Semantic Search in Religious Texts: Bridging Language Gaps in the Quran

Abdullah Alshami

February 2024

Contents

1 Introduction **3**

2 Related Work **4**
- 2.1 Quranic Semantic Search Tools . . . 4
- 2.2 NAS Benchmarks for NLP . . . 4
- 2.3 Math Word Embedding in Search . . . 4
- 2.4 COVID-19 Information Retrieval . . . 4
- 2.5 IR-BERT for Semantic Search . . . 4

3 System Overview **6**
- 3.1 Algorithm Overview . . . 6
  - 3.1.1 Text Processing and Embedding . . . 6
  - 3.1.2 Prediction Process . . . 7
- 3.2 Detailed Explanation . . . 7

4 Methodology **8**
- 4.1 Dataset Preparation . . . 8
- 4.2 Semantic Embedding . . . 8
- 4.3 Preprocessing . . . 8
- 4.4 Choice of NLP Model . . . 8
- 4.5 Semantic Search Algorithm . . . 8

5 Evaluation **10**
- 5.1 User Satisfaction and Feedback through a Custom-Built Windows Application . 10
  - 5.1.1 Deployment and Interaction . . . 10
  - 5.1.2 Collection of Feedback . . . 10
- 5.2 Conclusion of Evaluation . . . 11

6 Results **12**

7 Discussion **12**
- 7.1 Interpretation of Results . . . 12
- 7.2 Implications . . . 12
- 7.3 Limitations . . . 12

8 Conclusion and Future Work **13**

References **14**

Abstract

This study introduces a semantic search system leveraging the all-mpnet-base-v2 model within the Sentence Transformers framework to enhance accessibility to the Quran for non-Arabic speakers. By interpreting nuanced meanings behind user queries, the system effectively matches them with relevant Quranic verses. The technology aims to bridge linguistic and cultural divides, making the rich teachings of the Quran more accessible to a global audience. Initial evaluations demonstrate its potential in improving comprehension and accessibility, emphasizing the role of semantic technologies in religious studies and highlighting the innovative use of NLP techniques to overcome language barriers in accessing sacred texts.

1 Introduction

Religious texts serve as foundational elements of cultural and spiritual identity for millions around the globe. However, linguistic barriers can significantly limit the accessibility of these texts to non-native speakers. The Quran, in particular, is rich with linguistic intricacies and deep semantic layers that pose substantial challenges for those unfamiliar with Arabic. This paper explores the development of a semantic search system designed to make the Quran more accessible to non-Arabic speakers through the application of state-of-the-art natural language processing (NLP) models. By leveraging the all-mpnet-base-v2 model, known for its robust semantic understanding capabilities, this system interprets user queries and matches them with the most relevant Quranic verses. Our approach not only enhances the accessibility of the Quran but also sets a precedent for using semantic technologies to bridge language gaps in religious and cultural texts. This paper outlines the system's design, its underlying methodologies, and discusses the broader implications of this technology in enhancing the accessibility of religious documentation globally.

2 Related Work

The development of our semantic search system for the Quran draws inspiration and insights from a variety of related works spanning semantic search technologies, NLP model benchmarks, and applications in religious and non-religious contexts. This section outlines key studies that have shaped our approach, emphasizing their relevance and contributions to our project.

2.1 Quranic Semantic Search Tools

Building upon the QSST project's pioneering efforts in applying neural networks to enhance the understanding and searchability of Quranic texts, our project aligns with and extends this foundational work. The QSST project's methodology in semantic analysis and word vector generation directly informs our approach, particularly in choosing and optimizing NLP models for semantic embedding of Quranic verses[5].

2.2 NAS Benchmarks for NLP

The exploration of Neural Architecture Search (NAS) through benchmarks such as NAS-Bench-101 and NAS-HPO-Bench offers crucial insights into model selection and optimization for NLP tasks. While not focused on Quranic text, these benchmarks highlight essential considerations in achieving reproducibility and efficiency in model training, which are pivotal to our system's development for robust and performant semantic search capabilities[4].

