Multilingual Mini Search Engine for Documents

Project submitted for the partial fulfillment of the requirements for the course

CSE 466: INFORMATION RETREIVAL

Offered by the

Department Computer Science and Engineering

School of Engineering and Sciences

Submitted by

THUPAKULA ANUSREE- AP22110011599

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

[DECEMBER, 2025]

Table of Contents

ABSTRACT

In the present digital era, large volumes of information are generated and stored in document formats such as PDF across multiple languages. Retrieving relevant information from these multilingual documents poses a significant challenge for traditional search engines, as most of them are designed to operate within a single language and lack support for effective cross-language search and translation. This limitation becomes especially evident in academic institutions, government organizations, and multinational environments where documents exist in diverse regional and global languages.

This project presents the design and implementation of a **Super Mini Multilingual Search Engine** capable of handling multilingual PDF documents with advanced language processing capabilities. The system allows users to upload one or more PDF files written in different languages, including Indian regional languages such as Telugu and Hindi, as well as international languages. Once uploaded, the system extracts textual content from the PDF files using reliable text extraction techniques and automatically detects the document's original language through statistical language detection methods.

To enable efficient searching, the detected content is translated into English and indexed using well-established Information Retrieval models. The system employs **Term Frequency–Inverse Document Frequency (TF-IDF)**, **BM25 ranking**, and a **hybrid ranking approach** that combines the strengths of both methods to produce highly relevant search results. Users can submit queries through a web-based interface and receive ranked document results based on semantic similarity and relevance scores.

An important feature of the system is its multilingual output capability. While indexing is performed in English to ensure consistency and performance, the system allows users to view search snippets and full document translations in multiple target languages of their choice. This enables users who are unfamiliar with the original document language to easily understand the content without manual translation. The translation module is designed to handle errors, rate limits, and fallback scenarios gracefully to ensure uninterrupted user experience.

The proposed search engine is implemented using Python and Flask, integrating libraries for natural language processing, vector space modeling, and translation services. The modular architecture ensures scalability, ease of maintenance, and future extensibility, such as adding OCR support for scanned PDFs or integrating deep learning–based language models. Overall, this project demonstrates an effective approach to multilingual document retrieval, bridging language barriers and providing a practical solution for cross-language information access in real-world applications.

1. INTRODUCTION

The exponential growth of digital data has made document management and information retrieval increasingly complex. A vast portion of today’s information is stored in Portable Document Format (PDF) files, especially in domains such as education, research, government documentation, and enterprise systems. While these documents are easily accessible in digital form, extracting meaningful information from them remains a significant challenge, particularly when the content spans multiple languages. Conventional search engines are largely monolingual and fail to provide efficient search and understanding when documents are written in diverse regional or foreign languages.

Multilingual document retrieval has emerged as an important research and application area within Information Retrieval (IR) and Natural Language Processing (NLP). In a country like India, where multiple languages such as Telugu, Hindi, Tamil, Kannada, and Malayalam are officially used alongside English, documents are frequently created in local languages. Users searching such documents often face language barriers, as they may not be familiar with the language in which the document was authored. This creates a strong need for systems that not only retrieve relevant documents but also understand and translate multilingual content effectively.

Traditional keyword-based search systems depend on exact term matching and are ineffective in handling cross-language queries. For example, a user may search for a keyword in English while the relevant document exists in Telugu or Hindi. Without multilingual indexing and translation mechanisms, such documents remain undiscovered. Moreover, PDFs present additional complexity since text extraction quality may vary depending on encoding, fonts, and formatting styles. Handling such unstructured and multilingual data requires careful preprocessing, normalization, and intelligent indexing techniques.

To address these challenges, this project proposes a **Super Mini Multilingual Search Engine** capable of extracting, indexing, searching, and translating content from multilingual PDF documents. The system is designed to bridge the gap between users and documents written in different languages by incorporating automatic language detection and cross-language translation. Once a PDF is uploaded, the system extracts its text and identifies the language of the document using statistical language detection algorithms. This detected language information is stored as metadata and displayed to the user for transparency and usability.

