih al-Bukhari	97	9241 hadith	3450 chapter
2	al-Jami’ al-Shahih Muslim	81	3030 hadith	2411 chapter
3	Kitab as-Sunan Abu Dawud	35	5274 hadith	1871 chapter
4	Kitab Jami’ at-Tirmidziy	32	3956 hadith	50 chapter
5	Kitab as-Sunan an-Nasaa`i	41	5774 hadith	51 chapter
6	Kitab as-Sunan Ibn Majah	32	4000 hadith	1500 chapter

3.3. Paragraph

The dataset, consisting of a collection of documents, is processed in two different parts: document-level processing and paragraph-level processing. Both parts will undergo pre-processing, including tokenizing and stopword removal. This approach is taken because, in addition to requiring vector values and indices for each word in the document, indices or vectors for each document [24] or hadith verse are also needed to be retrieved when a user searches for information.

3.4. Vector Representation Embedding

The next process involves calculating word weights using Word Embedding to obtain vector values for each document. This stage is divided into two parts with different datasets, similar to the tokenizing and stopword removal processes. The first part is the vector weight calculation for the hadith dataset, and the second part is the vector weight calculation for the user-provided query. This method is employed because, in the hadith translation dataset, many words differ but have the same meaning. Thus, the embedding process will seek the distance or similarity in meaning of these words or find similarities among related and semantically equivalent words [25]. Keywords obtained from the query will then undergo text representation to generate vector values that can be processed by the computer before the Neural Network stage. In this study, text embedding uses the Dual Embedding Space Model (DESM) to convert text into vector form while also considering word similarity with other words of the same meaning [26].

3.5. Document Collection

The vector representations generated in the previous stage are then stored in the document collection as a Microsoft Excel file created using a Python program. This collection of documents will be used and compared with the queries provided by the user [24]. The data in this document collection serves as a reference and main database for the accurate retrieval of hadith information required by the user.

3.6. Information Retrieval Query

In the hadith information retrieval process, the user will enter a "query" or command based on what they are looking for, such as "show all hadiths about speaking ill of others" or the query "prohibition of gossip." The system will then use this input to search for relevant documents. The queries entered by the user are not limited in the number of words. Users can input just keywords, questions, or specific information retrieval commands.

3.7. Matching Proses With Neural Network

The final stage involves using the Neural Network model, which is chosen for its excellent performance with large datasets [27]. In the context of document ranking for information retrieval, this model offers advantages over other methods, such as RNN, because it does not have key limitations when checking each document. Unlike RNN, which checks only three to four words following the keywords or only three to four documents, the Neural Network does not have such limitations. It continuously searches for matches or the distance between the keywords and all documents in the dataset, resulting in more potential matches and relevant documents for subsequent ranking [28]. In this research, the Neural Network processes input vectors from the embedding of each word in the query and the word vectors in the dataset. These vectors are then combined using an interaction matrix with the Hadamard Product, followed by Fully Connected Layers for Matching, where the output is a score for each document to identify which is most relevant to the query [29].

In this process, the model does not only search for information by providing questions or keywords to positively-toned or positive-text documents. Instead, the process also involves checking and querying keywords in negative documents. This approach allows the model to learn from all aspects of the dataset to produce relevant documents accordingly [30]. The Neural Network process is illustrated in figure 2.

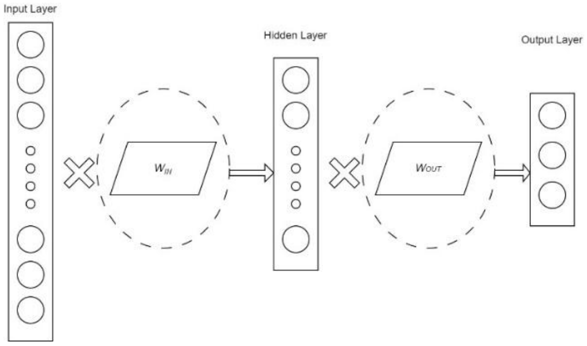


Figure 2. Neural Network Models

Figure 2 shows that in this stage, the input layer is determined by the vector results from the previous process, which are used as input data, and metric calculations are performed for each data input. The resulting metrics are then processed in the hidden layers to apply Dropout to unrelated data. After processing in the hidden layers, the output is categorized into three documents: relevant, not relevant, and significantly relevant. If the input text does not match any documents or if the command is unrecognized, the model will perform query reformulation.

