ML-BASED MEDICAL DIAGNOSIS SYSTEM

PRESENTED BY

Vineet Parmar

Omkar Rao

Hatim Sawai

UNDER GUIDANCE OF

Prof. Varsha Hole

November 22nd, 2024

Content Summary

1 Problem Statement

2 Objectives

3 Scope

4 Literature Survey

5 Design and Methodology

6 Implementation Details

7 Results and Discussion

8 References

Problem Statement

- In contemporary healthcare systems, **linguistic diversity poses a significant barrier** to efficient patient-doctor communication and accurate diagnosis, particularly in regions where languages like Hinglish are prevalent.

- Existing solutions often struggle to effectively process and analyze patient-reported medical conditions expressed in **mixed-language formats,** hindering timely diagnosis and treatment.

Objectives

- Develop a robust **language processing pipeline** capable of handling Hinglish input, including **language identification, phrase grouping, translation, and transliteration.**

- Implement accurate translation mechanisms from **Hinglish to Hindi Devnagri script and vice versa,** ensuring preservation of meaning and context.

- **Integrate specialized dictionaries or databases** to identify and replace **medical keywords in both Hindi and English,** enriching the input text with relevant medical symptoms.

- Design and deploy a **Named Entity Recognition (NER) module** tailored for identifying biological entities and medical symptoms within the processed text.

- Utilize **Large Language Models (LLMs)** for generating accurate diagnoses and suggesting appropriate remedies based on the identified symptoms and medical context.

Objectives (cont’d)

- Evaluate the performance of the language processing pipeline, **translation mechanisms**, **keyword identification**, **NER module**, and LLMs through rigorous testing **against diverse datasets** and real-world scenarios.

- Optimize the **computational efficiency and scalability** of the entire system to handle large volumes of patient data and ensure timely responses.

- Collaborate with **healthcare professionals** to **validate the accuracy** and effectiveness of the **generated diagnoses** and remedy suggestions, incorporating feedback to **improve system performance.**

- Document the entire development process, including methodologies, algorithms, and tools used, to facilitate reproducibility and future enhancements. Additionally, providing comprehensive user documentation to support adoption and usage by healthcare practitioners.

Scope

The Following listed will be the boundaries of the project that the project will comprise of:

The system caters specifically to users comfortable in Hinglish, a blend of Hindi and English.

It’s important to remember that this system serves as an informative tool, offering a preliminary analysis to empower patients but not replacing the need for professional medical evaluation and diagnosis.

Literature Survey

NAME & YEAR	AUTHORS	WORK	TECHNIQUES
BioBERT Based Named Entity Recognition in Electronic Medical Record, 2019	• X. Yu • W. Hu • S. Lu • X. Sun • Z. Yuan	They have covered codemixed input text summarization in a medical setting using MMCQs dataset , which combines Hindi-English codemixed medical queries with visual aids. They have introduced a framework named MedSumm that leverages the power of LLMs and VLMs for this task.	ML Models Used: • MedSumm
Classification of Patient Portal Messages with BERT-based Language Models, 2023	Y. Ren	This paper proposes a pipelined mechanism for machine translation of a bi-lingual language i.e. Hinglish to monolingual English in this paper.	Python Libraries Used: Nltk Spacy
Disease Prediction using Machine Learning, 2022	• N. Kosarkar • P. Basuri • P. Karamore • P. Gawali • P. Badole • P. Jumle	They have have proposed a Language Modelling (LM) based approach to text classification of Hinglish text. We approach this problem by building a Universal Language Model Fine-tuning using AWD-LSTM architecture on a Hindi-English code-switched (Hinglish) corpus collected from various blogging sites.	Architecture Used: AWD-LSTM
Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning, 2019	• R. B. Mathew • S. Varghese • S. E. Joy • S. S. Alex	They have created a python library for clinical texts, EHRKit. This library contains two main parts: MIMIC-III-specific functions and task-specific functions. The first part introduces a list of interfaces for accessing MIMIC-III NOTEEVENTS data, including basic search, information retrieval, etc.	NLP Libraries Used: • MIMIC-Extract • ScispaCy • medspaCy • Stanza Biomed • SciFive • EHRKit (ours)
Human Disease Prediction And Doctor Booking System, 2023	• Joel Roy • Reeju Koshy • Roshan Roy • Anjumol Zachariah	They have we propose a supervised learning method that can be used for much special domain NER tasks. The model consists of two parts, a multidimensional self-attention (MDSA) network and a CNN-based model.	ML Model Architecture Used: MDSA-CNN

