AI-powered Diagnosis System with Virtual Consults and Multimodal Analysis

submitted in partial fulfillment of the requirement
for the award of the Degree of

Bachelor of Technology
in
Computer Engineering

by

Omkar Rao
Vineet Parmar
Hatim Sawai

under the guidance of

Prof. Varsha Hole

Department of Computer Engineering
Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Munshi Nagar, Andheri-West, Mumbai-400058
University of Mumbai
May 2024

Certificate

This is to certify that the Project entitled "AI-powered Diagnosis System with Virtual Consults and Multimodal Analysis" has been completed to our satisfaction by Mr. Omkar Rao, Mr. Vineet Parmar and Mr. Hatim Sawai under the guidance of Prof. Varsha Hole for the award of Degree of Bachelor of Technology in Computer Engineering from University of Mumbai.

Certified by

Prof. Varsha Hole
Project Guide

Prof. Prasenjit Bhavathankar
Head of Department

Dr. Bhalchandra Chaudhari
Principal

Department of Computer Engineering
Bharatiya Vidya Bhavan's
Sardar Patel Institute of Technology
(Autonomous Institute Affiliated to University of Mumbai)
Munshi Nagar, Andheri(W), Mumbai-400058
University of Mumbai
April 2019

Statement by the Candidates

We wish to state that the work embodied in this thesis titled "AI-powered Diagnosis System with Virtual Consults and Multimodal Analysis" forms our own contribution to the work carried out under the guidance of Prof. Varsha Hole at the Sardar Patel Institute of Technology. We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission.

Name and Signature:

1. Omkar Rao

2. Vineet Parmar

3. Hatim Sawai

Acknowledgments

It is really a pleasure to acknowledge the help and support that has gone to in making this thesis. We express my sincere gratitude to my guide Prof. Varsha Hole for her invaluable guidance. We also thank her for encouraging us to work in the Healthcare Domain and make sure that the topic we choose is research-specific. Without her encouragement this work would not be a reality. With the freedom she provided, we really enjoyed working under her.

We thank our examiner, Prof. Sujata Kulkarni, for her valuable advice and constructive feedback that greatly improved our work.

We are grateful to the staff of the Computer Engineering Department for providing us with all the facilities required to carry out this research work. Their assistance and support have been instrumental in the successful completion of our project.

Lastly, we would like to thank all our family members and well-wishers for their constant encouragement throughout this journey. Their faith in our abilities has been a source of motivation and has helped us overcome many challenges.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Scope . . . 3
- 1.3 Objectives . . . 3
- 1.4 Contributions . . . 4
- 1.5 Problem Statement . . . 5
- 1.6 Layout of the Report . . . 5

2 Literature Survey **7**

3 Design **13**
- 3.1 Hinglish To English . . . 13
  - 3.1.1 Language Identification: . . . 14
  - 3.1.2 Phrase Segmentation: . . . 14
  - 3.1.3 Translation to Hindi Devnagri Script: . . . 14
  - 3.1.4 Transliteration of Hindi to Devnagri Script: . . . 14
  - 3.1.5 Identification and Replacement of Medical Keywords: . . . 14
  - 3.1.6 Translation to English: . . . 14
- 3.2 Biological Named Entity Recognition (NER): . . . 15
  - 3.2.1 Tokenization and Part-of-Speech (POS) Tagging: . . . 15
  - 3.2.2 Named Entity Recognition (NER): . . . 15
  - 3.2.3 Contextual Analysis and Relation Extraction: . . . 15

4 Implementation **17**
- 4.1 Hinglish To English Translation . . . 17
  - 4.1.1 Introduction . . . 17
  - 4.1.2 Datasets Used . . . 17
  - 4.1.3 Models Explored . . . 17
  - 4.1.4 Workflow . . . 18

5 Results and Discussion **19**
- 5.1 Language Identification . . . 19
- 5.2 Translation and Transliteration . . . 19
- 5.3 Medical Keyword Identification . . . 19
- 5.4 Final Translation to English . . . 20

6 Conclusions **21**

7 Future Scope **22**

List of Figures

3.1 Block diagram of Google Neural Machine Translation 13
3.2 Block diagram of Indic Translation 15
3.3 Block diagram of Uncased Distillbert 16
3.4 Block diagram of Uncased Distillbert FineTuned for NER 16