3.8. Query Reformulation

The process of matching words in the query with the document collection during the training phase is carried out using a Neural Network. If the information from the query matches documents in the collection and the documents are found, the documents will be ranked based on how well they meet the user's information needs. However, if no documents are found, the query will be reformulated by seeking alternative meanings of the user-provided query [31]. For example, if the user inputs the query "show all hadiths about gibah," alternative phrases with the same meaning, such as "speaking ill" or "disclosing faults," will be searched, and the process will be repeated. The output of this process includes rankings of relevant documents, significantly relevant documents, and irrelevant documents. Once the most relevant documents are identified, they will be sent directly to the user as the response to the information sought.

3.9. Document Evaluation

The Recall and Precision values will be calculated to evaluate the model and determine how relevant the documents found in the search process are to the given query. Recall is calculated by comparing the number of documents retrieved to the total number of relevant documents in the dataset. Precision is calculated by comparing the number of relevant documents to the total number of items retrieved for ranking [32]. The formulas for precision and recall are as follows:

$$R = \frac{\text{number of relevants item retrieved}}{\text{Total number of relevant items in collection}} \tag{1}$$

$$P = \frac{\text{number of relevants item retrieved}}{\text{Total number of items retrieved}} \tag{2}$$

4. Results and Discussion

This section explains all stages and results depicted in figure 2, starting from sample dataset examples, the processes conducted before information retrieval, including preprocessing and its examples, followed by the application of the neural network in matching queries with hadith documents, and ending with the performance results of the proposed model. All explanations are provided and detailed using tables.

4.1. Preprocessing Datasets

From the dataset mentioned in table 1, there are 31,275 records in raw data form with some duplicates. Therefore, before performing information retrieval, data preprocessing is required [33]. Additionally, the original dataset contains three fields: the book name, hadith verse, and hadith translation. For this study, only the "hadith translation text" field is used. An additional column is added as an ID to distinguish each document. Thus, the field structure of the dataset used in this research is shown in table 2. In the "hadith translation" field, this study removes the "sanad" and retains only the "matan" because the sanad refers to the chain of narrators of the hadith, including the Prophet, companions, and tabi'in, and does not affect the hadith retrieval process [34]. Therefore, the "sanad" content is excluded. We have kept the data in the table and the examples in Indonesian, in accordance with the dataset used in this research.

Table 2. The Example of Hadith Datasets

ID	Hadith Translation Text
D1	"You who believe with your words that have not yet reached your hearts, do not annoy the Muslims, do not speak ill of them, do not look for their disgrace. Whoever looks for the faults of his fellow Muslims, Allah will surely look for his faults. Whoever Allah looks for his faults, Allah will reveal them even in his house." (HR. At Tirmidzi)
D2	We were with the Prophet when we suddenly smelled an unpleasant stench. Then Rasulullah saw. said, 'Do you know what this smell is? This is the smell of those who backbite (gossip) the believers."
D3	When I was migrating, I passed a people who had copper nails scratching their faces and chests. I asked: 'Who are these, O Gabriel? Jibril answered: They are people who eat human flesh (gibah) and insult their honor'. (HR. Abu Daud)

4.2. Paragraph

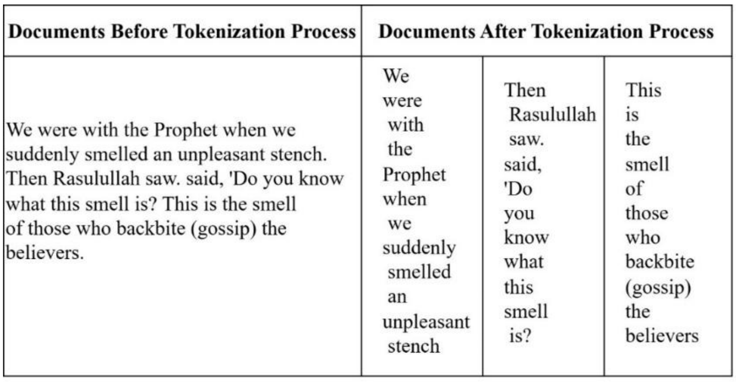
The original dataset contains two fields as described earlier. In this stage, an additional field is added to the dataset to store duplicates of the "hadith translation text" field. This is necessary because two different types of data are needed: one for direct indexing of the hadith translations and another for word-level indexing, which requires tokenization and stopword removal. A sample of the duplicated field is shown in table 3. The two columns will then be processed using different stages.

Table 3. Dataset Duplication (Paragraph)

ID	Hadith Translation Text	Duplicate of Hadith Translation Text
D1	We were with the Prophet when we suddenly smelled an unpleasant stench. Then Rasulullah saw. said, 'Do you know what this smell is? This is the smell of those who backbite (gossip) the believers."	We were with the Prophet when we suddenly smelled an unpleasant stench. Then Rasulullah saw. said, 'Do you know what this smell is? This is the smell of those who backbite (gossip) the believers."

