Semantic search engine for the holy Quran

Abstract:

With the rapid advancement of AI and language models, it has become increasingly important to leverage these technologies in serving Arabic content in general, and religious texts in particular. The Holy Qur'an, which stands as the primary source of Islamic legislation, holds immense linguistic and semantic richness. However, researchers often face challenges in analyzing its meanings and retrieving verses related to various synonymous or contextually linked terms.

Given the diversity of Arabic dialects and the variation in word usage across regions, there is a growing need for intelligent tools capable of understanding semantic meaning rather than relying solely on literal keyword matching. This paper introduces a research project that utilizes enhanced Arabic language models to develop an advanced semantic search system for the Qur'an. The goal is to assist scholars and students in accessing verses and their interpretations based on meaning, rather than on surface-level word matches alone.

Project Overview:

This project aims to develop a semantic retrieval system for the Holy Qur'an, capable of understanding abstract concepts and synonyms using pre-trained Arabic language models fine-tuned for this task.

The system allows users to:

Submit natural language queries (e.g., "punishment of Hell")

Interpret the intended meaning and retrieve related verses (e.g., "Sa'ir", "Hawiyah", "evil abode")

Display interpretations for each verse, drawing from trusted sources like Ibn Kathir and Al-Qurtubi

Target Users:

- Students of Islamic studies
- Researchers in Qur'anic interpretation and Arabic linguistics
- Anyone seeking deeper insight into Qur'anic meanings.

Dataset:

A custom JSON dataset was created for this project, comprising over 8,058 entries. It includes:

- Quranic verses
- Keywords and synonyms
- Interpretations from trusted sources

Texts were processed and converted into numerical embeddings using a fine-tuned AraBERT model, enhancing semantic linkage and retrieval accuracy.

Language Model (LLM Used) AraBERT:

- Parameters: 371 million
- Hidden size: 768
- Trained on: 200M+ Arabic sentences (77GB)
- Supports Modern Standard Arabic
- Optimized for medium-resource devices

The model was further fine-tuned with Qur'anic vocabulary and Islamic concepts to boost semantic understanding.

Technical Architecture:

- Sparse RAG for text retrieval
- LSTM + Attention for contextual understanding
- AraBERT-based embeddings

Related Work:

- Omar Ahmed (Finland, May 2024): Used asafaya/bert-base-arabic for Qur'anic search
- Customer sentiment analysis in Saudi Arabia: AraBERT applied to telecom tweets
- Medical record anonymization: AraBERT used to detect and protect sensitive data