Literature Survey

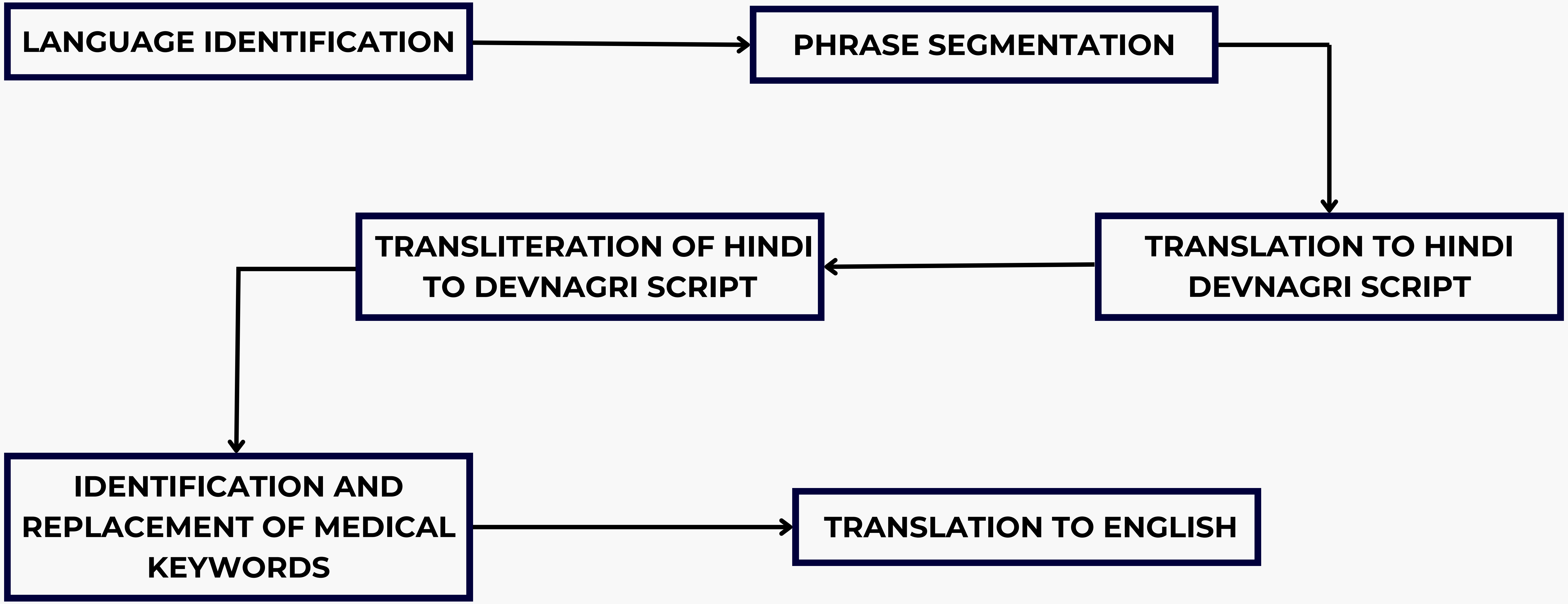
NAME & YEAR	AUTHORS	WORK	TECHNIQUES
MedSumm: A Multimodal Approach to Summarizing Code-Mixed Hindi-English Clinical Queries, 2024	• Akash Ghosh • Arkadeep Acharya • Prince Jha • Aniket Gaudgaul	They have used a recently introduced pre-trained language model BERT for named entity recognition in electronic medical records to solve the problem of missing context information and we add an extra mechanism to capture the relationship between words.	BERT-Based Named Entity Recognition in Chinese Electronic Medical Record
Code-Mixed Hinglish to English Language Translation Framework, 2022	IEEE Conference Publication	This paper examines if using semantic features and word context improves portal message classification. Materials and methods: ortal messages were classified into the following categories: informational, medical, social, and logistical. We constructed features from portal messages including bag of words, bag of phrases, graph representations, and word embeddings	• random forest • logistic regression classifiers • convolutional neural network (CNN) with a softmax output.
Machine Learning based Language Modelling of Code Switched Data, 2020	IEEE Conference Publication	they have In introduced a system which is trained on sentences consisting of various symptoms and later by using the dataset consisting of disease and the set of symptoms they possess the most probable disease the user may be suffering from is determined.	NLP Techniques used: • NER • SVM
EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts, 2023	• Irene Li • Keen You • Yujie Qiao • Lucas Huang	they have In introduced a system which is trained on sentences consisting of various symptoms and later by using the dataset consisting of disease and the set of symptoms they possess the most probable disease the user may be suffering from is determined.	NLP Techniques used: • NER • SVM
Multidimensional self-attention for aspect term extraction and biomedical named entity recognition, 2020	• X. Song • A. Feng • W. Wang • Z. Gao	This project aims to develop a portal for predicting disease according to the symptoms which is given by the user and an option for consulting doctor.	• Decision Tree • Naive Bayes • Random Forest