4.3. Tokenizing and Stopword Removal

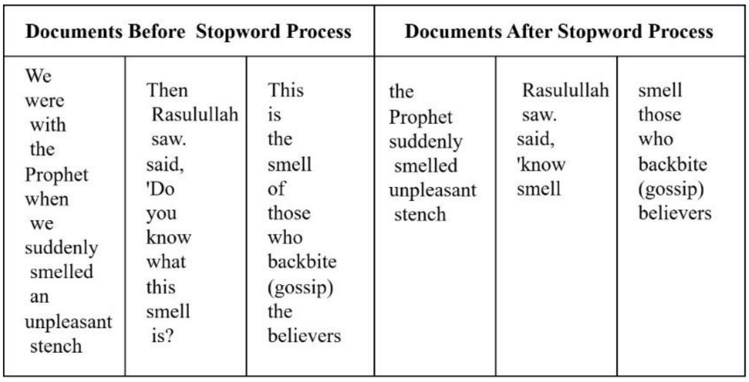
This stage is divided into two parts with two different datasets. The first part involves tokenizing and stopword removal for the hadith dataset, while the second part deals with the data provided by the user’s query. The dataset consists of complete hadith texts along with their Indonesian translations. In this case, the dataset is very large and impacts the model's performance during the search for relevant documents [35]. This stage aims to obtain the vector values for each word and document. Before the documents undergo training, text processing is performed in two stages: word-level processing and document-level vector search. The first process involves tokenization to separate each word based on delimiters, followed by stopword removal to eliminate words that do not affect the search, such as "and," "or," and "says." In the query content, this process removes words that are not considered "keywords" and common words in the provided query. Before this weighting is done, the dataset, which is still in paragraph or single-sentence form in the "Duplicate of Hadith Translation Text" column, needs to be prepared. First, this column needs to be split into individual words using a delimiter. In this case, the separation is based on space delimiters. An example of the tokenizing process is shown in Figure 3.



Documents Before Tokenization Process	Documents After Tokenization Process		
We were with the Prophet when we suddenly smelled an unpleasant stench. Then Rasulullah saw. said, 'Do you know what this smell is? This is the smell of those who backbite (gossip) the believers.	We were with the Prophet when we suddenly smelled an unpleasant stench	Then Rasulullah saw. said, 'Do you know what this smell is?	This is the smell of those who backbite (gossip) the believers

Figure 3. Duplicate Paragraph

Stopword removal is then performed to eliminate words that are less meaningful and almost always present in every document in the dataset, such as "and," "or," "for," and "has." In this study, stopword removal is carried out using the Sastrawi library. This process uses the column resulting from tokenization in Figure 3. A sample of the dataset after stopword removal is shown in Figure 4.



Documents Before Stopword Process			Documents After Stopword Process		
We were with the Prophet when we suddenly smelled an unpleasant stench	Then Rasulullah saw. said, 'Do you know what this smell is?	This is the smell of those who backbite (gossip) the believers	the Prophet suddenly smelled unpleasant stench	Rasulullah saw. said, 'know smell	smell those who backbite (gossip) believers

Figure 4. Documents After Stopword Removal Process

4.4. Vector Representation Embedding

The dataset, which includes both hadith documents and queries, undergoes a weight of words search process that significantly impacts the model's performance. In this study, embedding involves finding the distance and relationships between one document and all other documents in the dataset. For example, if the dataset is as shown in table 1, during both training process and testing process, the document being searched will be compared with all documents in the dataset. For instance, if the keyword is “ghibbah,” the search will include all words in documents with meanings such as “ ghibbah,” “gossip,” and other related terms, as well as “give charity” and “set aside assets”.

This step is performed in collaboration with hadith experts. In this stage, the researcher examines several documents with similar meanings but different content to determine how embedding assigns weight to each word and document. From the three documents in table 2, words with similar meanings “to speak ill of”, “to seek disgrace”, “to gossip”, “to backbite”, and “to eat human flesh”. These words share a negative connotation related to discussing others. The embedding process identifies semantic similarities and relationships among these words so that during information retrieval, related documents are categorized as relevant. This stage involves assigning vector weights and comparing the occurrence of these words in each document. An example of how embedding works is on Figure 5. First, observe where these words appear across the three sample. We have kept the data in the table and the examples in Indonesian, in accordance with the dataset used in this research.

