Semantic Summarization of Religious Texts: A Centroid-Based Approach with Word Embeddings

Abstract
This research presents a tool that uses AI to summarize long religious texts, articles, research papers, and so on. It focuses on the main ideas, creating a shorter version of it while keeping the important parts of the reading. By using the "word embeddings" method, the tool understands word meanings based on context, helping it pick and understand the sentences that represent key themes without repeating information. This AI tool can save time for students, researchers, and scholars and offer them a quick summary of the main points in long-complex religious texts.

Introduction

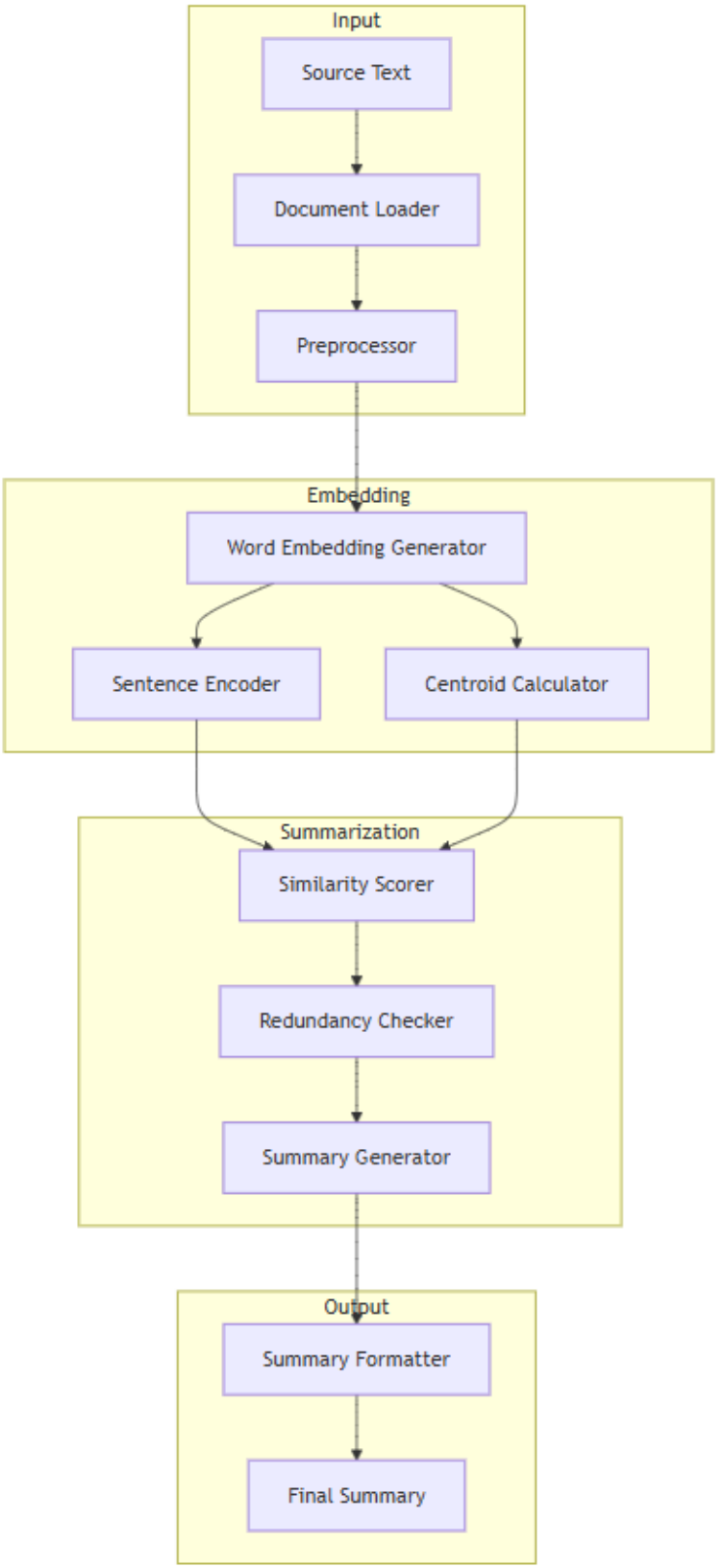
Religious texts are often very long, detailed, and sometimes hard to understand. Scholars, researchers, or whoever wants to study this kind of text—need to spend a lot of time reading to understand the main ideas. This can be challenging, especially if they need to get to the main points quickly. A tool that can create summaries of religious texts would make this much easier.