Literature Survey

NAME & YEAR	AUTHORS	WORK	TECHNIQUES
Design and Application of Intelligent Language Translation Software, 2023	• Akash Ghosh • Arkadeep Acharya • Prince Jha • Aniket Gaudgaul	The paper introduces intelligent English translation software utilizing natural language processing and artificial neural networks. Techniques include text analysis, matching calculations, and database management within a B/S architecture. Experiments show improved accuracy and speed compared to traditional tools, emphasizing AI's role in enhancing translation software.	Preliminary text analysis, Matching calculations for intelligent translation
Text Translation for Indian Languages, 2023	IEEE Conference Publication	The paper presents an LSTM based language translation model for Indian languages, aiming to bridge linguistic barriers. It employs an encoder-decoder architecture trained on a large dataset, showcasing improved translation accuracy	LSTM-based model, Encoder-decoder architecture, Large dataset training
Machine Learning Based Komering Language Translation Engine with Bidirectional RNN Model Algorithm, 2023	IEEE Conference Publication	Data collection involves scanning a Komering dictionary and distributing questionnaires. Pre-processing includes stopword removal, normalization, tokenization, and padding. A bidirectional RNN model is then trained using the preprocessed data.	Bidirectional RNN for modeling, data collection via questionnaires and scanning
Evolution of Machine Translation for Indian Regional Languages using Artificial Intelligence , 2023	• Irene Li • Keen You • Yujie Qiao • Lucas Huang	The research team developed a Neural Machine Translation (NMT) model focusing on English to Bengali, Punjabi, and Tamil transliteration. Training utilized parallel corpora for each language pair, with careful adjustment of hyperparameters to optimize performance.	Language detection, text normalization
Many-to-Many Multilingual Translation Model for Languages of Indonesia, 2023	• X. Song • A. Feng • W. Wang • Z. Gao	Developing a many-to-many multilingual translation model for Indonesian languages involves fine-tuning the pretrained mT5 model on religious texts, followed by further specialization on social media texts. Training employs a text-to-text approach, with evaluation using SacreBLEU metric.	Fine-tuning pretrained models, text-to-text translation framing, sequence alignment for verse pairs