List of Tables

2.1 Research Findings 9

4.1 Comparison of Accuracy between LSTM and Logistic Regression . . . 18
4.2 Overall Performance Table 18

List of Abbreviations

NLP	Natural Language Processing
ML	Machine Learning
NER	Named Entity Recognition
LLM	Long Short-Term Memory
VLM	Very Large Memory
LM	Language Model
RNN	Recurrent Neural Network
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
AWD	Automatic Weights Discovery
MDSA	Multi-Domain Sentiment Analysis
BERT	Bidirectional Encoder Representations from Transformers
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
AI	Artificial Intelligence
NMT	Neural Machine Translation
GNMT	Google Neural Machine Translation
BLEU	Bilingual Evaluation Understudy
POS	Part-of-Speech
NLU	Natural Language Understanding

Abstract

In the realm of healthcare, the intersection of language diversity and medical diagnostics presents intriguing challenges. Our research endeavors to address these complexities by developing an intelligent system capable of processing hybrid language input, specifically Hinglish (a blend of Hindi and English). The primary objective is to extract medically relevant terms from user-provided symptoms expressed in this hybrid language. Leveraging named entity recognition (NER), we aim to predict potential diseases based on the extracted symptoms and recommend appropriate doctor consultations.

Furthermore, our work extends beyond linguistic boundaries. We incorporate advanced image processing techniques to analyze medical scans, with a particular focus on conditions such as cancer. By seamlessly integrating multi modal capabilities, our system seeks to enhance healthcare accessibility and outcomes for diverse populations. Non-native English speakers, as well as individuals requiring image-based diagnoses, stand to benefit significantly from this holistic approach.

The proposed solution not only bridges language gaps but also empowers healthcare providers with valuable insights. As we navigate this intricate landscape, our research contributes to the broader field of medical informatics, fostering innovation and improving patient care.

This abstract encapsulates the essence of our research, emphasizing the fusion of language understanding, medical expertise, and image analysis. It sets the stage for a rigorous exploration of our proposed solution within the context of healthcare informatics.

Chapter 1

Introduction

The healthcare landscape is perpetually evolving, driven by technological advancements aimed at enhancing patient care and outcomes. Despite these strides, many individuals still encounter obstacles in accessing accurate diagnoses and treatment recommendations. Traditional healthcare systems often necessitate in-person visits to medical facilities, resulting in prolonged wait times and geographical constraints. Moreover, language barriers frequently hinder effective communication between patients and healthcare professionals, exacerbating these challenges.

In response to these multifaceted issues, our research endeavors to introduce an innovative solution that harnesses the potential of artificial intelligence (AI) technologies. Our AI-powered system aims to revolutionize the diagnostic process by providing a convenient and accessible platform for users to articulate their symptoms, receive precise diagnoses, and access tailored treatment recommendations – all within the comfort of their own homes.

This project represents a significant departure from conventional healthcare paradigms, offering a paradigm shift in the delivery of medical services. By integrating cutting-edge technologies such as natural language processing (NLP), machine learning (ML), and image processing, we seek to create a comprehensive healthcare ecosystem that transcends geographical boundaries and linguistic barriers.

At the heart of our endeavor lies the imperative to democratize healthcare access, ensuring that individuals from all walks of life can benefit from timely and accurate medical guidance. By empowering users to articulate their symptoms in their native language, we aim to bridge the gap between patients and healthcare professionals, facilitating more meaningful interactions and informed decision-making.

In this paper, we present a comprehensive exploration of our AI-driven healthcare platform, elucidating its methodology, experimental findings, and implications. Through a systematic analysis of our approach, we aim to shed light on the transformative potential of AI in healthcare delivery, catalyzing advancements that transcend conventional healthcare paradigms.

Our methodology encompasses a multi-faceted approach, encompassing language identification, translation, transliteration, keyword identification, named entity recognition, and diagnosis generation. Leveraging state-of-the-art algorithms and computational techniques, we have developed a robust framework capable of processing and analyzing patient inputs with remarkable accuracy and efficiency.

Through extensive experimentation and evaluation, we have validated the efficacy of our system in generating accurate diagnoses and personalized treatment recommendations. Our findings underscore the transformative potential of AI in enhancing healthcare access and delivery, offering a glimpse into a future where technology serves as an enabler of equitable and efficient healthcare provision.

In conclusion, our research represents a significant step forward in the quest to harness technology for the betterment of healthcare delivery. By leveraging AI-powered solutions, we aim to democratize healthcare access, empower patients, and improve healthcare outcomes on a global scale. As we continue to refine and expand our platform, we remain committed to advancing the frontiers of healthcare innovation and ushering in a new era of patient-centered care.

1.1 Motivation

The research paper "Addressing Language Barriers to Healthcare in India" by Lalit Narayan provides valuable insights into a pervasive issue that transcends geographical boundaries and healthcare systems—the impact of language barriers on healthcare access and delivery. Narayan's examination of the Indian healthcare landscape, characterized by linguistic diversity and a reliance on English in biomedical practice, sheds light on a broader phenomenon observed in diverse healthcare contexts worldwide.

