Kingdom of Saudi Arabia
Ministry of Education
University of Jeddah
College of Computer Science and Engineering
Department of Computer Science and Artificial Intelligence

المملكة العربية السعودية
وزارة التعليم
جامعة جدّة
كلية علوم وهندسة الحاسب
قسم علوم الحاسب والذكاء الاصطناعي

Project report:

Semantic Search Engine for Hadith

Supervisor:
Ms. Alaa Alharithi

Prepared by:

Ryouf Alghamdi 2110489
Rama Alyoubi 2110112
Rimas Alshehri 2110240

May 7, 2024

Contents

List of Figures

Abstract

Traditional search methods for religious texts, including Hadiths, often rely on simplistic keyword matching, hindering efficient and accurate retrieval of relevant information. In response to this limitation, we propose the development of a Semantic Search Engine for Hadiths. By harnessing the power of advanced AI models and semantic analysis techniques, our solution aims to enhance the accessibility and understanding of Hadiths by enabling searches based on the semantic meaning of queries rather than just keywords. This project seeks to bridge the gap between traditional search methods and the complex nature of Hadith texts, empowering users with the tools they need to access and comprehend these sacred texts more effectively.

1 Introduction

Accessing and comprehending religious texts, particularly Hadiths, plays a significant role in the lives of millions around the world. However, traditional search methods often fall short in providing efficient and accurate results due to their reliance on simplistic keyword matching. This limitation poses challenges for individuals seeking specific guidance or insights from Hadiths, hindering their ability to navigate these sacred texts effectively. In response to this challenge, we propose the development of a Semantic Search Engine for Hadiths. By harnessing the power of advanced AI models and semantic analysis techniques, our solution aims to enhance the accessibility and understanding of Hadiths by enabling searches based on the semantic meaning of queries rather than just keywords. This introduction sets the stage for our project, highlighting the importance of improving search capabilities for religious texts and outlining our approach to addressing this crucial need.

2 Background

The study and interpretation of religious texts, including Hadiths, are foundational to the spiritual and moral guidance of individuals within Islamic communities worldwide. Hadiths, which consist of the sayings, actions, and approvals of the Prophet Muhammad (peace be upon him), serve as a crucial source of Islamic jurisprudence and ethical teachings. However, accessing and navigating Hadith collections can be a daunting task, especially for individuals who may not have extensive knowledge of Arabic or Islamic scholarship. Traditional search methods often rely solely on keyword matching, which may yield incomplete or irrelevant results, hindering users' ability to find the specific guidance they seek. In recent years, advancements in artificial intelligence and natural language processing have opened up new possibilities for enhancing search capabilities, particularly in the realm of semantic analysis. Semantic search engines have the potential to understand the contextual meaning of queries and texts, thereby enabling more accurate and relevant results.[3]
The document titled **ISWSE: Islamic Semantic Web Search Engine.** It was published in the International Journal of Computer Applications in February 2015. The paper discusses the development of an Islamic Semantic Web Search Engine that focuses on searching and retrieving information from

the Holy Quran. The authors propose the ISWSE system, which is based on Islamic Ontology and uses Azhary as a lexical ontology for the Arabic language. The paper also mentions the challenges in understanding and extracting information from the Quran due to its complexity and large size. It discusses the importance of developing a semantic search engine to facilitate accurate and easy retrieval of information from the Quran. The document provides an introduction to the Semantic Web, discusses related works in the field of semantic search engines, and presents the proposed system architecture and its implementation. It also includes experimental results and a performance evaluation of the ISWSE system. However, the content of the document is truncated, and the remaining sections are not available. [3]
The document titled **Specialized Quranic Semantic Search Engine: Features and Architecture** appears to be a research paper published in the International Journal of Computer Science and Information Security (IJCSIS) in February 2019. The paper is authored by Moulay Ibrahim El-Khalil Ghembaza from the Department of Computer Science and IT Research Center for the Holy Quran and Its Sciences at Taibah University in Medina, Kingdom of Saudi Arabia. The abstract of the paper outlines the purpose of the study, which is to examine Quranic search engines and their services that support detailed queries within the Holy Quran. The paper discusses the challenges of developing specialized Quran search engines and proposes new features to enhance the search experience. It also introduces a semantic search engine specialized in the Arabic language of the Quran, which includes morphological analysis and ontology construction. The paper emphasizes the importance of specialized search engines for Quranic research and highlights the limitations of general-purpose search engines in handling Arabic and Quranic words effectively. It mentions the need for customized Quranic search engines that provide accurate and meaningful results for users interested in Quranic sciences. The introduction and motivation section of the paper discusses the use of computer technology in Quranic word processing and digital storage of Quranic content. It mentions the evolution of Quranic applications, starting from simple text storage to incorporating features such as Tajweed rules, Uthmanic script style diacritics, and audio recitations. The paper highlights the significance of search engines in the digital world and their role in retrieving information quickly. It emphasizes the need for efficient search engines for Arabic and Quranic words, considering their unique linguistic and semantic characteristics. The author suggests that specialized Quranic search engines can provide more accurate and faster