DESIGN AND METHODOLOGY

1. HINGLISH TO ENGLISH ✓



LANGUAGE IDENTIFICATION

PHRASE SEGMENTATION

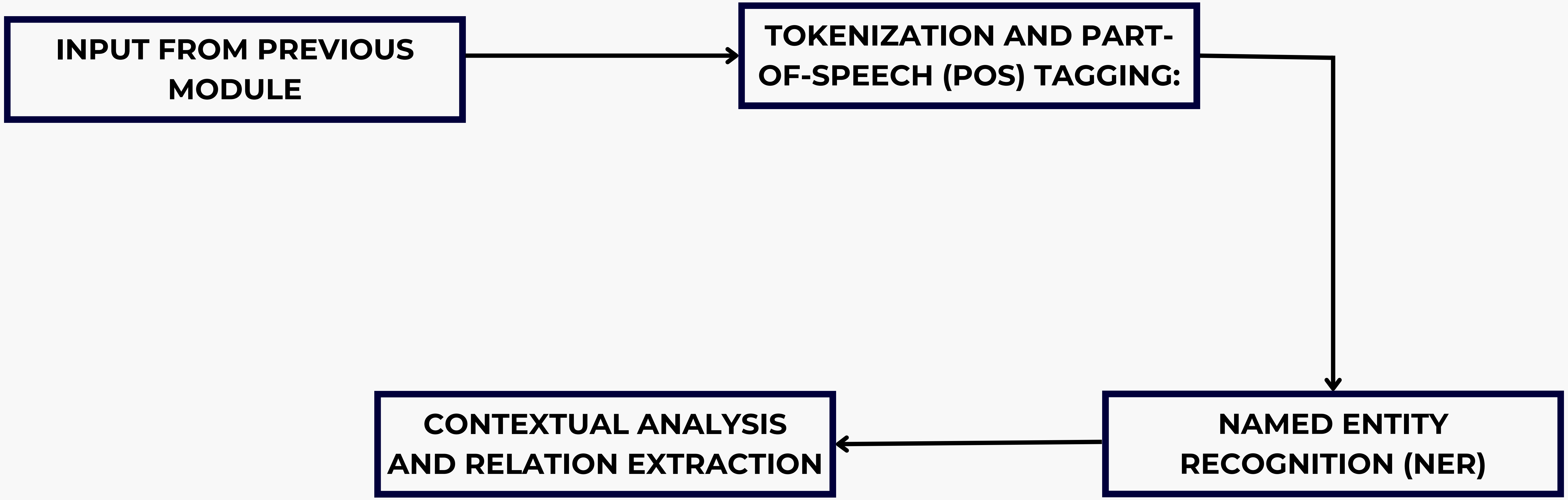
TRANSLATION TO HINDI DEVNAGRI SCRIPT

TRANSLITERATION OF HINDI TO DEVNAGRI SCRIPT

IDENTIFICATION AND REPLACEMENT OF MEDICAL KEYWORDS

TRANSLATION TO ENGLISH

2. Biological Named Entity Recognition (NER)



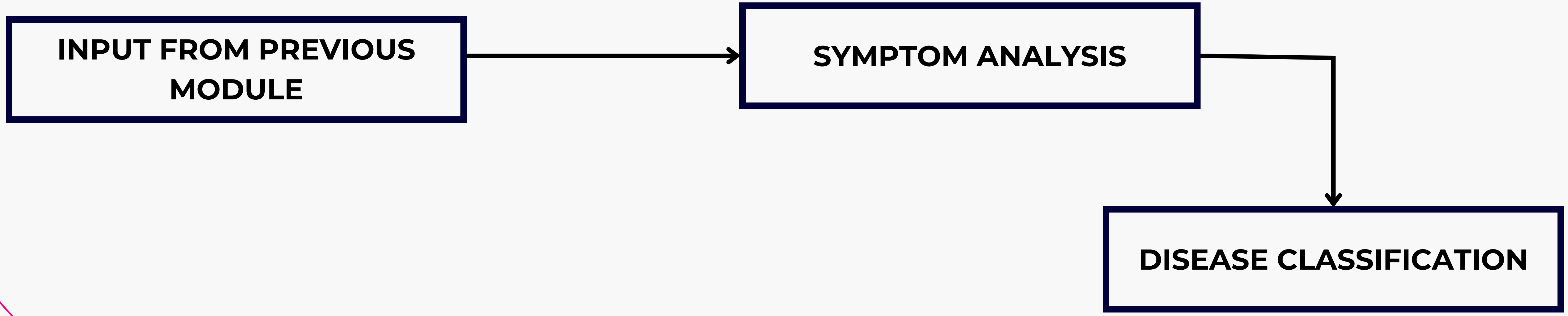
INPUT FROM PREVIOUS MODULE

TOKENIZATION AND PART-OF-SPEECH (POS) TAGGING:

NAMED ENTITY RECOGNITION (NER)