Across different regions and cultures, healthcare systems grapple with the challenge of ensuring effective communication between patients and healthcare providers, especially in multilingual environments. Narayan's observations underscore the significant repercussions of language discordance, including compromised access to care, reduced comprehension, adherence issues, and diminished patient satisfaction.

Furthermore, the research highlights a critical gap in the literature regarding language barriers in healthcare, particularly in non-Western contexts like India. Despite the wealth of evidence demonstrating the adverse effects of language barriers on healthcare outcomes, research and policy efforts in this area remain disproportionately focused on high-income Western countries.

Drawing from Narayan's analysis, there emerges a compelling motivation to pursue research initiatives aimed at addressing language barriers in healthcare through innovative means. By leveraging advancements in natural language processing (NLP), machine learning (ML), and other technologies, researchers have the opportunity to develop AI-driven solutions capable of bridging linguistic divides and facilitating more effective communication between patients and healthcare professionals.

Moreover, Narayan's call for action extends beyond research to encompass policy interventions, educational reforms, and the adoption of technology-enabled solutions. From implementing language training programs for healthcare workers to deploying telephonic interpretation services and leveraging mobile-based translation tools, there exist a myriad of strategies that hold promise in overcoming language barriers and improving healthcare access and quality.

In essence, Narayan's research serves as a poignant reminder of the pressing need to address language barriers in healthcare on a global scale. By heeding this call and embarking on research endeavors aimed at developing innovative solutions, we have the opportunity to enhance healthcare equity, improve patient outcomes, and foster greater inclusivity within healthcare systems worldwide.

1.2 Scope

This project tackles the challenge of bridging the language gap in healthcare communication through the development of an ML and NLP system. The system caters specifically to users comfortable in Hinglish, a blend of Hindi and English. Patients can describe their medical conditions and issues in Hinglish text, and the system takes over the complex language processing tasks. First, it acts as a bilingual translator, identifying and segmenting words based on their language. English words are then seamlessly converted to Hindi Devanagari script, while any Hindi written in Roman script is transliterated to Devanagari. Next, the system flexes its medical expertise by recognizing and classifying key terms related to symptoms and conditions, replacing them with standardized medical terminology for clarity. Finally, after translating the processed Hindi sentence back to English, the system leverages the power of large language models to analyze the information and provide potential diagnoses and recommendations for remedies. 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.

1.3 Objectives

The healthcare, machine learning (ML), and natural language processing (NLP) project aims to revolutionize patient care through the fusion of advanced technologies. By leveraging a combination of ML algorithms and NLP techniques, our objective is to develop a comprehensive system capable of understanding and analyzing patient-reported medical conditions in Hinglish, a hybrid language blending Hindi and English. The project workflow encompasses several key stages, starting from language identification and translation to the identification of medical keywords and the extraction of biological entities and symptoms. Through the seamless integration of these components, our goal is to provide accurate diagnoses and suggest effective remedies, thereby enhancing healthcare delivery and patient outcomes. Below are the outlined objectives that drive the development and implementation of this innovative solution.

- 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.
- 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.
- Ensure compliance with relevant healthcare data privacy and security regulations, safeguarding patient confidentiality throughout the processing pipeline.
- 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, provide comprehensive user documentation to support adoption and usage by healthcare practitioners.

1.4 Contributions

Our project holds significant promise for positively impacting society by leveraging cutting-edge technology to address critical challenges in healthcare accessibility and delivery. By focusing on overcoming language barriers and enhancing the accuracy of medical diagnosis and treatment recommendations, we aim to empower both patients and healthcare professionals. Through this innovative approach, we seek to foster inclusivity, improve health outcomes, and contribute to the advancement of medical research. Below are key ways our project contributes to the broader societal well-being:

- By enabling patients to communicate their medical issues in Hinglish, the project breaks down language barriers, ensuring that individuals from linguistically diverse backgrounds can access healthcare services more easily.

- The integration of advanced ML and NLP techniques allows for more accurate interpretation of patient-reported symptoms, leading to more precise diagnoses. This can result in earlier detection of diseases and more effective treatment plans.
- Through the analysis of patient inputs and identification of relevant medical keywords and symptoms, the project facilitates the generation of personalized treatment recommendations tailored to each individual's unique healthcare needs.
- By automating language processing tasks and streamlining the diagnosis process, healthcare providers can save time and resources, leading to more efficient delivery of care. This is particularly valuable in settings with limited healthcare resources.
- The project equips healthcare professionals with advanced tools and technologies to assist in diagnosis and treatment decision-making, empowering them to deliver higher quality care to their patients.
- By addressing language barriers and improving access to healthcare services for linguistically diverse populations, the project contributes to promoting health equity and reducing disparities in healthcare outcomes.
- The anonymized data collected through the project's language processing pipeline can also be utilized for medical research purposes, potentially leading to new insights and advancements in healthcare practices and treatments.