For effective search functionality, the extracted text is translated into English and indexed using state-of-the-art information retrieval models. English serves as the common indexing language due to the availability of robust NLP tools and libraries. The system employs **TF-IDF (Term Frequency–Inverse Document Frequency)** and **BM25**, both of which are widely used ranking algorithms in modern search engines. Additionally, a hybrid ranking approach is implemented to combine the strengths of both methods, ensuring accurate and relevant search results even for short or ambiguous queries.

A key feature of the proposed system is its multilingual output support. Users are not limited to viewing results in English alone; instead, they can choose their preferred target language for viewing translated snippets and full document content. This feature enhances accessibility for users from diverse linguistic backgrounds and significantly reduces the effort required to manually translate documents. Error handling mechanisms are included to manage translation failures, rate limits, and mixed-language documents, thereby increasing system robustness.

The project is implemented using Python with the Flask web framework, making it lightweight, modular, and easy to deploy. Libraries such as scikit-learn, NLTK, and translation APIs are integrated to support text processing, ranking, and multilingual translation. The system architecture is modular, allowing future enhancements such as OCR support for scanned PDFs, deep learning–based semantic search, and support for additional languages.

In summary, this project focuses on building a practical, user-friendly multilingual search engine that overcomes language and format barriers commonly encountered in document retrieval systems. By combining language detection, translation, and advanced ranking techniques, the system provides an efficient solution for accessing multilingual PDF content. The proposed approach is especially beneficial for academic institutions, government organizations, and multilingual knowledge repositories, where information accessibility across languages is crucial.

2.METHODOLOGY

.This chapter describes the systematic approach followed in designing and implementing the Super Mini Multilingual Search Engine. The methodology covers document ingestion, text extraction, language detection, multilingual processing, indexing, ranking, translation, and result presentation. The objective is to ensure accurate retrieval of relevant documents while supporting multilingual input and output.

2.1 System Overview

The proposed system follows a pipeline-based architecture where each stage processes the input and passes it to the next stage. The complete flow begins with PDF upload and ends with multilingual search results. The major stages include PDF preprocessing, text extraction, language detection, translation, indexing, ranking, and result visualization. Flask is used as the backend framework to coordinate all modules efficiently.

2.2 PDF Upload and Document Management

Users are allowed to upload one or multiple PDF documents through a web interface. Uploaded files are stored securely with unique identifiers to prevent filename conflicts. The system validates file formats and ensures only PDF documents are processed. Metadata such as filename, document ID, and file path are stored for indexing and future reference.

2.3 Text Extraction from PDF Files

The uploaded PDF documents are processed using PDF text extraction libraries capable of handling Unicode text. This step converts PDF content into raw text while preserving character encoding. Special care is taken to handle multilingual scripts such as Telugu and Hindi. Empty or corrupted extractions are filtered to avoid indexing invalid documents.

2.4 Language Detection of Documents

Once text is extracted, automatic language detection is performed. The system analyzes the largest chunk of extracted text to reduce misclassification caused by mixed-language content. Statistical language detection techniques are applied to identify the dominant language of each document. The detected language is stored as document metadata and displayed in the search results.

2.5 Text Normalization and Preprocessing

Before indexing, the extracted text undergoes normalization to improve retrieval effectiveness. This includes:

- Converting text to lowercase
- Unicode-aware tokenization
- Removal of noise characters
- Handling of mixed-language tokens

Unlike traditional systems that remove only English stopwords, the system avoids aggressive stopword elimination for non-English languages to prevent empty vocabulary errors. Token fallback techniques are applied to retain meaningful content even for low-resource languages.

2.6 Translation for Multilingual Indexing

To enable cross-language search, all extracted document text is translated into English, which serves as the base indexing language. Automatic translation APIs are used with `auto` source language detection. Translation exceptions and rate-limit failures are handled gracefully by falling back to the original text. This ensures uninterrupted indexing.

2.7 Index Construction

The system builds indexes using two ranking models:

2.7.1 TF-IDF Indexing

TF-IDF computes term importance based on frequency and inverse document occurrence. It is effective for identifying distinctive terms within documents and is suitable for keyword-driven searches.

2.7.2 BM25 Indexing

BM25 improves upon TF-IDF by handling document length normalization and term saturation. It provides better relevance ranking, especially for short search queries.