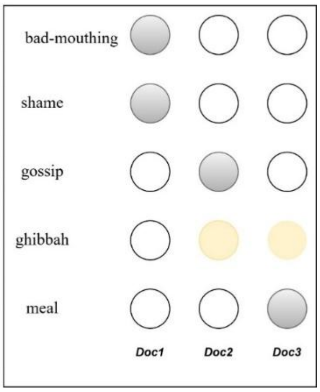


Figure 5. Example Of How Embedding Works

From the five words, it can be seen that the uncolored sections have a value of zero (0), indicating that the word does not appear in the document. The gray areas indicate a non-zero value, while the yellow areas signify that the word has more than one non-zero section, suggesting a strong relationship and overlap between the two elements or vectors [36]. Next, the distance and relationships between several words adjacent to the five mentioned words are examined. The process is as follows on Figure 6.

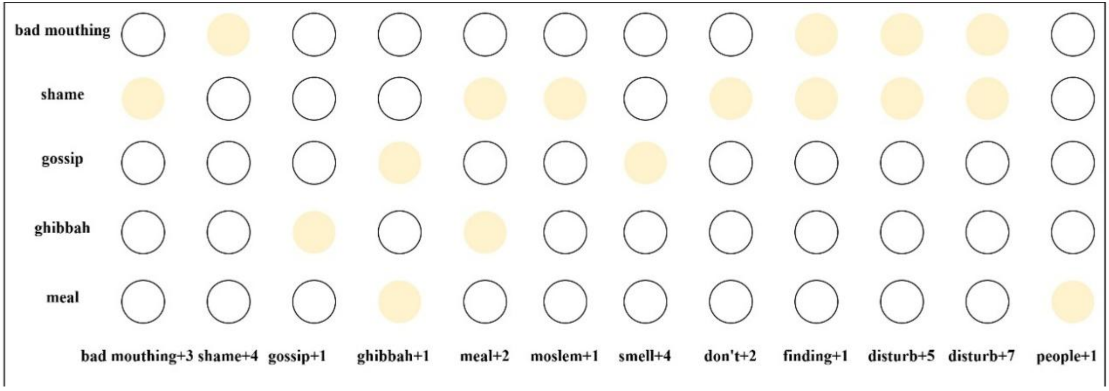


Figure 6. Example Of How Embedding Works Second Process

The process of finding the nearest distance for each word or feature can inform the model that these five words are closely related based on the distance calculations with other words, whether they appear on the right or left side. This is evidenced by the fact that non-zero values appear frequently, such as the feature “aib” appearing after “finding,” or “dont” appearing or being close to “bad-mouthing,” flanked by two other words. This process aids the model when a user inputs a query for retrieving hadith information. In the FastText embedding process, the window size has a significant impact on the distance calculation between words [37].

Table 4 illustrates a sample of weight values from the documents where this process is applied From this table, a close relationship or distance can be observed between the words “bakbite” and “gossip,” with weight values of -2.787 and -2.782 respectively, compared to other words.

Table 4. Vector Representation Embedding Value

No	Documents After Stopword Removal Process	Vector Value
1	The	-5.140
	Prophet	3.300
	When	-2.550
	Suddenly	-3.660
	smelled	-3.360
	unpleasant	-2.550
	stench.	-5.730
	Rasulullah	2.960

saw.	-1.250
said,	-5.140
'Do	3.300
know	-2.560
smell	-3.660
?	-3.360
the	-2.510
smell	-4.250
of	-5.140
those	3.300
who	-2.787
backbite	-2.782
(gossip)	-3.360
the	-2.440
believers	-2.440

4.5. Information Retrieval System

This section explains the results of the Information Retrieval process, from entering the query for the information sought to the system providing the output or document that answers the query. It also discusses the performance of the proposed model and highlights the advantages of this research compared to previous studies with similar themes.

4.6. Matching Document With Neural Network

After the user inputs the query, the next step is to match the query with hadith documents in the data collection. This process involves several stages, starting with preprocessing the provided query and generating vectors through the embedding process. These vectors are then calculated and matched with the documents in the data collection. For instance, if the query yields the keyword "gossip," the system will search for words related to this keyword. Documents containing these words will have their weights calculated, considering both the words following (to the right) and preceding (to the left) the document. The documents related to the query will then be ranked as either relevant, significantly relevant, or not relevant. Metrics for each document will be computed using the Hadamard product, based on the number of relevant (positive) documents compared to irrelevant (negative) documents. In addition to checking word-by-word, the model will also examine keywords within the document dataset. The following is an example of the document search process and query matching, using 5 sample documents from the hadith dataset in table 5.