results compared to general-purpose search engines.[2]

3 Approach

In addressing the challenge of accessing and interpreting Hadiths, we developed a Semantic Search Engine that utilized advanced AI models and semantic analysis techniques. Our solution employed Natural Language Processing (NLP) to analyze both the literal and deeper contextual meanings of the Arabic text within the Hadiths [1]. The system integrated vector space modeling using the AraGPT, a GPT-2 architecture optimized for Arabic, renowned for its ability to recognize and generate linguistic patterns. This model played a crucial role in processing input queries and Hadith texts to compute semantic similarities that surpassed basic keyword matching, thereby enhancing the search engine's ability to decipher complex user queries and deliver pertinent results [4]. Initially, our development focused on fine-tuning the language model for optimal performance, configuring the AdamW optimizer to specifically update the model's head parameters. This phase involved rigorous training and validation cycles to minimize loss and prevent overfitting, ensuring the model's proficiency in generating accurate nuances of Arabic text in Hadiths. However, we primarily relied on our second approach, embedding generation, which proved to be exceptionally effective. By converting Hadith texts into dense vector representations, this method captured the intricate semantic relationships within the texts. The embeddings, derived from the model's last hidden state, represented texts in a high-dimensional space, essential for tasks like semantic analysis and clustering. This approach not only facilitated efficient data handling but also deepened our semantic understanding, which was pivotal for the search engine's functionality. Ultimately, both strategies were integral to creating a search engine that not only retrieved but also thoroughly understood and contextualized Hadiths to provide meaningful search experiences.

```
GPT2Model(
  (wte): Embedding(64000, 768)
  (wpe): Embedding(1024, 768)
  (drop): Dropout(p=0.1, inplace=False)
  (h): ModuleList(
    (0-11): 12 x GPT2Block(
      (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
      (attn): GPT2Attention(
        (c_attn): Conv1D()
        (c_proj): Conv1D()
        (attn_dropout): Dropout(p=0.1, inplace=False)
        (resid_dropout): Dropout(p=0.1, inplace=False)
      )
      (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
      (mlp): GPT2MLP(
        (c_fc): Conv1D()
        (c_proj): Conv1D()
        (act): NewGELUActivation()
        (dropout): Dropout(p=0.1, inplace=False)
      )
    )
  )
  (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)
)
```

Figure 1: Model configuration

To ensure the Semantic Search Engine operates optimally, we conducted comprehensive Exploratory Data Analysis (EDA) and visualization of the Hadith texts. This preliminary phase was crucial for understanding the distribution and characteristics of the data, aiding in the fine-tuning of our model.
We employed various visualization techniques to represent these insights graphically, facilitating a more intuitive understanding of patterns in the data.

- The Histogram shown in figure 2 illustrates the distribution of text lengths in our dataset, comparing Arabic and English texts. The x-axis represents the length of the text, while the y-axis indicates the frequency of texts with specific lengths. The graph reveals that both Arabic and English texts primarily cluster within a narrow range of shorter text lengths, as indicated by the tall peaks circled near the origin. This suggests that the majority of Hadith texts in both languages are relatively concise, typically under 1,000 characters in length. The frequency dramatically decreases as the text length increases, indicating fewer lengthy texts within the dataset. The overlay of Arabic (in

blue) and English (in red) text lengths allows us to observe slight variations in text length distribution between the two languages. This is crucial for understanding any systematic differences in translation or text structuring that may impact how our model process and analyze the texts.