2.7.3 Hybrid Ranking Approach

A combined ranking strategy is implemented, where TF-IDF and BM25 scores are merged. This hybrid approach improves accuracy and relevance consistency across diverse query types.

2.8 Query Processing and Search Execution

When a user submits a query, the system preprocesses it using the same normalization pipeline as the documents. The query is translated into English if required and searched across the indexed corpus. Top-ranked documents are retrieved based on the selected ranking algorithm (TF-IDF, BM25, or hybrid).

2.9 Snippet Generation and Keyword Highlighting

For each retrieved document, a relevant snippet is extracted that best matches the user query. Query terms are highlighted within the snippet to provide visual emphasis and improve user understanding. This enhances result interpretability and speeds up information discovery.

2.10 Multilingual Result Translation

Users can select a target output language from the available options. The system translates:

- The result snippet
- The full document text (on demand)

Translation is performed dynamically, enabling users to read content in their preferred language regardless of the document’s original language.

2.11 Result Presentation and User Interface

Search results are displayed through a responsive web interface. Each result shows:

- Document title

- Detected language
- Relevance score
- Translated snippet
- Options to view original text or full translated text

The interface is designed to be intuitive and accessible for non-technical users.

2.12 Error Handling and System Robustness

Several safeguards are implemented to ensure system stability:

- Handling empty or invalid PDF text
- Fallback tokenization to prevent empty vocabulary errors
- Robust translation exception handling
- Graceful failure recovery without crashing the application

These measures make the system reliable under real-world conditions.

2.13 Implementation Tools and Technologies

- **Backend:** Python, Flask
- **Text Processing:** NLTK, Unicode Regex
- **Indexing & Ranking:** TF-IDF, BM25
- **Language Detection:** Statistical language detection
- **Translation:** Multilingual translation APIs
- **Frontend:** HTML, CSS, Jinja Templates

2.14 Methodology Summary

The methodology integrates multilingual processing with classical information retrieval techniques to build a powerful yet lightweight search engine. By combining language detection, translation, and hybrid ranking, the system enables efficient cross-language access to PDF-based knowledge sources.

3.PROJECT WORKFLOW

The Super Mini Multilingual Search Engine is a web-based information retrieval system designed to search, analyze, and translate content from PDF documents written in multiple languages. The primary motivation behind the project is to overcome language barriers in document search systems and enable users to retrieve relevant information regardless of the original language of the document.

Traditional search engines are often limited to monolingual content and fail when documents are written in regional or non-English languages such as Telugu, Hindi, or Tamil. Additionally, most systems do not support cross-language searching or full-text translation of retrieved documents. This project addresses these limitations by integrating language detection, multilingual translation, and advanced ranking models into a unified search framework.

The system allows users to upload one or more PDF documents containing text in any supported language. Once uploaded, the documents are processed through a multilingual pipeline that extracts text, detects the original language, translates the content for indexing, and stores the processed text in an efficient search index. Users can then submit search queries and receive ranked results along with translated snippets or full translated documents in their preferred language.

Key Objectives of the Project

The major objectives of the Super Mini Multilingual Search Engine are:

- To automatically detect the language of uploaded PDF documents.
- To support searching across documents written in different languages.
- To translate both search results and full document content into a user-selected language.
- To implement efficient ranking algorithms for accurate and relevant search results.
- To provide a lightweight and user-friendly web-based search interface.

System Architecture Overview

The system follows a modular architecture where each component performs a specific function. These components work together to ensure smooth data flow and reliable search performance.

1. **User Interface Module**
   Allows users to upload PDF documents, enter search queries, choose ranking methods, and select output languages.
2. **PDF Processing Module**
   Extracts Unicode text from uploaded PDF files and prepares it for further processing.
3. **Language Detection Module**
   Identifies the dominant language of each document using statistical analysis of textual content.
4. **Translation Module**
   Translates document text into English for indexing and translates search results into the user's chosen language.
5. **Search Engine Core**
   Implements TF-IDF, BM25, and hybrid ranking models for effective document retrieval.
6. **Result Presentation Module**
   Displays ranked results with language labels, scores, highlighted snippets, and translation options.