In this research, we introduce a tool that uses AI to create short summaries of religious texts. It works by finding the “centroid,” or main focus, of the document. Right after identifying this main topic of the text, the tool uses the "word embeddings" method to understand how words are related based on the content of that text. This helps the tool find sentences that best represent the main theme. After putting these sentences together, the tool creates a short, easy-to-read summary.

The good thing about this approach is that it captures deeper meanings in complex texts. This approach is not just counting words, it looks at how words and ideas are connected to each other.

This tool is especially helpful for summarizing long, detailed texts in religion and other humanities fields. It allows scholars, researchers, and anyone interested in these topics to quickly grasp the main ideas without having to read the whole text.

Method and Architecture:



Our approach for summarizing religious texts uses AI to find the main themes and ideas in long, detailed documents. The tool focuses on understanding the key points without including too much repetition. The system's design is organized into a few main steps, each with a specific role in processing the text and creating a clear, concise summary.

1. Input

Source Text: We start with the main document we want to summarize.

Document Loader: This part loads the text into the system so it can go through the summarization process.

Preprocessor: The preprocessor cleans up the text, removing things like extra spaces, stopwords (common but unimportant words), and making everything lowercase. This will allow us to ensure that only relevant content affects the summary.

2. Embedding

Word Embedding Generator: This part of the system represents each word as a "vector" (a kind of data format that helps the AI understand word meanings based on context). There will be pre-trained embeddings so that the system has a basic understanding.

Sentence Encoder: The sentence encoder combines the vectors of all the words in a sentence to create a

single representation of that sentence. This helps the system consider each sentence as a whole.

Centroid Calculator: The centroid calculator finds the main theme of the document (text, pdf, any kind of text) by calculating a “centroid” or central point.

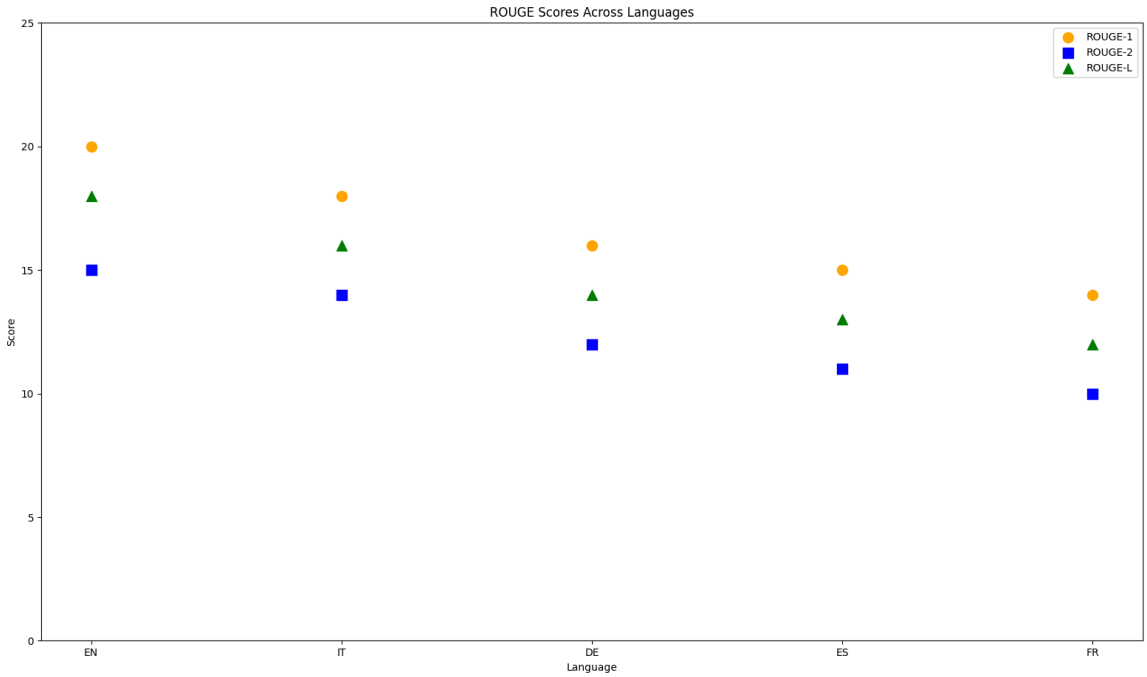
3. Summarization

Similarity Scorer: This scorer compares each sentence to the document’s main theme (the centroid) and gives each sentence a similarity score the score will define how relevant those sentences are to the main topic the higher the score is, the more relevant the sentence is to the main theme.