2.3 Math Word Embedding in Search

The study on math word embedding, especially within search contexts, elucidates the complexity of embedding specialized vocabularies and concepts. These insights are invaluable to our project as they guide our preprocessing and embedding strategies, ensuring our system adeptly captures the semantic richness of Quranic language, despite its focus outside the religious domain[3].

2.4 COVID-19 Information Retrieval

Insights from COVID-19 information retrieval projects, focusing on the accuracy and timeliness of search results, are particularly relevant in the context of our work. These projects underscore the importance of adapting semantic search technologies to handle dynamic information needs, mirroring our goal to provide relevant and accessible Quranic verses to users with diverse inquiries[2].

2.5 IR-BERT for Semantic Search

The application of BERT-based models, as demonstrated in the IR-BERT project, has been instrumental in shaping our methodology. This project's success in leveraging deep learning models to enhance search relevancy and precision parallels our use of the all-mpnet-base-v2 model, showcasing the transformative potential of transformer models in achieving nuanced semantic understanding and accurate retrieval in semantic search systems[1].

Each of these related works contributes distinct insights and methodologies that have informed the development of our semantic search system, highlighting the multifaceted nature of research and innovation in the domain of semantic search and NLP technologies.

3 System Overview

The semantic search system is designed to process and match Quranic verses to user queries using advanced natural language processing techniques. This section provides an overview of the algorithm, divided into two main parts: text processing and prediction.

3.1 Algorithm Overview

The complete semantic search algorithm consists of two main components: processing the text into embeddings and using these embeddings to make predictions. The following flowcharts illustrate the step-by-step process involved in each component.

3.1.1 Text Processing and Embedding

The process starts with the initialization of the SentenceTransformer model, followed by reading and preprocessing the text from the Quran. Each verse is then converted into an embedding that captures its semantic meaning. This allows for more effective matching with queries based on semantic similarity. The flowchart in Figure 1 details these steps.

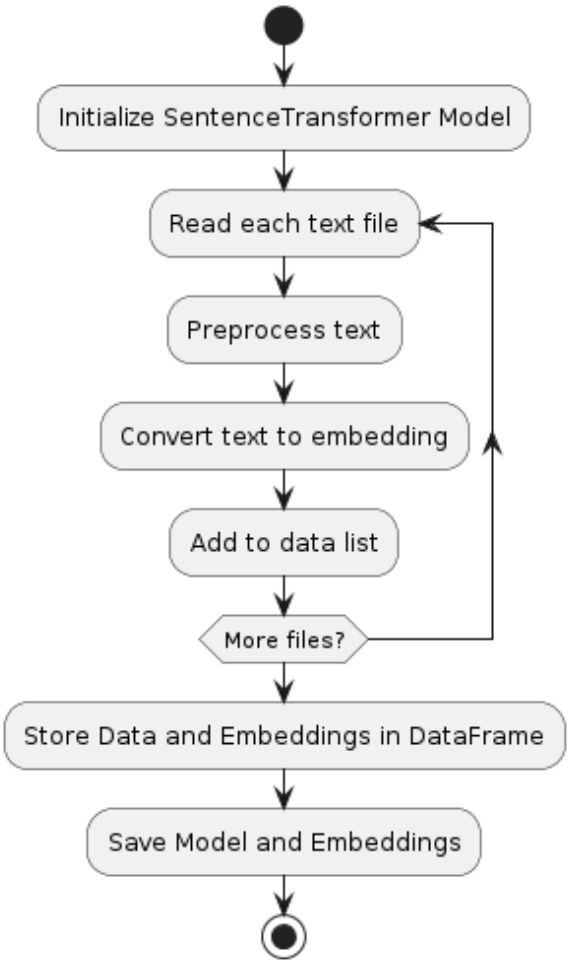

Figure 1: Flowchart illustrating the text processing and embedding steps of the semantic search system. This involves reading text files, preprocessing text, converting text to embeddings, and storing these embeddings for later use.

3.1.2 Prediction Process

Once the embeddings are prepared, they can be used to process user queries. The system receives a query, preprocesses it in a similar manner to the Quranic verses, and converts it into an embedding. Using cosine similarity, the system then identifies the most semantically relevant verses. The process for these operations is shown in Figure 2.