Working Flow of the System

1. The user uploads one or more PDF documents.
2. The system extracts text from each PDF.
3. The document’s original language is automatically detected.
4. The extracted text is translated to English for indexing.
5. TF-IDF and BM25 models index the translated text.
6. The user enters a search query.
7. The system retrieves and ranks relevant documents.
8. Snippets and full text are translated into the selected output language.
9. Results are displayed in a clear and readable format.

Technologies Used

- **Backend:** Python, Flask
- **Text Extraction:** PDF text extraction libraries
- **Language Detection:** Statistical language detection techniques
- **Translation:** Multilingual translation APIs
- **Search Models:** TF-IDF, BM25
- **Frontend:** HTML, CSS, Jinja Templates

Uniqueness of the Project

What makes this project unique is its ability to handle **cross-language document search**, where users can search in one language and retrieve results from documents written in another language. Additionally, it provides both snippet-level and full-document translation, making it especially useful for educational, legal, and research applications.

Applications of the System

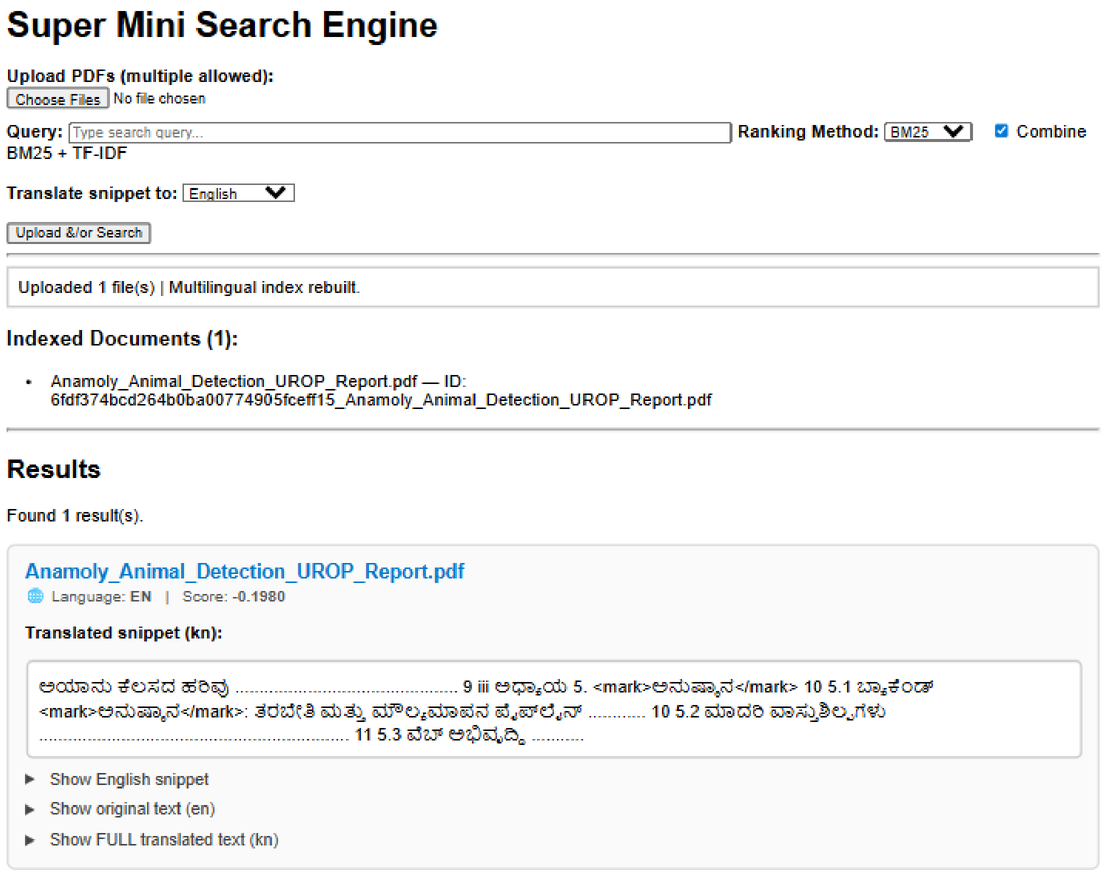
- Educational institutions for multilingual study materials
- Government document archives
- Digital libraries
- Research organizations
- Knowledge management systems

Project Outcome

The outcome of this project is a fully functional multilingual search engine capable of processing real-world PDF documents. It demonstrates how classical information retrieval techniques can be enhanced using multilingual processing to create more inclusive and accessible search systems.