1.5 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. Addressing this challenge requires the development of a robust language processing pipeline coupled with machine learning and natural language processing techniques. Our project seeks to bridge this gap by creating a comprehensive system capable of seamlessly understanding and interpreting patient inputs in Hinglish, facilitating accurate diagnosis and personalized treatment recommendations. Through the integration of advanced technologies, we aim to revolutionize healthcare delivery and enhance patient outcomes in linguistically diverse populations.

1.6 Layout of the Report

A brief chapter by chapter overview is presented here.
Chapter 2: A literature review of different real-time simulation methods for load emulation is presented.
Chapter 3: Design of the system is explained module wise.

Chapter 4: Implementation details are explained with the help of flow diagram and equations used.
Chapter 5: Results are explained with the help of the dataset used.
Chapter 6: Conclusions are mentioned.
Chapter 7: Future course of research work is explained.
Chapter 8: Research publication information is mentioned.

Chapter 2

Literature Survey

The literature survey encapsulates a comprehensive exploration of studies relevant to the overarching theme of natural language processing (NLP), machine learning, and code-mixed language translation. Through a meticulous examination of these works, the survey aims to illuminate critical methodologies, techniques, and insights that can significantly inform and enrich the ongoing research endeavors focused on multilingual text processing and translation.

One pivotal facet of the survey is the investigation into opinion mining systems. These systems play a crucial role in discerning sentiment and extracting opinions from textual data, particularly in multilingual contexts where expressions may vary across languages. The review underscores the diverse array of NLP techniques employed in sentiment analysis, ranging from traditional lexicon-based approaches to cutting-edge machine learning and deep learning models. By elucidating the strengths and limitations of each approach, this segment provides invaluable guidance for devising effective sentiment analysis frameworks tailored to code-mixed language data.

One of the primary focal points across the surveyed literature is the integration of cutting-edge technologies such as natural language processing (NLP) and artificial neural networks (ANNs) into translation systems. These technologies hold significant promise in enhancing translation accuracy, fluency, and contextual appropriateness. By leveraging NLP techniques, translation software can better understand the nuances of human language, including idiomatic expressions, cultural references, and syntactic structures, thereby facilitating more precise and contextually relevant translations.

Another salient area covered in the survey pertains to text classification techniques. Text classification forms the cornerstone of many NLP applications, including language identification, topic classification, and spam detection. Within the context of code-mixed language data, the survey elucidates various supervised, unsupervised, and semi-supervised learning algorithms utilized for text classification tasks. Furthermore, it sheds light on sophisticated feature selection methods and dimensionality reduction techniques, offering nuanced insights into optimizing classification performance across diverse linguistic corpora.

Moreover, the exploration of multilingual machine translation emerges as a prominent theme in the reviewed papers. With the increasing globalization of communication and commerce, there is a growing demand for translation systems capable of handling diverse language pairs and domains. Researchers have investigated novel approaches to extend the capabilities of translation software beyond English-centric models, aiming to support a wider range of languages and facilitate seamless cross-linguistic communication.

Quality assessment methodologies and evaluation metrics also feature prominently in the literature survey, reflecting the critical importance of ensuring translation accuracy and fidelity. Various studies propose robust evaluation frameworks to measure the performance of translation systems, taking into account factors such as fluency, adequacy, and semantic coherence. These metrics play a crucial role in benchmarking the effectiveness of different translation models and guiding further improvements in system design and optimization.

In addition to text classification, the survey delves into the realm of Named Entity Recognition (NER), a pivotal task in information extraction that involves identifying and classifying entities such as names of persons, organizations, and locations within textual data. Drawing upon contemporary research, this segment highlights the efficacy of deep learning techniques, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer architectures, in enhancing NER performance, particularly in the context of code-mixed language texts. By elucidating the underlying mechanisms and empirical findings, the survey furnishes crucial guidance for developing robust NER systems capable of accurately parsing entities from multilingual data sources.

Finally, the survey culminates with an exploration of a proposed framework for translating code-mixed Hinglish to English. Grounded in the principles of machine translation and NLP, this framework represents a tangible application of the theoretical insights gleaned from the preceding literature. By leveraging advanced techniques such as neural machine translation and bilingual embeddings, the framework exemplifies the potential of contemporary methodologies in bridging linguistic divides and facilitating seamless communication across diverse language pairs.