CONTEXTUAL ANALYSIS AND RELATION EXTRACTION

3. Disease Diagnosis



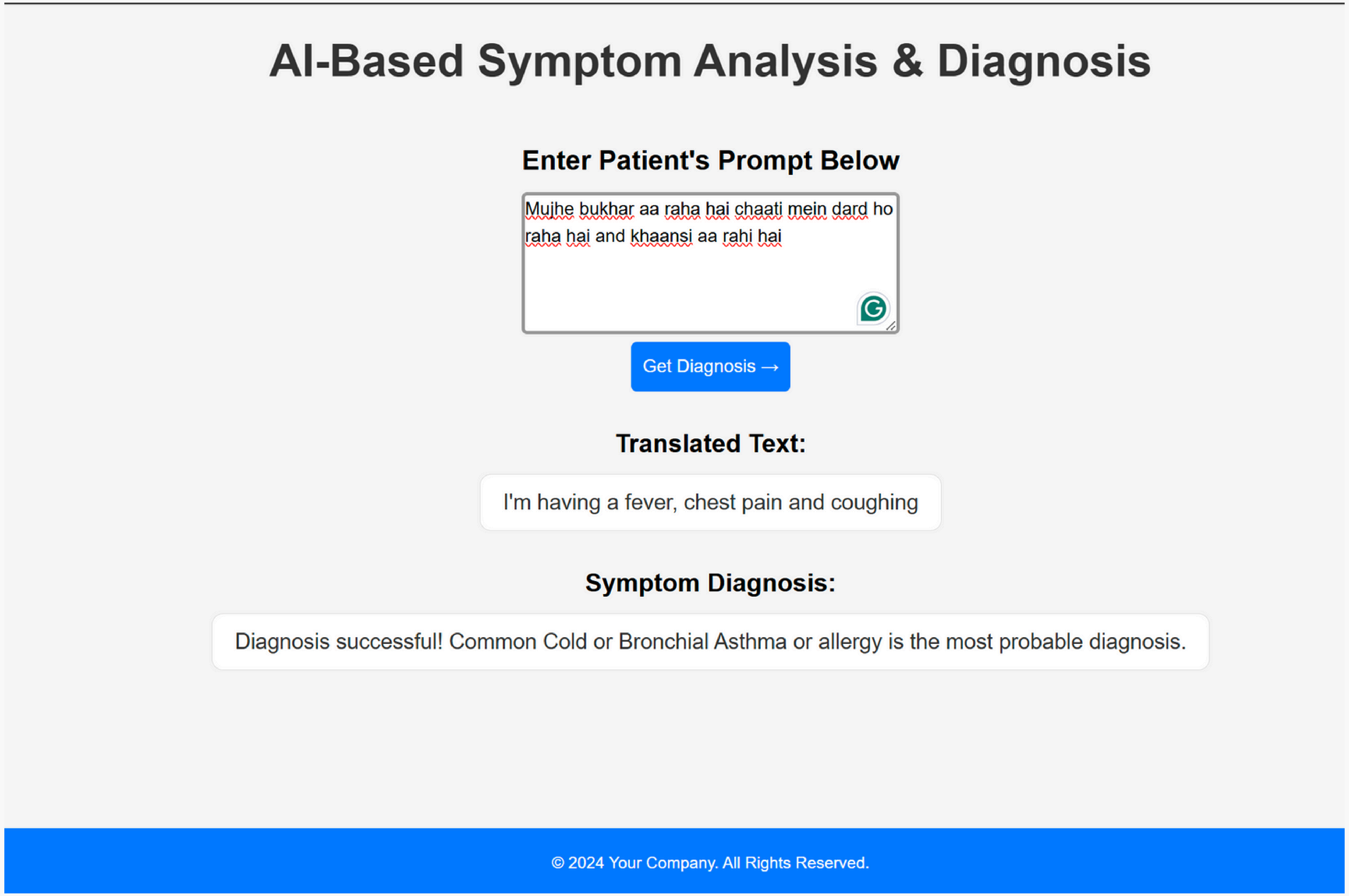
INPUT FROM PREVIOUS MODULE

SYMPTOM ANALYSIS

DISEASE CLASSIFICATION

IMPLEMENTATION DETAILS

Frontend Design



AI-Based Symptom Analysis & Diagnosis

Enter Patient's Prompt Below

Mujhe bukhar aa raha hai chaati mein dard ho raha hai and khaansi aa rahi hai

Translated Text:

I'm having a fever, chest pain and coughing

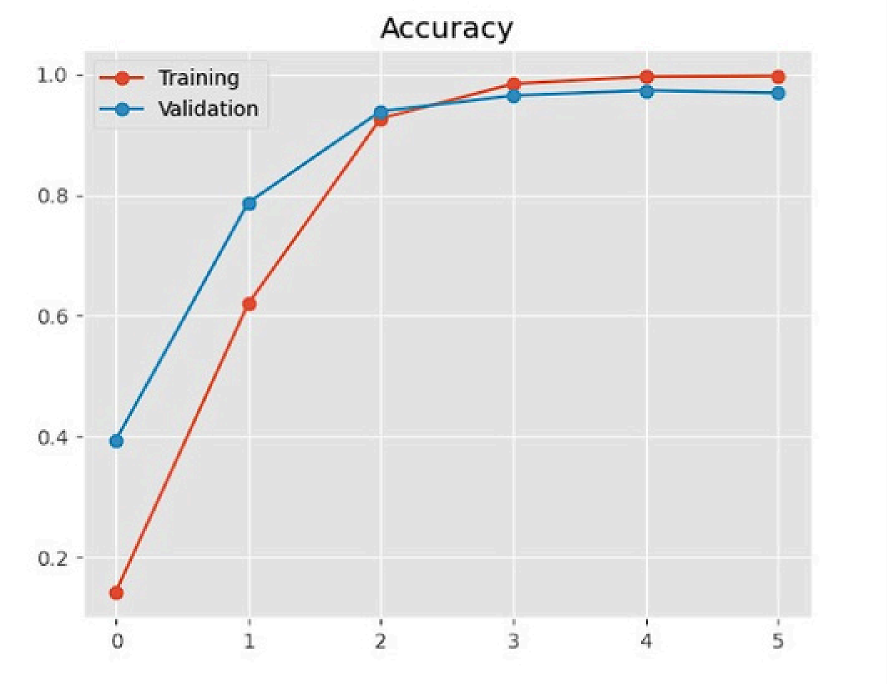
Symptom Diagnosis:

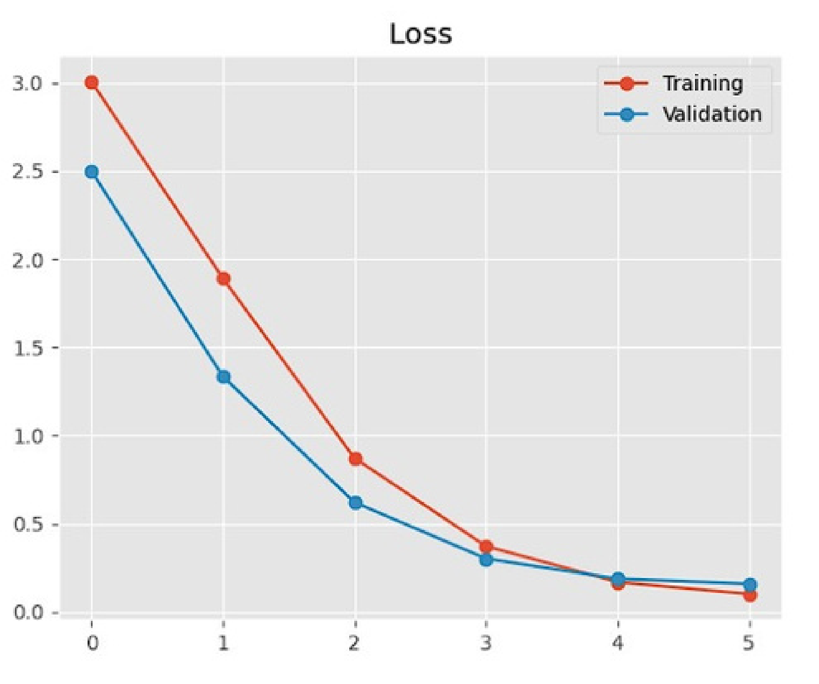
Diagnosis successful! Common Cold or Bronchial Asthma or allergy is the most probable diagnosis.

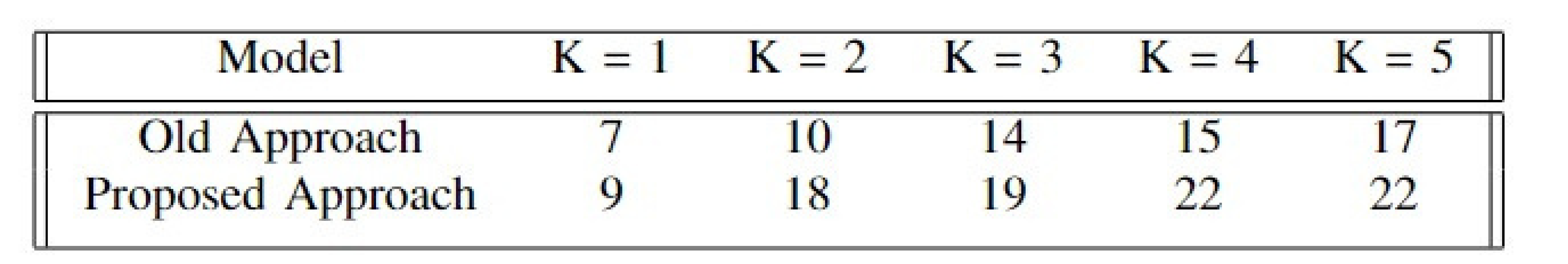
© 2024 Your Company. All Rights Reserved.