4.OUTPUT

.



Super Mini Search Engine

Upload PDFs (multiple allowed):

Choose Files No file chosen

Query: Type search query... Ranking Method: BM25 ☑ Combine BM25 + TF-IDF

Translate snippet to: English

Upload &/or Search

Uploaded 1 file(s) | Multilingual index rebuilt.

Indexed Documents (1):

- Anamoly_Animal_Detection_UROP_Report.pdf — ID: 6fdf374bcd264b0ba00774905fceff15_Anamoly_Animal_Detection_UROP_Report.pdf

Results

Found 1 result(s).

Anamoly_Animal_Detection_UROP_Report.pdf

Language: EN | Score: -0.1980

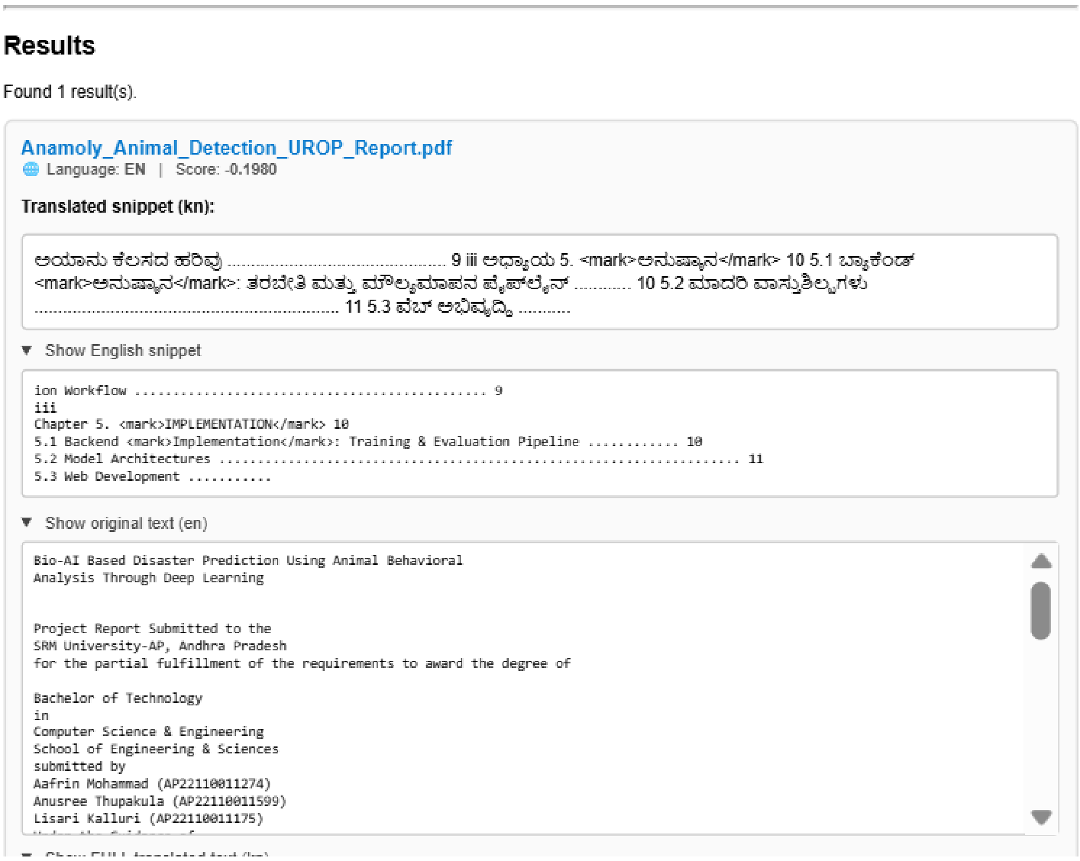
Translated snippet (kn):

ಅಯಾನು ಕೆಲಸದ ಹರಿವು .... 9 iii ಅಧ್ಯಾಯ 5. <mark>ಅನುಷ್ಠಾನ</mark> 10 5.1 ಬ್ಯಾಕೆಂಡ್ <mark>ಅನುಷ್ಠಾನ</mark>: ತರಬೇತಿ ಮತ್ತು ಮೌಲ್ಯಮಾಪನ ಪೈಪ್‌ಲೈನ್ .... 10 5.2 ಮಾದರಿ ವಾಸ್ತುಶಿಲ್ಪಗಳು .... 11 5.3 ವೆಬ್ ಅಭಿವೃದ್ಧಿ ....