Table 5. Matching and Ranking Documents Process

Document	Query / Keyword		Vector
	Bad-Mouthing	Gossip	
D1	10.794	3.285	14.079
D2	15.047	3.87 9	18.926
D3	8.3996	2.898	11.2976
D4	9.761	3.124	12.885
D5	9.93	3.152	13.082

Based on table 7, it can be seen that the document ranking first for information regarding hadiths with meanings related to "menjelek-jelekkan" or "menggosip" is Document Number 2. Therefore, this document is classified as relevant and will be presented as an answer to the user. The next highest weights are for Document Number 1, Number 5, Number 4, and the last document is Document Number 5. These five documents are categorized as "relevant documents," while documents with lower weights are classified as "significantly relevant," and documents with a weight of 0 are categorized as "not relevant." The total weights are calculated based on the formula in Equation.

4.7. Query Reformulation

Query Reformulation This process is conducted only when the information sought does not match any documents. Query reformulation is performed by examining the weights of the given query and finding weights that are close to the query in each document. In this study, we used Deep Reinforced Query Reformulation for generate alternative query. The DRQR model's task can be formally defined as follows: it takes a user query X = [x1, x2, ..., xN], consisting of N terms, and a paraphrase of that query Y = [y1, y2, ..., yM], which has M terms. The model is trained to generate a reformulated query Yˆ = [yˆ1, yˆ2, ..., yˆM]. This generated query Yˆ is intended to help a retrieval system find documents that are relevant to the original query X. If the model finds a close weight, it will perform word matching both to the right and to the left to determine the relevance of the query with the found document. Subsequently, the value and words are returned to the query for restructuring. For example, if the query is "Are there hadiths about prohibiting the rebuking of orphans?" and this term is not found, it will be reformulated to "Are there hadiths recommending the honoring of orphans?" Another case might be "Mention hadiths about bank interest?" which would be reformulated to "Mention hadiths about usury." Table 6 shows the example of query reformulation cases.

Table 6. Example Of Query Reformulation Cases

No	Query	
	Original	**Reformulation**
1	What is the hadith about cheating?	What is the hadith about adultery?
2	What is the hadith about eating pork?	What types of meat are forbidden?
3	Hadith about transgender	Hadith about resembling men or women
4	Hadith about pregnancy outside of marriage.	hadith about adultery
5	Hadith about the virtues of tooth brushing	Hadith about the virtues of siwak

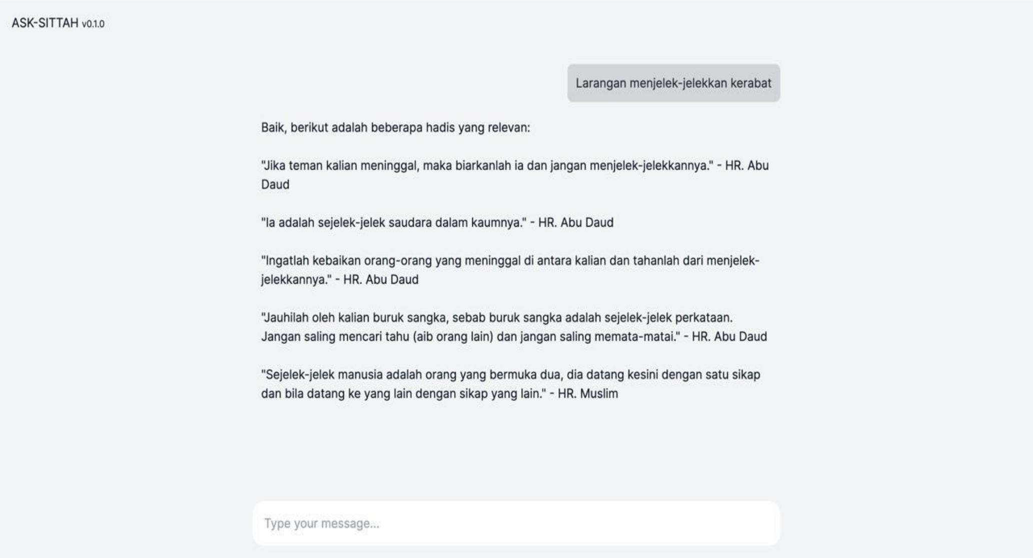
4.8. Document Evaluation

Document Evaluation Based on the evaluation results of the information retrieval process for the query "Prohibition of speaking ill of others," the Recall and Precision values were obtained using formulas (2) and (3) with two different experiments, altering the proportion of training and testing data. Precision is measured by the number of correct and relevant hadith predictions based on the query, while Recall is determined by the number of correct hadith present in the dataset. When the training data proportion was 80% and the testing data proportion was 20%, the results were a Recall value of 0.8721 and a Precision value of 0.79112. In contrast, with a training data proportion of 70% and a testing data proportion of 30%, the Recall value was 0.7721 and the Precision value was 0.75112. The author also found that FastText significantly impacts the recall value. Using another model like Word2Vec the results were less relevant, with a recall value of 0.5671.