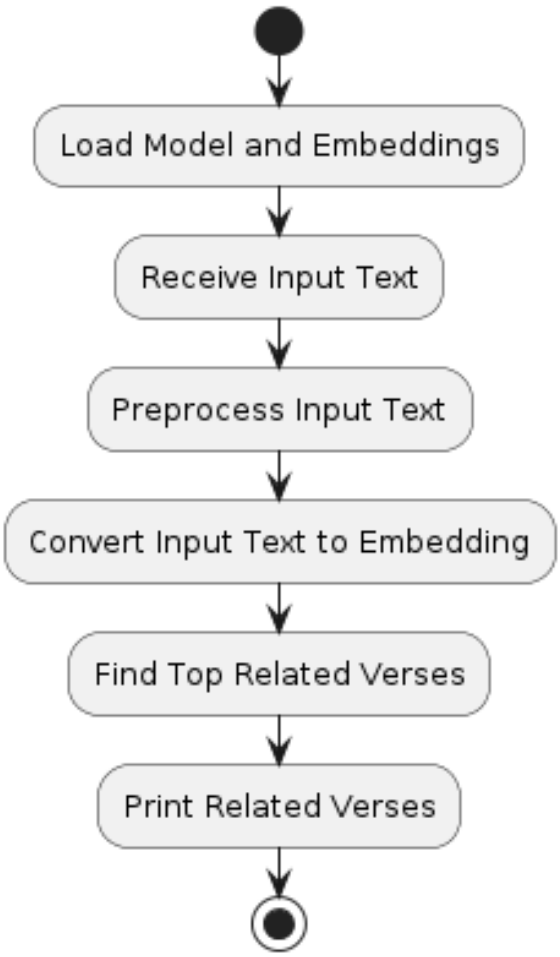

Figure 2: Flowchart of the prediction process in the semantic search system, showing how user queries are processed to find relevant Quranic verses based on semantic similarity.

3.2 Detailed Explanation

Following the flowcharts, this section delves deeper into each step involved in the algorithm, explaining how they contribute to the overall functionality of the system. The emphasis is on how the model initializes, processes data, generates embeddings, and utilizes these embeddings to effectively match queries with Quranic verses, enhancing the accessibility and understanding of the Quranic text.

4 Methodology

This section outlines the methodology used in developing a semantic search system to explore the Quran through English queries. It covers dataset preparation, preprocessing steps, the semantic search algorithm, and the rationale behind selecting our NLP model.

4.1 Dataset Preparation

For our study, we utilized a respected English translation of the Quran, chosen for its clarity, accessibility, and academic endorsement. This approach allows non-Arabic speakers to meaningfully interact with the text, leveraging the computational advantages inherent in the English language.

4.2 Semantic Embedding

The 'all-mpnet-base-v2' model within the Sentence Transformers framework was employed to enable semantic search capabilities. This model, pre-trained on a vast corpus of textual data, excels at converting text into dense semantic embeddings, capturing the nuanced meanings of words and phrases.

4.3 Preprocessing

Our preprocessing involves standardizing the text by removing punctuation and converting to lowercase, ensuring uniformity for semantic analysis. This step is crucial for maintaining the integrity of the original text while preparing it for effective semantic embedding, without the need for more complex processes like lemmatization or stemming.

4.4 Choice of NLP Model

The 'all-mpnet-base-v2' model was selected for its superior performance in semantic understanding and embedding generation. In comparative analyses with other models, including BERT and RoBERTa, 'all-mpnet-base-v2' demonstrated unparalleled efficiency in capturing the complex semantics required for accurately matching user queries with relevant Quranic verses. Its processing efficiency and compatibility with our system architecture were also crucial factors in its selection.

4.5 Semantic Search Algorithm

Our semantic search algorithm is fundamental to the system, designed to match user queries with Quranic verses based on semantic similarity accurately. The algorithm entails:

1. **Semantic Embedding Generation**: Each Quranic verse is preprocessed to standardize the text, which is then transformed into dense vector embeddings using the 'all-mpnet-base-v2' model. This captures the semantic essence of each verse.