▶ Show English snippet

▶ Show original text (en)

▶ Show FULL translated text (kn)



Results

Found 1 result(s).

Anamoly_Animal_Detection_UROP_Report.pdf
Language: EN | Score: -0.1980

Translated snippet (kn):

ಅಯಾನು ಕೆಲಸದ ಹರಿವು .... 9 iii ಅಧ್ಯಾಯ 5. <mark>ಅನುಷ್ಠಾನ</mark> 10 5.1 ಬ್ಯಾಕೆಂಡ್ <mark>ಅನುಷ್ಠಾನ</mark>: ತರಬೇತಿ ಮತ್ತು ಮೌಲ್ಯಮಾಪನ ಪೈಪ್‌ಲೈನ್ .... 10 5.2 ಮಾದರಿ ವಾಸ್ತುಶಿಲ್ಪಗಳು .... 11 5.3 ವೆಬ್ ಅಭಿವೃದ್ಧಿ ....

▼ Show English snippet

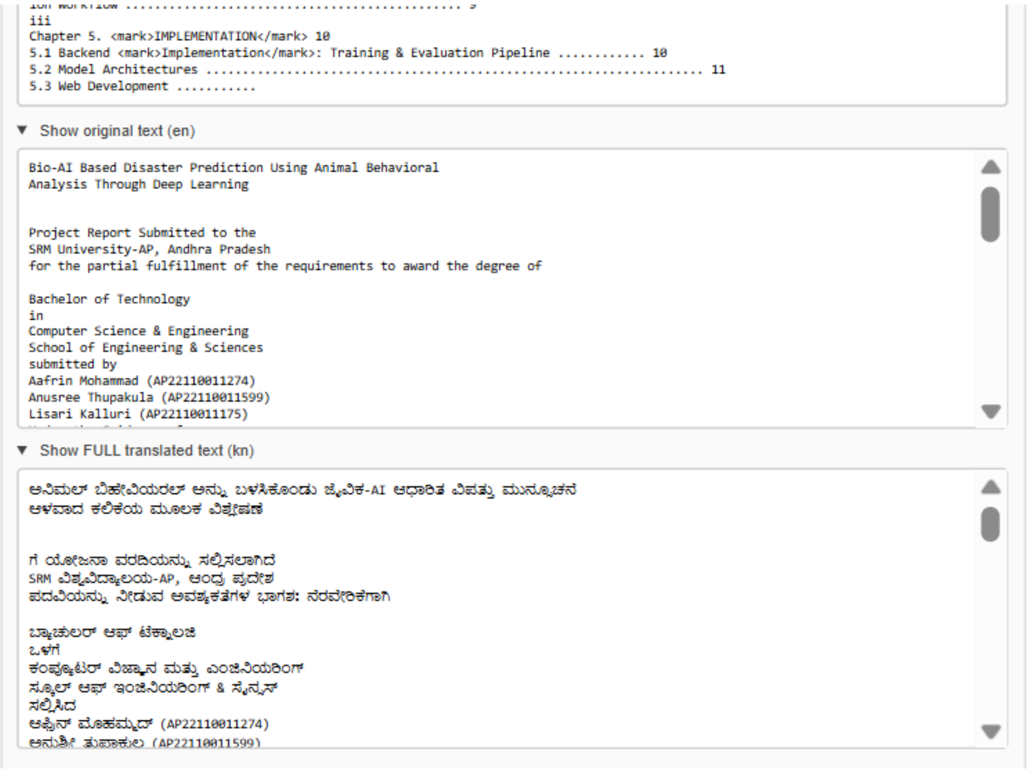
```
ion Workflow ............................................... 9
iii
Chapter 5. <mark>IMPLEMENTATION</mark> 10
5.1 Backend <mark>Implementation</mark>: Training & Evaluation Pipeline ............ 10
5.2 Model Architectures ..................................................................... 11
5.3 Web Development ...........
```