In the test case, when we search for information regarding "imam" and "prayer," it refers to the virtues of prayer and the imam. The FastText model understands the intended query and retrieves hadith related to both the imam and prayer, as well as the virtues of these two aspects. When using the Word2Vec model, the results provided are not very relevant. When searching for "imam" and "prayer," the model instead retrieves information about prayer itself.

The final stage in the model development involved testing to prove that FastText performs better than Word2Vec by using keywords related to "imam and prayer." The results indicated that FastText produced hadith that were more accurate and relevant compared to Word2Vec, which yielded less relevant hadith.

After the model was completed and demonstrated good performance, it was subsequently implemented into a website-based system. This website is designed similarly to Google's "Gemini" or "ChatGPT" products, using the Neural Network model developed and proposed in this research. The system will be published to make it accessible to a global audience. Below is a sample of hadith search using the website system that has been created, as shown in figure 7.



Larangan menjelek-jelekkan kerabat

Baik, berikut adalah beberapa hadis yang relevan:

"Jika teman kalian meninggal, maka biarkanlah ia dan jangan menjelek-jelekkannya." - HR. Abu Daud

"Ia adalah sejelek-jelek saudara dalam kaumnya." - HR. Abu Daud

"Ingatlah kebaikan orang-orang yang meninggal di antara kalian dan tahanlah dari menjelek-jelekkannya." - HR. Abu Daud

"Jauhilah oleh kalian buruk sangka, sebab buruk sangka adalah sejelek-jelek perkataan. Jangan saling mencari tahu (aib orang lain) dan jangan saling memata-matai." - HR. Abu Daud

"Sejelek-jelek manusia adalah orang yang bermuka dua, dia datang kesini dengan satu sikap dan bila datang ke yang lain dengan sikap yang lain." - HR. Muslim

Figure 7. The System of Information Retrieval for Hadith Web Based

To use the web in figure 7, users can enter commands or questions in the input column that contains the writing "Type your message" and when the user presses enter, the system will search for hadiths that match the command entered. The system will produce the 5 most relevant hadiths that occupy the top ranking.

5. Conclusion

Based on the points and explanations provided earlier, this research has conducted an information retrieval process with the proposed model performing quite well. This study achieved better results compared to previous research on the same topic by increasing the dataset size from a maximum of 4,000 Sahih hadiths to 31,275. While previous research achieved a recall of 0.781, this study reached a higher recall of 0.8721. Additionally, this research enables the wider community, particularly Muslims, to search for hadiths more easily, accurately, and according to the core keywords sought. This improvement is expected to enhance Muslims' knowledge in distinguishing Sahih hadiths, prevent errors in religious understanding and beliefs, and reduce deviations in religious practices.

This study performs hadith information retrieval based on keywords with an effective process and achieves good performance. The use of FastText for embedding significantly impacts the ability to generate relevant documents for the query. This is because FastText performs well in finding similar or related meanings of different words. The ranking and document matching process using Neural Networks also yields good performance in this study, even with a relatively large dataset. This is due to the model not only searching for the exact words input as the query but also identifying patterns in sentences and examining words that occur before or after the query. Future research could innovate by using Arabic as the field for the model and exploring libraries beyond Sastrawi. Additionally, newer Neural Network models could be tested and developed in multiple languages to enable global applicability

6. Declarations

6.1. Author Contributions

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

6.2. Data Availability Statement

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

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

[1] T. Azhar, “Hadith Isnad Study In The Discovery Of Islamic Law (Critique of the Thoughts of Goldziher and Schacht),” *Eduvest-Journal of Universal Studies*, vol. 4, no. 05, pp. 4170–4183, 2024.

[2] D. Rasyid, A. D. Rasyid, A. Lubis, M. A. W. F. B. M. Balwi, and B. D. Rasyid, “The writing of hadith in the era of prophet muhammad A Critique on Harun Nasution’s Thought,” *Al-Jami’ah*, vol. 59, no. 1, pp. 191–220, 2021, doi: 10.14421/ajis.2021.591.191-220.

[3] N. S. B. N. Fauzi, M. Hoque, and K. A. Kadir, “Spreading Hadith Maudhu’ Via Information And Communication Technology: Reasons And Suggestions,” *Journal Of Hadith Studies*, vol. 7, no. 1, pp. 160–167, Jun. 2022

[4] K. Jamal, A. I. Mauliddin, and D. B. Dalimunthe, “The Implication of Asbabun Nuzul for Al-Quran Verses Interpretation,” *Kawanua International Journal of Multicultural Studies*, vol. 3, no. 1, pp. 12–17, Jun. 2022.