DISEASE DIAGNOSIS

The fine-tuned BERT model was trained and the quantitative results are shown below, also compares the diagnosis provided by the model after mapping medical words with the earlier model, which did not have specialized mapping for top K predictions, from K = 1 to 5.


Accuracy


Loss



Model	K = 1	K = 2	K = 3	K = 4	K = 5
Old Approach	7	10	14	15	17
Proposed Approach	9	18	19	22	22

Datasets used

ENGLISH WORDS DATASET

We utilized the Google Most Frequent Words dataset, containing a comprehensive list of commonly used English words.

HINDI ROMANIZED WORDS DATASET

The Dakshina Dataset was employed to identify Romanized Hindi words within the input text.

MEDICAL KEYWORDS DATASET

For identifying medical keywords in both Hindi and English, we utilized the Hindi Health Dataset.

Datasets used Cont.

CROWD INDIC TRANSLITERATION DATASET

It contains Hinglish-English transliteration pairs, to accurately map phonetic representations of Hinglish terms to English spellings, enabling seamless preprocessing for NLP tasks.

SYMPTOM2DISEASE DATASET

It provides symptom descriptions linked to diseases, as a foundation for mapping symptoms to potential diagnoses in healthcare analysis.

GENERATED DATASET OF HINGLISH SYMPTOM DESCRIPTION

Contains Hinglish symptom descriptions with diagnoses, to train and evaluate models for Hinglish-specific symptom recognition and disease prediction tasks.

Models Explored

LANGUAGE IDENTIFICATION

Two models were explored for language identification: Long Short-Term Memory (LSTM) networks and Logistic Regression.

TRANSLATION AND TRANSLITERATION:

- **English to Hindi Devnagri Script Translation:** Google Neural Machine Translation (GNMT) model.
- **Hindi Romanized to Devnagri Script Transliteration:** IndicTrans model.
- **Hindi Devnagri Script to English Translation:** GNMT model.

REFERENCES

[1] Ghosh, A., & Acharya, A. (2024, January 3). MedSumm: A Multimodal Approach to Summarizing Code-Mixed Hindi-English Clinical Queries. Retrieved from https://arxiv.org/pdf/2401.01596.pdf

Code-Mixed Hinglish to English Language Translation Framework. (2022, April 7). IEEE Conference Publication — IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/document/9760834

Machine Learning based Language Modelling of Code Switched Data.(2020, July 1). IEEE Conference Publication — IEEE Xplore. Retrieved from https://ieeexplore.ieee.org/document/9155695

Li, You, & Qiao. (2023, June 28). EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts. Retrieved from https://arxiv.org/pdf/2204.06604.pdf

X. Song, A. Feng, W. Wang and Z. Gao, 'Multidimensional self-attention for aspect term extraction and biomedical named entity recognition' Mathematical Problems in Engineering, vol. 2020, pp. 1–6, Dec. 2020.

REFERENCES

1. X. Yu, W. Hu, S. Lu, X. Sun and Z. Yuan, ‘BioBERT Based Named Entity Recognition in Electronic Medical Record’ 2019 10th International Conference on Information Technology in Medicine and Education (ITME), Qingdao, China, 2019, pp. 49–52, doi: 10.1109/ITME.2019.00022.
2. Y. Ren et al., ‘Classification of Patient Portal Messages with BERT-based Language Models’ 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, USA, 2023, pp. 176–182, doi:10.1109/ICHI57859.2023.00033.
3. 8. N. Kosarkar, P. Basuri, P. Karamore, P. Gawali, P. Badole and P. Jumle, ‘Disease Prediction using Machine Learning’ 2022 10th International Conference on Emerging Trends in Engineering and Technology – Signal and Information Processing (ICETET-SIP-22), Nagpur, India, 2022, pp. 1–4, doi: 10.1109/ICETET-SIP-2254415.2022.9791739.
4. R. B. Mathew, S. Varghese, S. E. Joy and S. S. Alex, ‘Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning’ 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019
5. Mr. Joel Roy, Mr. Reeju Koshy, Mr. Roshan Roy, Ms. Anjumol Zachariah, 2023, Human Disease Prediction And Doctor Booking System, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH and TECHNOLOGY (IJERT), Volume 11, Issue 01 (June 2023)