▼ Show original text (en)

```
Bio-AI Based Disaster Prediction Using Animal Behavioral
Analysis Through Deep Learning


Project Report Submitted to the
SRM University-AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
in
Computer Science & Engineering
School of Engineering & Sciences
submitted by
Aafrin Mohammad (AP22110011274)
Anusree Thupakula (AP22110011599)
Lisari Kalluri (AP22110011175)
```



```
iii
Chapter 5. <mark>IMPLEMENTATION</mark> 10
5.1 Backend <mark>Implementation</mark>: Training & Evaluation Pipeline ............ 10
5.2 Model Architectures ..................................................................... 11
5.3 Web Development ...........
```

▼ Show original text (en)

```
Bio-AI Based Disaster Prediction Using Animal Behavioral
Analysis Through Deep Learning


Project Report Submitted to the
SRM University-AP, Andhra Pradesh
for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology
in
Computer Science & Engineering
School of Engineering & Sciences
submitted by
Aafrin Mohammad (AP22110011274)
Anusree Thupakula (AP22110011599)
Lisari Kalluri (AP22110011175)
```

▼ Show FULL translated text (kn)

```
ಅನಿಮಲ್ ಬಿಹೇವಿಯರಲ್ ಅನ್ನು ಬಳಸಿಕೊಂಡು ಜೈವಿಕ-AI ಆಧಾರಿತ ವಿಪತ್ತು ಮುನ್ಸೂಚನೆ
ಆಳವಾದ ಕಲಿಕೆಯ ಮೂಲಕ ವಿಶ್ಲೇಷಣೆ


ಗೆ ಯೋಜನಾ ವರದಿಯನ್ನು ಸಲ್ಲಿಸಲಾಗಿದೆ
SRM ವಿಶ್ವವಿದ್ಯಾಲಯ-AP, ಆಂಧ್ರ ಪ್ರದೇಶ
ಪದವಿಯನ್ನು ನೀಡುವ ಅವಶ್ಯಕತೆಗಳ ಭಾಗಶಃ ನೆರವೇರಿಕೆಗಾಗಿ

ಬ್ಯಾಚುಲರ್ ಆಫ್ ಟೆಕ್ನಾಲಜಿ
ಒಳಗೆ
ಕಂಪ್ಯೂಟರ್ ವಿಜ್ಞಾನ ಮತ್ತು ಎಂಜಿನಿಯರಿಂಗ್
ಸ್ಕೂಲ್ ಆಫ್ ಇಂಜಿನಿಯರಿಂಗ್ & ಸೈನ್ಸ್
ಸಲ್ಲಿಸಿದ
ಆಫ್ರಿನ್ ಮೊಹಮ್ಮದ್ (AP22110011274)
ಅನುಶ್ರೀ ತುಪಾಕುಲ (AP22110011599)
```

5.CONCLUSION

The Super Mini Multilingual Search Engine successfully addresses the challenges associated with searching and understanding documents written in multiple languages. By integrating language detection, translation, and advanced information retrieval techniques, the system enables users to search across multilingual PDF documents with improved accessibility and usability. This project demonstrates that traditional search methodologies can be effectively enhanced to support cross-language information retrieval.

The system accurately detects the original language of uploaded documents and ensures that the indexing process remains consistent by translating all content into a common language. The use of TF-IDF and BM25 ranking algorithms ensures relevant and meaningful search results, while the hybrid ranking approach improves accuracy further. Additionally, providing translated snippets and full translated documents enhances user comprehension and makes the system practical for real-world applications.

One of the key strengths of the project is its modular and scalable design, which allows easy integration of additional languages, ranking models, or document formats in the future. The web-based interface further simplifies interaction, making the system accessible to users with minimal technical knowledge.

Overall, this project highlights the importance of multilingual support in modern information systems and serves as a strong foundation for advanced multilingual search engines. It proves that effective document retrieval and translation can coexist within a single platform, offering a valuable solution for educational, research, and organizational environments where language diversity is a significant challenge.