[5] R. Abbas, “Nahwu Al Fiqh Al Jadid: Controversy Surrounding Jamal Al banna’s Thought About Hadith Narrated by the Companions of the Prophet,” *Europan Journal for Philosophy of Religion*, vol. 15, no. 3, pp. 331–345, 2023.

[6] S. N. H. Adam, A. M. A. A. Sulyman, and N. Al-Makki, “Role of Family and Masjid in Preserving Morals and Thoughts from Pollution,” *Social Science Journal*, vol. 13, no. 2, pp. 3524–3533, Jan. 2023, Accessed: Nov. 20, 2024.

[7] M. Zulqarnain, R. Ghazali, Y. M. M. Hassim, and M. Rehan, “A comparative review on deep learning models for text classification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 325–335, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp325-335.

[8] I. U. Haq, M. Pifarré, and E. Fraca, “Natural Language Processing Approach to Evaluate Real-Time Flexibility of Ideas to Support Collaborative Creative Process,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 19, no. 5, pp. 93–107, Jun. 2024, doi: 10.3991/ijet.v19i05.47465.

[9] E. D. S. Mulyani, N. N. Febriani, A. Darmawan, R. A. Wiyono, R. D. Saputra, and D. Rohpandi, “Keyword-Based Hadith Grouping Using Fuzzy C-Means Method,” in *2020 2nd International Conference on Cybernetics and Intelligent System, ICORIS 2020*, Manado: Institute of Electrical and Electronics Engineers Inc., vol. 2, no. Oct., pp. 1-7, Oct. 2020. doi: 10.1109/ICORIS50180.2020.9320796.

[10] E. A. Olivetti *et al.*, “Data-driven materials research enabled by natural language processing and information extraction,” *Appl Phys Rev*, vol. 7, no. 4, pp. 1–12, Dec. 2020.

[11] J. Guo *et al.*, “A Deep Look into neural ranking models for information retrieval,” *Inf Process Manag*, vol. 57, no. 6, pp. 1–20, Nov. 2020, doi: 10.1016/j.ipm.2019.102067.

[12] P. Ansari, “Reigious Moderation in Caring for Indonesian Plurality (Analysis of Mental Revolution from The Qur’an Perspective),” *PUSPITUR: International Journal of Academic Research (PIJAR)*, vol. 1, no. 1, pp. 55–62, 2024, Accessed: Nov. 19, 2024.

[13] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, “Dense Text Retrieval Based on Pretrained Language Models: A Survey,” *ACM Trans Inf Syst*, vol. 42, no. 4, pp. 1–60, Feb. 2024.

[14] C. N. Tulu, "Experimental Comparison of Pre-Trained Word Embedding Vectors of Word2Vec, Glove, FastText for Word Level Semantic Text Similarity Measurement in Turkish," *Advances in Science and Technology Research Journal*, vol. 16, no. 4, pp. 147–156, Sep. 2022, doi: 10.12913/22998624/152453.

[15] M. F. Ahamed *et al.*, "A review on brain tumor segmentation based on deep learning methods with federated learning techniques," *Computerized Medical Imaging and Graphics*, vol. 110, no. 1, pp. 1-12, Nov. 2023, doi: 10.1016/j.compmedimag.2023.102313.

[16] S. Bruch, C. Lucchese, and F. M. Nardini, "Efficient and Effective Tree-based and Neural Learning to Rank," *Foundations and Trends in Information Retrieval*, vol. 17, no. 1, pp. 1–123, 2023, doi: 10.1561/1500000071.

[17] H.-T. Nguyen, M.-K. Phi, X.-B. Ngo, V. Tran, L.-M. Nguyen, and M.-P. Tu, "Attentive Deep Neural Networks for Legal Document Retrieval," *Springer Nature*, vol. 35, no. 1, pp. 57–86, Dec. 2022, doi: 10.1007/s10506-022-09341-8.

[18] G. Faggioli, T. Formal, S. Marchesin, S. Clinchant, N. Ferro, and B. Piwowarski, "Query Performance Prediction for Neural IR: Are We There Yet?," in *45th European Conference on Information Retrieval*, Dublin: ECIR, vol. 45, no. Apr., pp. 232–248, 2023.

[19] S. E. Pratama, W. Darmalaksana, D. Sa'adillah Maylawati, H. Sugilar, T. Mantoro, and M. A. Ramdhani, "Weighted inverse document frequency and vector space model for hadith search engine," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 2, pp. 1004–1014, 2020, doi: 10.11591/ijeecs.v18.i2.pp1004-1014.

[20] B. Sulistio, A. Ramadhan, E. Abdurachman, M. Zarlis, and A. Trisetyarso, "The utilization of machine learning on studying Hadith in Islam: A systematic literature review," *Educ Inf Technol*, vol. 29, no. 5, pp. 5381–5419, Dec. 2023.

