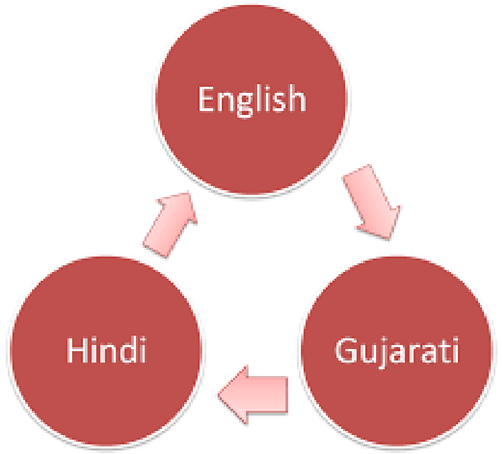
CROSS-LINGUAL INFORMATION RETRIEVAL SYSTEM FOR SHRIMAD BHAGAVAD GITA

ABOUT THE PROJECT



- A cross-lingual information retrieval mechanism for searching relevant documents related to Shrimad Bhagavad Gita.
- Will allow the user to enter queries in any of the three user-selected Indian languages: English, Hindi, and Gujarati, and retrieve documents in any of those languages as well.
- Uses a multilingual database of shlokas and returns optimal documents to the user.

01 PROBLEM STATEMENT

- Lack of reliable CLIR system for retrieving "Shrimad Bhagavad Gita" information in languages other than the text's language identified.
- Shlokas and their meanings present in different Indian languages, leading to differences in meaning of the same shloka.
- Search engine application required to ease the process of finding required shlokas and their meanings.
- Cross-lingual search engine needed to retrieve meanings in the user's mother tongue language.

02 LITERATURE REVIEW

CLIR SYSTEM

- **The main advantage of CLIR system to retreive information from Shrimad Bhagavad Gita in different languages is that enables users to access inforamtion that may not be available in their native language .**

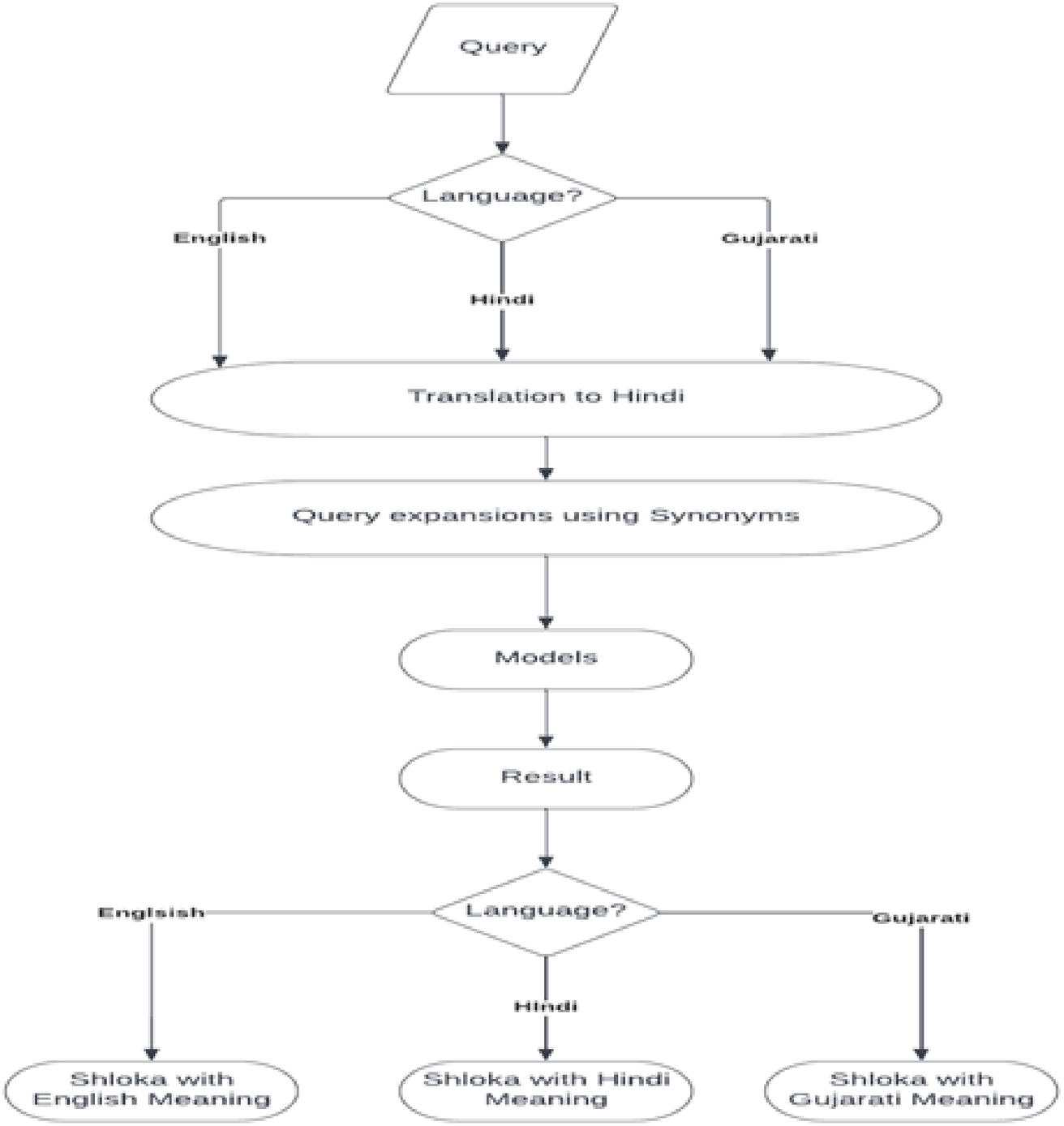
- **This will be helpful for researchers,scholars who want to study text in original language(Sanskrit) but may not be proficient have access to translations.**

