Medical Report Text Summarization

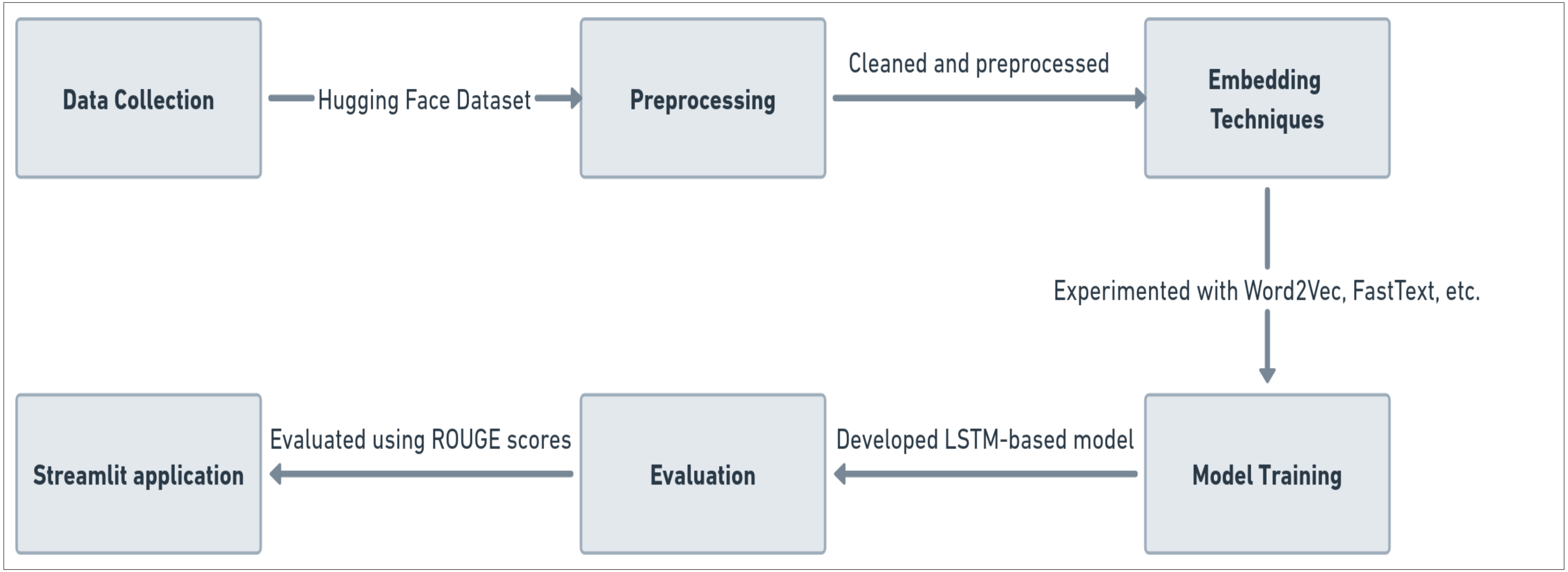
TEAM MEMBERS

- Keerthana G – AI & DS-*A*
- Karthick N G – AI & DS-*A*

Project Description

- Our project aims to address the challenge of efficiently summarizing medical reports using Natural Language Processing (NLP) techniques.
- The project involves experimenting with both extractive and abstractive summarization methods to find the most effective approach.
- As a learning project, we tried various embedding techniques and approaches such as page rank algorithm, Deep Learning techniques, etc.,
- ***Final Application:*** We developed a user-friendly web application that allows users to input medical reports and receive concise summaries generated by our model. This application takes length of the expected summary and medical report as input.

Pipeline



Data Collection
Hugging Face Dataset
Preprocessing
Cleaned and preprocessed
Embedding Techniques
Experimented with Word2Vec, FastText, etc.
Model Training
Developed LSTM-based model
Evaluation
Evaluated using ROUGE scores
Streamlit application

Embedding Techniques Used

1. Word2Vec:

- Explored both Continuous Bag of Words (CBOW) and Skip-gram variants of Word2Vec embedding.
  - ***CBOW***: Focuses on predicting a target word given its context, suitable for capturing semantic relationships within medical texts.
  - ***Skip-gram***: Predicts context words given a target word, beneficial for handling polysemy and capturing diverse word usage.