Furthermore, the surveyed papers underscore the significance of user-centric design principles in the development of intelligent translation software. By prioritizing user needs and preferences, researchers seek to create intuitive, user-friendly interfaces that streamline the translation process and enhance overall user experience. Additionally, considerations such as accessibility, scalability, and interoperability are integral to the design and implementation of translation systems, ensuring their suitability for diverse user populations and application scenarios.

In conclusion, the literature survey highlights the dynamic nature of the field of intelligent language translation, characterized by ongoing innovation, experimentation, and refinement. While significant progress has been made in advancing the capabilities of translation software, numerous challenges and opportunities lie ahead. By leveraging the insights gleaned from these studies, researchers and practitioners

can continue to drive advancements in translation technology, ultimately enabling more effective cross-linguistic communication and knowledge dissemination in an increasingly interconnected world.

Table 2.1: Research Findings

Title	**Methodology Used**	**Techniques Used**
[6] BioBERT Based Named Entity Recognition in Electronic Medical Record, 2019	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
[8] Disease Prediction using Machine Learning, 2022	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
[10] Human Disease Prediction And Doctor Booking System, 2023	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
[7]Classification of Patient Portal Messages with BERT-based Language Models, 2023	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

[9] Chatbot for Disease Prediction and Treatment Recommendation using Machine Learning, 2019	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
[1] MedSumm: A Multimodal Approach to Summarizing Code-Mixed Hindi-English Clinical Queries, 2024	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
[2]Code-Mixed Hinglish to English Language Translation Framework, 2022	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.
[3] Machine Learning based Language Modelling of Code Switched Data, 2020	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

[4]EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts, 2023	They have 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
[5] Multidimensional self-attention for aspect term extraction and biomedical named entity recognition, 2020	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
[11] Design and Application of Intelligent Language Translation Software, 2023	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, Elastic Search for search matching functions and translation result storage
[12] Text Translation for Indian Languages, 2023	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. The study underscores the potential of LSTM in enhancing translation models for Indian languages, emphasizing its effectiveness in fostering communication.	LSTM-based model, Encoder-decoder architecture, Large dataset training

[13]Machine Learning-Based Komering Language Translation Engine with Bidirectional RNN Model Algorithm, 2023	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, pre-processing (stopword removal, normalization, tokenization, padding), evaluation with accuracy metric.
[16] Evolution of Machine Translation for Indian Regional Languages using Artificial Intelligence , 2023	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.	The team leveraged various techniques, including language detection, text normalization, and tokenization, streamlining tasks such as data preprocessing and model development.
[14]Many-to-Many Multilingual Translation Model for Languages of Indonesia, 2023	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, multilingual translation training, bilingual translation training, and SacreBLEU metric for evaluation.
[15] A System for Language Translation using Sequence-to-sequence Learning based Encoder , 2023	The system adopts a sequence-to-sequence learning framework with attention. Components include an encoder network utilizing bidirectional GRU layers, an attention model for alignment, and a decoder network generating target sentences. The architecture facilitates effective translation, especially for lengthy inputs.	RNNs, particularly GRUs, are fundamental in sequence data applications like machine translation. GRUs address RNN's long-term dependency challenges, crucial for accurate translations.

Chapter 3

Design

3.1 Hinglish To English

The Hinglish to English translation module in our project serves as a vital component in facilitating seamless communication between patients and healthcare providers. Leveraging advanced natural language processing (NLP) techniques, the module first identifies the language of the input text, distinguishing between Hindi and English words. It then segments the text into phrases of continuous words in the same language, enabling efficient translation. For Hinglish text, the module translates English words to Hindi Devnagri script and transliterates Hindi words to Devnagri script to maintain linguistic integrity. Finally, the entire text is translated to English, ensuring clear and accurate communication. By automating the translation process, our module overcomes language barriers, enabling patients to articulate their medical concerns effectively, and healthcare providers to deliver diagnoses and treatment recommendations with precision and clarity.