- **Also CLIR system will be helpful to identify and retreive information across multiple languages which can be useful in cross lingual research.**

TECHNIQUES AND ALGORITHMS

- **Web Scraping and Data Cleaning: To collect and pre-process the data related to the texts of Gita in multiple languages.**
- **Query Translation: To translate user queries into the language of the text using machine translation and translation memory.**
- **Document Indexing: To create an inverted index of the preprocessed data and retrieve relevant documents based on the user's query.**
- **Relevance Ranking: To rank the retrieved documents based on their relevance to the user's query.**
- **Presentation: To present the results to the user in a user-friendly manner.**

04 STEPS


Query
Language?
English
Hindi
Gujarati
Translation to Hindi
Query expansions using Synonyms
Models
Result
Language?
Englsish
Hindi
Gujarati
Shloka with English Meaning
Shloka with Hindi Meaning
Shloka with Gujarati Meaning

Scoring Models Used (TF-IDF)

05

The formula used is as follows:

$$TF = \text{no of time a word occurs in a document}$$

$$idf = \ln \frac{N+1}{df_t + 1}$$

$$tfidf = tf * idf$$

where,

$$tf = \text{term frequency}$$

$$idf = \text{ document frequency}$$

$$N = \text{total no of documents}$$

$$df_t = \text{no of document in which a particular term is presnt}$$

Scoring Models Used (BM25)

06

The formula used is as follows:

$$\sum_{\forall t \in q} (1 + \ln(\frac{n - df_t + 0.5}{df_t + 0.5})) . \frac{(k1 + 1).(tf_d)}{tf_d + k1 * (1 - b + b * \frac{L_d}{L_{avg}})}$$

n – total documents

df_t – total no of ducuments in which term is present

tf_d – frequency of term in a document d

$k1, b$ – tuning parameters

RELATED WORK

07

Existing CLIR systems for the Bhagavad Gita. Some examples are Bhagavad Gita Search (https://bhagavadgitasearch.com/), Gita SuperSite (https://www.gitasupersite.iitk.ac.in/), Gita Press (https://www.gitapress.org/).

USER INTERFACE(WEB APP USING STREAMLIT)

- The Frontend web app will be developed using streamlit to sum up the entire work.
- User Interface will have plenty of features.
- User will be able to select the language in which he/she wishes to type the query.
- User will be able to choose the language in which he wishes to see the retrieved documents.

Thank You

Group-15

Manik Arora(2020519)
Amit Malik(2020493)
Talah Samar(2020343)
Shivanshu Pandey(2020333)
Ujjwal Goel(2020545)
Praveen Singh Samota(2020104)