2. FastText:

  - Utilized FastText embedding to handle out-of-vocabulary words and enhance representation learning.

Justification:

- These embedding techniques were chosen for their ability to capture the semantic and syntactic nuances present in medical reports, which is crucial for effective summarization.

Word2Vec for Extractive Summarization:

- Word2Vec's ability to capture semantic relationships between words makes it well-suited for extractive summarization, where identifying important sentences based on word similarity is key. This approach allows us to extract sentences that closely match the context of the original text, thereby producing more coherent summaries.

Model Used and Its Architecture

Model Architecture:

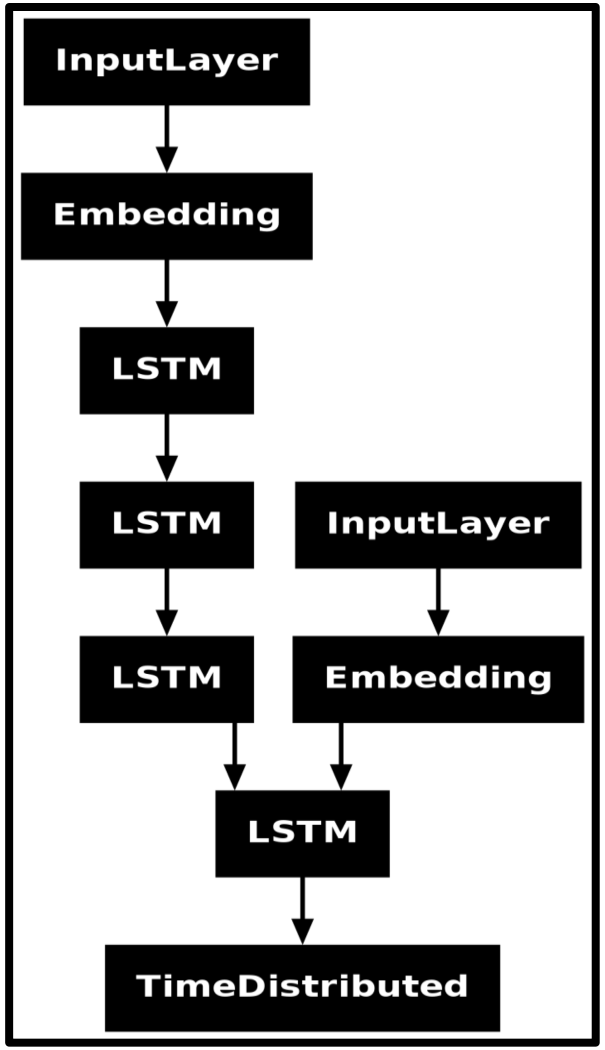
- Implemented a sequence-to-sequence model, a popular architecture for tasks involving sequential data processing.
- Utilized Long Short-Term Memory (LSTM) layers for both the encoder and decoder components.
- LSTM layers were chosen for their ability to capture long-range dependencies and handle sequential data effectively.

Encoder:

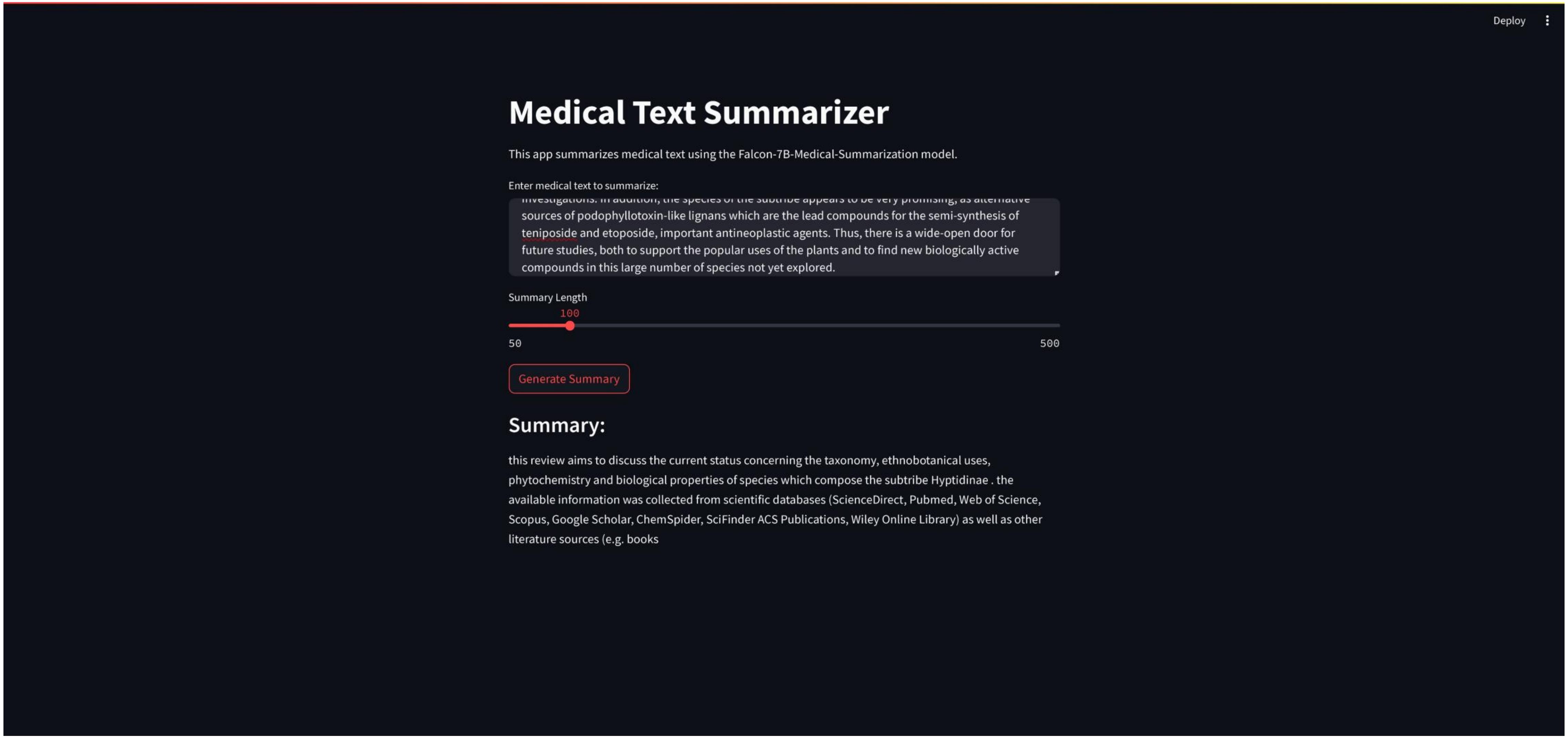
- The encoder component captures contextual information from input medical reports by processing the input sequence step-by-step.
- It converts the input text into a fixed-dimensional representation, which serves as the context vector for the decoder.

Decoder:

- The decoder component generates summary tokens based on the encoded representation of input sequences provided by the encoder.
- It uses the context vector to generate summaries token by token, incorporating the contextual information learned by the encoder.



Results


Deploy
Medical Text Summarizer
This app summarizes medical text using the Falcon-7B-Medical-Summarization model.
Enter medical text to summarize:
sources of podophyllotoxin-like lignans which are the lead compounds for the semi-synthesis of teniposide and etoposide, important antineoplastic agents. Thus, there is a wide-open door for future studies, both to support the popular uses of the plants and to find new biologically active compounds in this large number of species not yet explored.
Summary Length
100
50
500
Generate Summary
Summary:
this review aims to discuss the current status concerning the taxonomy, ethnobotanical uses, phytochemistry and biological properties of species which compose the subtribe Hyptidinae . the available information was collected from scientific databases (ScienceDirect, Pubmed, Web of Science, Scopus, Google Scholar, ChemSpider, SciFinder ACS Publications, Wiley Online Library) as well as other literature sources (e.g. books

Future Scope

Fine-tuning:

- Explore opportunities to fine-tune the model with additional medical data to further enhance summarization quality.

Advanced Architectures:

- Investigate the integration of advanced deep learning architectures such as transformer-based models for improved performance.

Domain-Specific Knowledge:

- Incorporate domain-specific knowledge or ontologies to enhance the model's understanding of medical terminology.

Customization Options:

- Provide customization options within the UI, allowing users to adjust summarization parameters or preferences based on their specific needs.
- Include options for selecting preferred language, or domain-specific terminology like identifying symptoms to diseases to tailor the summary according to user requirements.

Real-time Feedback:

- Implement real-time feedback mechanisms to provide users with insights into the summarization process.