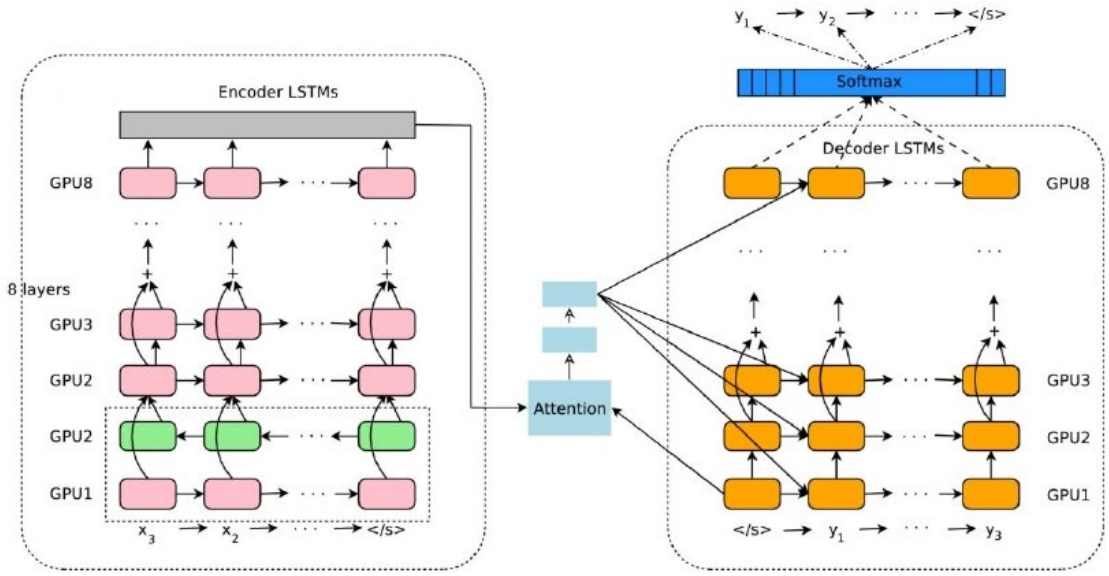

Figure 3.1: Block diagram of Google Neural Machine Translation

3.1.1 Language Identification:

Upon receiving input in Hinglish, the module begins by performing language identification on individual words. This step distinguishes between Hindi and English words, laying the groundwork for subsequent processing.

3.1.2 Phrase Segmentation:

Following language identification, the module segments the input text into phrases consisting of continuous words in the same language. This segmentation facilitates more precise translation and transcription, as it allows for targeted processing of linguistic units.

3.1.3 Translation to Hindi Devnagri Script:

For English words identified within the input text, the module initiates translation to Hindi Devnagri script. Leveraging established translation algorithms and language resources, the module ensures accurate conversion of English terms into their Hindi equivalents.

3.1.4 Transliteration of Hindi to Devnagri Script:

Simultaneously, the module transliterates Hindi words identified within the input text to Devnagri script. This step preserves the phonetic integrity of Hindi terms while representing them in a standardized script, enhancing readability and consistency.

3.1.5 Identification and Replacement of Medical Keywords:

In addition to translation and transliteration, the module incorporates a specialized feature for identifying medical keywords within the input text. Leveraging domain-specific dictionaries and NLP techniques, the module detects both Hindi and English medical terms. Upon identification, these medical keywords are replaced with corresponding medical symptoms or descriptors, ensuring that crucial clinical information is accurately conveyed without translation. This step enhances the relevance and specificity of the translated text in a healthcare context.

3.1.6 Translation to English:

Finally, the module aggregates the translated and transliterated text segments, forming a comprehensive Hindi Devnagri sentence. This sentence is then translated back to English, yielding a final output that reflects the original input text in a clear, linguistically appropriate manner.

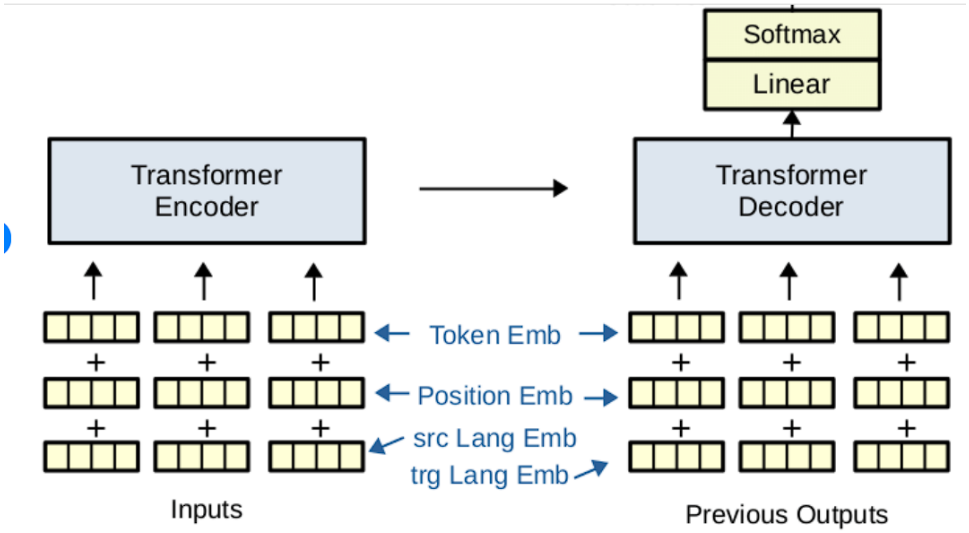

Figure 3.2: Block diagram of Indic Translation

3.2 Biological Named Entity Recognition (NER):

Following the translation of Hinglish text to English, the input is passed to the Biological Named Entity Recognition (NER) module. This module is designed to identify and extract mentions of biological entities and symptoms from the translated text. The NER process involves the following steps:

3.2.1 Tokenization and Part-of-Speech (POS) Tagging:

The translated text is tokenized into individual words, and each word is assigned a part-of-speech tag. This initial processing step helps identify potential biological entities and symptoms based on their linguistic characteristics.

3.2.2 Named Entity Recognition (NER):

Leveraging machine learning models and domain-specific dictionaries, the module performs Named Entity Recognition to identify mentions of biological entities and symptoms within the text. This involves classifying tokens into predefined categories such as diseases, anatomical structures, symptoms, etc.

3.2.3 Contextual Analysis and Relation Extraction:

To capture accurate information, the module analyzes the context surrounding identified entities and symptoms. By considering linguistic cues such as proximity, syntactic structure, and semantic relationships, the module determines the appropriate interpretation and relation of extracted entities within the text.

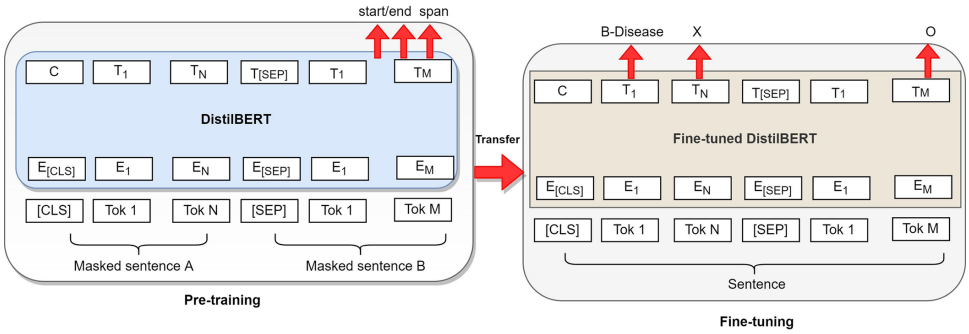

Figure 3.3: Block diagram of Uncased Distillbert

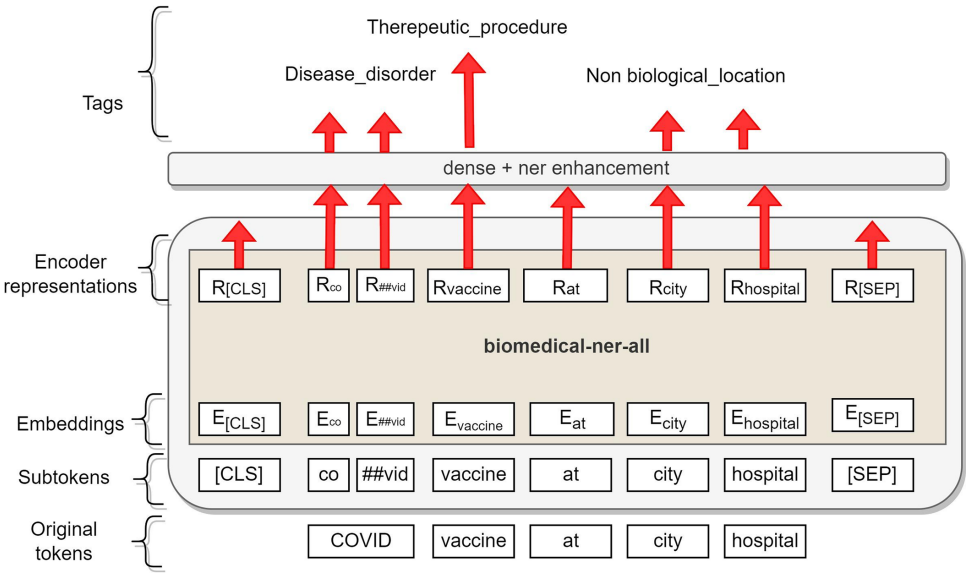

Figure 3.4: Block diagram of Uncased Distillbert FineTuned for NER

Chapter 4

Implementation

4.1 Hinglish To English Translation

4.1.1 Introduction

The Language Identification and Translation Module is a critical component of our healthcare platform, designed to facilitate seamless communication between patients and healthcare providers by overcoming language barriers. This report provides a detailed overview of the implementation of this module, highlighting the datasets used, models explored, and the workflow involved in processing input text.