Redundancy Checker: In order to keep the summary clear and make sure there is no repetition in the summary we introduced the redundancy checker to review the sentences. If it finds sentences that are too similar, it keeps only one of them. This makes sure the summary stays focused and doesn’t repeat the same ideas.

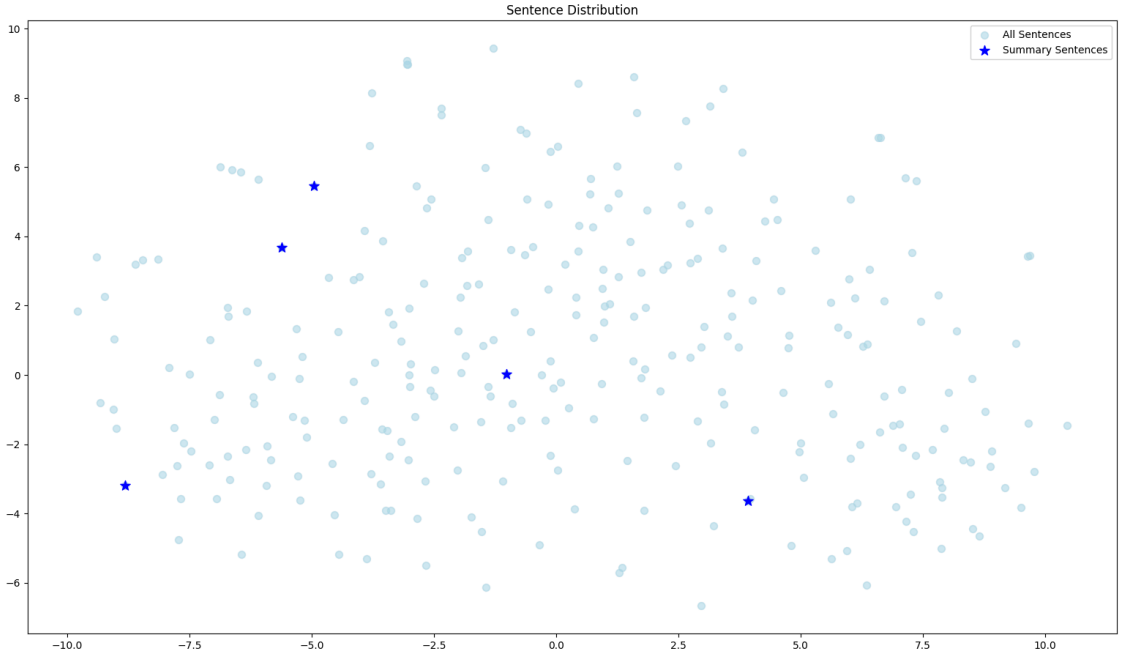
4. Output

Final Summary: After going though all these steps, we end up with a potential summarize version of the original text that will highlight the key points of our text.



2. **ROUGE Scores Plot** (Figure 1):

- This shows the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores across different languages
- Three different ROUGE metrics are shown:
    - ROUGE-1 (orange circles): Measures unigram overlap
    - ROUGE-2 (blue squares): Measures bigram overlap
    - ROUGE-L (green triangles): Measures longest common subsequence
- The x-axis shows different languages (EN, IT, DE, ES, FR)
- The y-axis shows scores from 0 to 25
- Trend shows:
    - Highest scores for English (EN)
    - Gradual decrease in scores for other languages
    - ROUGE-1 consistently scores highest, followed by ROUGE-L, then ROUGE-2



1. **Sentence Distribution Plot** (Figure 2):

- Light blue dots represent all sentences from the document
- Blue stars (*) represent the 5 sentences selected for the summary
- The axes range from -10 to 10 (x-axis) and -6 to 10 (y-axis)
- The scattered pattern shows how sentences are related to each other semantically:
    - Sentences that are closer together are more semantically similar
    - The 5 summary sentences (blue stars) are distributed across different regions, suggesting they capture different topics/aspects of the document
    - The dense areas indicate common themes or topics in the text