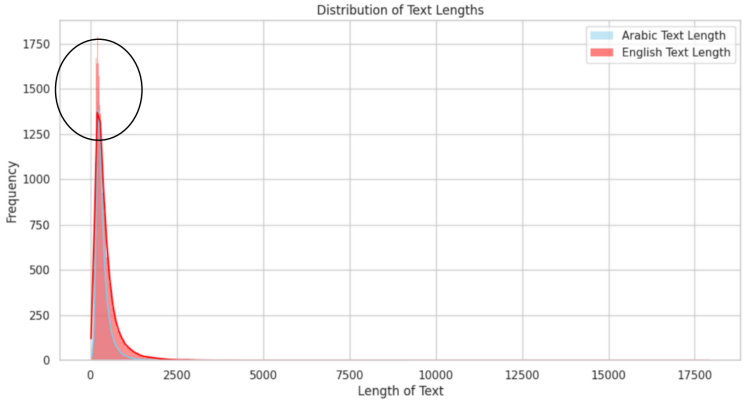

Figure 2: Distribution of Text Lengths in Arabic and English

- The Bar graph shown in figure 3 represents the number of Hadiths based on their source. The x-axis represents the different sources of the Hadiths, while the y-axis represents the count of Hadiths. Each source is represented by a distinct color, with corresponding labels in the legend. It can be inferred that Sahih Muslim is the primary source for the majority of hadiths, while Jami' al-Tirmidhi contributes to a smaller portion.

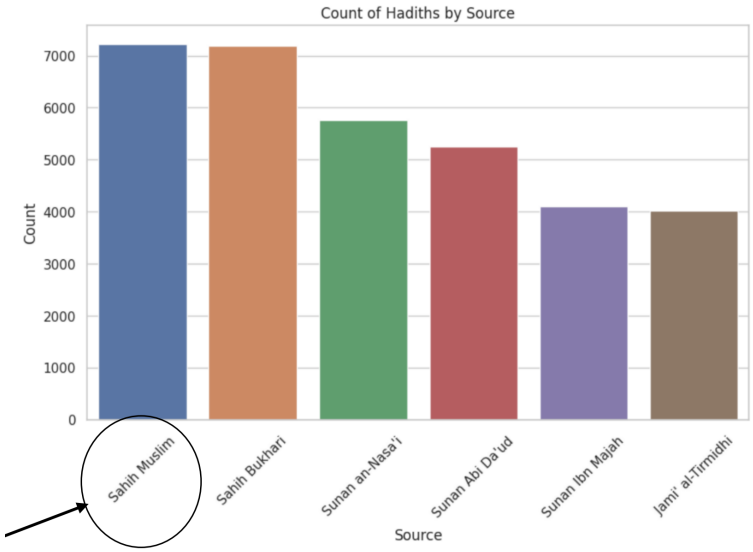

Figure 3: Distribution of Hadith Sources

- The Word Cloud shown in figure 4 presents a visually captivating representation of English Hadith texts. At its center, the word "Allah" stands out prominently, encircled by other significant terms like "peace," "prayer," and "narrated." Through thoughtful design, this image effectively captures the essential themes and concepts inherent in the texts, offering a concise and easily understandable visualization of their content.

Figure 4: WordCloud that captures the essential themes in hadiths

- The Stacked Area Chart shown in figure 5 effectively illustrates the cumulative number of Hadiths across different chapters. This type of graph allows for understanding the distribution and emphasis of teachings within the Hadith corpus, potentially guiding more detailed studies into specific chapters or themes based on the density and distribution of Hadiths.

Figure 5: Stacked Area Chart of number of Hadiths across different chapters

4 Experiments

4.1 Dataset Used

The dataset, titled all_hadiths_clean.csv, is sourced from Kaggle and contains 34,441 entries. The dataset includes 9 columns Descriptive statistics and frequency distribution of the sources and chapters are provided. Notably, the dataset includes texts from major Hadith collections like Sahih Bukhari and Sahih Muslim.

4.2 Experiment Execution

The experimental execution encompassed two distinct approaches utilizing the AraGPT model, each tailored to optimize model performance and feature extraction for Arabic text processing.

In the first approach, we focused on efficiently encoding textual data into embeddings using the function 'encode_text_with_labels'. The AraGPT model was configured with a maximum sequence length of 128 tokens, adhering to an 80-20 train-validation split. We processed the data in batches of 32, employing padding and truncation to maintain input consistency across batches. The embeddings were generated from the last hidden state of the model in evaluation mode, ensuring a robust dataset of features ready for subsequent modeling tasks.