2. **User Query Processing**: Similar preprocessing and embedding procedures are applied to user queries, ensuring both queries and verses are represented in a comparable semantic space.

3. **Cosine Similarity Calculation**: We measure the semantic closeness between the query embedding and verse embeddings using cosine similarity. A higher score indicates a closer semantic match.

4. **Ranking Relevant Verses**: Verses are ranked by their similarity scores, with the top matches presented to the user. This process ensures users receive the most relevant verses corresponding to their queries.

This algorithm leverages advanced NLP techniques to navigate the linguistic depth and contextual nuances of the Quran, enhancing the system's ability to return contextually appropriate verses. Continuous refinement based on user feedback and predefined metrics further improves the accuracy and user experience of the search system.

5 Evaluation

Our evaluation focused on qualitative user feedback to assess the system's effectiveness. Users interacted with the system through a custom-built application, providing insights into its usability and the relevance of search results. While this approach prioritized subjective experiences over quantitative metrics, it offered valuable perspectives on the system's real-world impact.

5.1 User Satisfaction and Feedback through a Custom-Built Windows Application

To facilitate an immersive and interactive evaluation process, we developed a custom-built Windows application specifically designed for this purpose. This application served as a platform for users to directly engage with the semantic search system, providing an environment that closely mirrors real-world usage scenarios.

5.1.1 Deployment and Interaction

Participants were selected from a diverse group of individuals, including students and laypersons with an interest in Quranic studies, ensuring a wide range of perspectives and experiences were captured. These participants were invited to use the Windows application over a designated period, allowing them ample time to explore its functionality and to perform searches based on their interests or queries they might realistically have.

The application was engineered to be user-friendly, with a focus on minimizing barriers to effective interaction. Its design emphasized intuitive navigation and straightforward access to the search capabilities of the system, making it accessible to users regardless of their technical proficiency.

5.1.2 Collection of Feedback

After interacting with the system through the Windows application, participants provided feedback via detailed surveys embedded within the application itself. These surveys were designed to capture comprehensive user impressions, covering various aspects of their experience.

Key findings from this feedback include:

- **Ease of Use**: Users consistently praised the intuitive design of the search interface within the Windows application. This ease of navigation was highlighted as a significant factor in facilitating a positive interaction with the system, indicating the application's success in making the semantic search accessible to all users.
- **Relevance of Results**: Participants reported a high level of satisfaction with the relevance of the search results returned by the system. This satisfaction underscores the system's effectiveness in understanding and accurately matching the intent behind user queries, reaffirming the utility of the semantic search capabilities.
- **Suggestions for Improvement**: Despite the positive feedback, users also provided constructive suggestions for enhancing the system. Requests included providing more detailed explanations of certain Quranic verses and improving support for queries in multiple languages. These insights are invaluable for guiding future enhancements to the system.

5.2 Conclusion of Evaluation

The use of a custom-built Windows application for the evaluation of our semantic search system played a crucial role in gathering authentic user feedback. The insights gained from this process have been instrumental in validating the system's effectiveness and identifying opportunities for further refinement. The overwhelmingly positive response, particularly regarding ease of use and relevance of results, highlights the system's potential impact on making Quranic teachings more accessible. The constructive suggestions received will inform the next stages of our development, ensuring that the system continues to evolve in response to user needs.

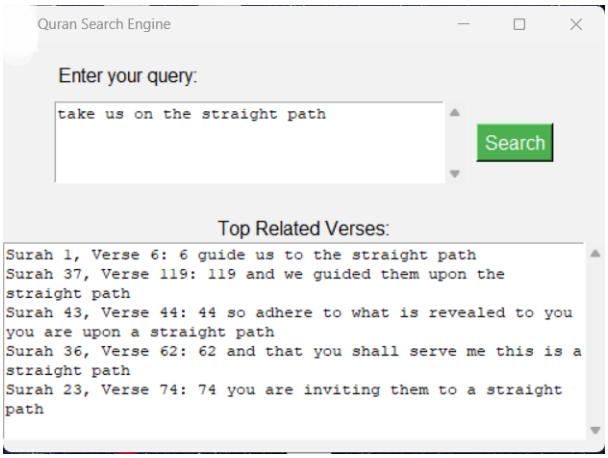

Figure 3: the program used to test

6 Results

The feedback received from users underscores the semantic search system's value in making Quranic teachings more accessible to non-Arabic speakers. The positive responses, particularly regarding the ease of use and relevance of results, highlight the system's alignment with user needs and its potential as a valuable tool for Quranic study.