[21] A. Abdi, S. Hasan, M. Arshi, S. M. Shamsuddin, and N. Idris, "A question answering system in hadith using linguistic knowledge," *Comput Speech Lang*, vol. 60, no. 1, pp. 1–13, Mar. 2020, doi: 10.1016/j.csl.2019.101023.

[22] Z. Abbasiantaeb and S. Momtazi, "Text-based question answering from information retrieval and deep neural network perspectives: A survey," *WIRES Data Mining and Knowledge Discovery*, vol. 11, no. 6, pp. 1–40, Nov. 2021, doi: 10.1002/widm.1412.

[23] E. M. Dharma, F. L. Gaol, H. L. H. S. W. Warnars, and B. Soewito, "The Accuracy Comparison Among Word2vec, Glove, And Fasttext Towards Convolution Neural Network (Cnn) Text Classification," *J Theor Appl Inf Technol*, vol. 100, no. 2, pp. 1-10, Jan. 2022, Accessed: Nov. 20, 2024.

[24] A. Esteva *et al.*, "COVID-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization," *NPJ Digit Med*, vol. 4, no. 1, pp. 1–9, Apr. 2021, doi: 10.1038/s41746-021-00437-0.

[25] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 1, pp. 1–20, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[26] K. Hambarde and H. Proenca, "Information Retrieval: Recent Advances and Beyond," *IEEE Access*, vol. 11, no. 1, pp. 76581–76604, Jan. 2023, doi: 10.1109/ACCESS.2023.3295776.

[27] S. Namasudra, S. Dhamodharavadhani, and R. Rathipriya, "Nonlinear Neural Network Based Forecasting Model for Predicting COVID-19 Cases," *Neural Process Lett*, vol. 55, no. 1, pp. 171–191, Feb. 2023, doi: 10.1007/s11063-021-10495-w.

[28] M. M. Taye, "Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions," *Computation*, vol. 11, no. 3, pp. 1–23, Mar. 2023, doi: 10.3390/computation11030052.

[29] Z. Li, X. Yang, L. Zhou, H. Jia, and W. Li, "Text Matching in Insurance Question-Answering Community Based on an Integrated BiLSTM-TextCNN Model Fusing Multi-Feature," *Entropy*, vol. 25, no. 4, pp. 1–16, Apr. 2023, doi: 10.3390/e25040639.

[30] M. Ciaperoni, A. Gionis, and H. Mannila, "The Hadamard decomposition problem," *Data Min Knowl Discov*, vol. 38, no. 1, pp. 2306–2347, May 2024, doi: 10.1007/s10618-024-01033-y.

[31] N. Kaur and H. Aggarwal, "Query reformulation approach using domain specific ontology for semantic information retrieval," *International Journal of Information Technology*, vol. 13, no. 1, pp. 1745–1753, May 2021.

[32] R. Kumar and S. C. Sharma, "Hybrid optimization and ontology-based semantic model for efficient text-based information retrieval," *Journal of Supercomputing*, vol. 79, no. 2, pp. 2251–2280, Feb. 2023, doi: 10.1007/s11227-022-04708-9.

[33] Y. Ramdhani, D. F. Apra, and D. P. Alamsyah, "Feature selection optimization based on genetic algorithm for support vector classification varieties of raisin," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 192–199, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp192-199.

[34] M. Yahya, D. Puyu, Ilyas, Z. Alwi, and M. Z. A. Nawas, "Comparative Critical Analysis of Methodologies for Establishing the Validity of Hadith Among Sunni and Shia," *International Journal of Religion*, vol. 5, no. 6, pp. 777–792, Apr. 2024, doi: 10.61707/31ec2561.

[35] A. Aldoseri, K. N. Al-Khalifa, and A. M. Hamouda, "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges," *Applied Sciences (Switzerland)*, vol. 13, no. 12, pp. 7082-7090, Jun. 2023, doi: 10.3390/app13127082.

[36] G. van Capelleveen, C. Amrit, H. Zijm, D. M. Yazan, and A. Abdi, "Toward building recommender systems for the circular economy: Exploring the perils of the European Waste Catalogue," *J Environ Manage*, vol. 277, no. 1, pp. 1-12, Jan. 2021, doi: 10.1016/j.jenvman.2020.111430.

[37] F. Torregrossa, V. Claveau, N. Kooli, G. Gravier, and R. Allesiardo, "On the Correlation of Word Embedding Evaluation Metrics," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille: European Language Resources Association (ELRA), vol. 12, no. May, pp. 4789–4797, 2020. Accessed: Nov. 20, 2024.