The second approach employed a hands-on training and validation cycle within a controlled execution context. We initialized the tokenizer and model from 'aubmindlab/aragpt2-base', adding a padding token to the tokenizer if it was absent, and resizing the model's token embeddings accordingly. After setting all model parameters to a non-trainable state, we selectively unfroze the language modeling head to concentrate updates on text generation capabilities. Through the custom training function 'train_model', the model underwent multiple epochs of training, focusing on minimizing loss and refining performance, with each epoch followed by a validation phase using 'validate_model' to assess and monitor the model's effectiveness and learning progress.

Both methods were built on a standardized data handling framework with a

consistent sequence length and split ratio, ensuring comparability and reliability of the results. The first approach emphasized advanced feature extraction, while the second approach focused on targeted fine-tuning of the model to enhance its language generation proficiency. Together, these strategies demonstrated a comprehensive and nuanced understanding of model training and feature handling, tailored specifically for the complexities of Arabic natural language processing.

4.3 Evaluation Metrics

In the fine-tuning approach, we employed perplexity calculations to evaluate the performance of the model. This measure helped us assess the coherence and accuracy of the generated Arabic text, providing insights into its quality and linguistic fluency. However, in the subsequent embedding approach, our focus shifted towards utilizing cosine similarity scores to determine the semantic relevance of search results in relation to the query. This method proved to be highly suitable for semantic search applications, as it allowed us to identify texts that exhibited contextual similarity to the query. By leveraging cosine similarity, we could measure the semantic overlap between the query and the search results, providing a reliable indicator of their relevance and similarity in meaning.

4.4 Results

The results demonstrated the model's ability to retrieve and rank Hadith texts according to their semantic relevance. For instance, a sample query returned Hadith texts with similarity scores ranging from approximately 0.52 to 0.41. This suggests that the model successfully identified and ranked the Hadith texts in order of relevance to the input query.

4.5 Description of Results

In this section, we analyze the outputs of a semantic search algorithm applied to the study of hadith texts, focusing on the influence of narrator chains on the relevance scores assigned to these texts. By leveraging modern computational techniques, this analysis offers a new perspective on traditional Islamic scholarship. The scores and accompanying texts demonstrate how the semantic search model determines the relevance of each hadith to the

specified query. Importantly, the findings underscore variations in scores that appear to be affected by differences in the texts attributed to various chains of narrators.

For example, a hadith with a score of 0.525, narrated through a chain involving Muhammad bin Bashar and Yahya bin Sa'id, details the Prophet Muhammad's teachings on the proper conduct and alignment during prayer. This score suggests a high degree of relevance to the query, reflecting the strong connection made by the semantic model. Conversely, lower-scoring hadiths indicate either less relevance or weaker links in the semantic encoding process, which may be a result of the variations in the reliability and fidelity of the narrator chains. Through such analysis, we can better understand the nuances of hadith authenticity and the potential application of computational tools in Islamic studies.

5 Conclusion

In developing the Semantic Search Engine for Hadiths, we've discovered the power of advanced AI models and semantic analysis in enhancing access to religious texts. Our project underscores the importance of user experience design and the effectiveness of semantic search in providing more accurate and meaningful results. Looking ahead, future ideas include expanding data sources, integrating multimodal features, implementing personalized recommendations, and leveraging collaborative annotation. By continuing to innovate, we can further empower individuals to engage with sacred texts in a transformative manner.

References

[1] Wahyudin Darmalaksana et al. “Latent semantic analysis and cosine similarity for hadith search engine”. In: *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 18.1 (2020), pp. 217–227.

[2] MIEK Ghembaza. “Specialized Quranic Semantic Search Engine”. In: *International Journal of Computer Science and Information Security (IJCSIS)* 17.2 (2019).

[3] Hossam Ishkewy and Hany Harb. “Iswse: Islamic semantic web search engine”. In: *International Journal of Computer Applications* 112.5 (2015).

[4] Septya Egho Pratama et al. “Weighted inverse document frequency and vector space model for hadith search engine”. In: *Indones. J. Electr. Eng. Comput. Sci* 18.2 (2020), pp. 1004–1014.