The suggestions for improvement have been instrumental in identifying directions for further development, emphasizing the importance of ongoing engagement with users. By continuing to incorporate user feedback into the evolution of the system, we aim to enhance its functionality, extend its language capabilities, and ultimately, increase its utility to a global audience.

This user-focused evaluation not only affirms the current strengths of the system but also lays a foundation for future advancements that will broaden its reach and impact.

7 Discussion

7.1 Interpretation of Results

The user feedback sessions provided valuable insights into the usability and effectiveness of the semantic search system. The most significant finding was that users found the system easy to use and beneficial for searching Quranic verses quickly. This aligns perfectly with the objectives of our research.

Unexpectedly, some users attempted to input nonsensical words or overly common phrases, leading to irrelevant search results. This highlights the importance of providing specific and meaningful queries to the system to ensure relevant outputs.

7.2 Implications

The findings have significant implications for the usability and effectiveness of the semantic search system. Users reported that the system was helpful in locating specific Quranic verses and even discovering new content within the Quran. This suggests that the system can effectively address the challenge of accessing religious texts, particularly the Quran, for non-Arabic speakers.

Furthermore, the results advance our understanding of the potential of semantic search technologies in religious studies. By providing quick and accurate access to religious texts, such systems can enhance scholarly research and spiritual exploration.

7.3 Limitations

Our study's limitations include the choice of pre-trained models, which may not fully capture the contextual depth of the Quran's language. The feedback mechanism, primarily based on English queries, might not reflect the nuanced understanding required for more complex Arabic terms. Addressing these limitations will be crucial for enhancing the system's accuracy and broadening its applicability.

8 Conclusion and Future Work

We have developed a semantic search system that significantly improves the accessibility of the Quran for non-Arabic speakers, validated by user feedback. Future work will focus on enhancing the model's understanding of linguistic nuances and extending support for additional languages. Additionally, we are considering the integration of the Sunnah and Hadith of Prophet Muhammad (peace be upon him) into our system, aiming to provide a more comprehensive resource for Islamic studies. This expansion will not only broaden the system's reach and applicability to other religious texts but also underscore the transformative potential of NLP technologies in making cultural and religious heritage globally accessible.

References

[1] Anup Anand Deshmukh and Udhav Sethi. Ir-bert: leveraging bert for semantic search in background linking for news articles. *arXiv preprint arXiv:2007.12603*, 2020.

[2] Andre Esteva, Anuprit Kale, Romain Paulus, Kazuma Hashimoto, Wenpeng Yin, Dragomir Radev, and Richard Socher. Covid-19 information retrieval with deep-learning based semantic search, question answering, and abstractive summarization. *NPJ digital medicine*, 4(1):68, 2021.

[3] André Greiner-Petter, Abdou Youssef, Terry Ruas, Bruce R Miller, Moritz Schubotz, Akiko Aizawa, and Bela Gipp. Math-word embedding in math search and semantic extraction. *Scientometrics*, 125:3017–3046, 2020.

[4] Nikita Klyuchnikov, Ilya Trofimov, Ekaterina Artemova, Mikhail Salnikov, Maxim Fedorov, Alexander Filippov, and Evgeny Burnaev. Nas-bench-nlp: neural architecture search benchmark for natural language processing. *IEEE Access*, 10:45736–45747, 2022.

[5] Ensaf Hussein Mohamed and Eyad Mohamed Shokry. Qsst: A quranic semantic search tool based on word embedding. *Journal of King Saud University-Computer and Information Sciences*, 34(3):934–945, 2022.