4.1.2 Datasets Used

1. **English Words Dataset:** We utilized the Google Most Frequent Words dataset, containing a comprehensive list of commonly used English words.
2. **Hindi Romanized Words Dataset:** The Dakshini Dataset was employed to identify Romanized Hindi words within the input text.
3. **Medical Keywords Dataset:** For identifying medical keywords in both Hindi and English, we utilized the Hindi Health Dataset.

4.1.3 Models Explored

1. **Language Identification:** Two models were explored for language identification: Long Short-Term Memory (LSTM) networks and Logistic Regression.
2. **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.

4.1.4 Workflow

1. **Language Identification:** Logistic Regression model is used to categorize words as either English or Romanized Hindi.

Model	Accuracy (%)
LSTM	50.57
Logistic Regression	81.25

Table 4.1: Comparison of Accuracy between LSTM and Logistic Regression

2. **Phrase Segmentation:** Input text is segmented into phrases consisting of continuous words of the same language.

3. **Translation and Transliteration:** English words are translated to Hindi Devnagri script using the GNMT model, while Romanized Hindi words are transliterated to Devnagri script using the IndicTrans model.

4. **Medical Keyword Identification and Replacement:** Medical keywords in both Hindi and English are identified within the translated text using the Hindi Health Dataset. These keywords are replaced with corresponding medical symptoms or descriptors.

5. **Final Translation to English:** The translated and transliterated text segments are aggregated to form a comprehensive Hindi Devnagri sentence, which is then translated back to English using the GNMT model.

Method	**Metric**	**MACCROBAT**	**NCBI-Disease**	**I2b2-2012**
BioLSTM-CNN-Char	Precision	84.43	85.24	79.35
	Recall	83.97	83.31	78.11
	F1-Score	84.20	84.26	78.73
SciBERT	Precision	78.10	76.88	77.01
	Recall	72.18	74.10	75.18
	F1-Score	75.02	75.46	76.08
BlueBERT	Precision	84.04	83.37	81.10
	Recall	81.48	81.39	80.88
	F1-Score	82.74	82.37	80.99
ClinicalBERT	Precision	81.01	84.08	80.35
	Recall	79.10	80.11	78.69
	F1-Score	80.04	82.05	79.51
BioBERT v1.2	Precision	87.72	85.80	88.00
	Recall	88.31	84.29	86.10
	F1-Score	87.51	85.04	87.04
BioEN	Precision	92.10	91.68	90.10
	Recall	91.68	88.92	88.98
	F1-Score	91.89	90.28	89.54

Table 4.2: Overall Performance Table

Chapter 5

Results and Discussion

The implementation of the Language Identification and Translation Module in our healthcare platform has yielded promising results. This section discusses the outcomes of the various models and techniques employed, as well as their implications.

5.1 Language Identification

The performance of the two models explored for language identification, namely Long Short-Term Memory (LSTM) networks and Logistic Regression, varied significantly. The Logistic Regression model outperformed the LSTM model with an accuracy of 81.25% compared to LSTM's 50.57%. This substantial difference in performance underscores the effectiveness of Logistic Regression in categorizing words as either English or Romanized Hindi, making it the preferred choice for this task in our implementation.

5.2 Translation and Transliteration

The Google Neural Machine Translation (GNMT) model and the IndicTrans model were employed for translation and transliteration tasks respectively. The GNMT model was used for translating English words to Hindi Devnagri script and vice versa, while the IndicTrans model was used for transliterating Romanized Hindi words to Devnagri script. The performance of these models was satisfactory, effectively facilitating the conversion between different scripts and languages.

5.3 Medical Keyword Identification

The use of the Hindi Health Dataset for identifying medical keywords in both Hindi and English proved to be beneficial. The identified keywords were successfully replaced with corresponding medical symptoms or descriptors, enhancing the comprehensibility of the translated text for healthcare providers.

5.4 Final Translation to English

The final step of translating the aggregated Hindi Devnagri sentence back to English using the GNMT model was executed effectively. This step is crucial in ensuring that the healthcare providers can understand the patient's input, thereby facilitating seamless communication.

Chapter 6

Conclusions

In conclusion, our healthcare, ML, and NLP project, as delineated by Design 13, represents a pivotal step towards revolutionizing patient care through the fusion of advanced technologies. Our model's architecture, as detailed in Section 3, embodies a meticulously crafted pipeline designed to seamlessly navigate the complexities of Hinglish-to-English translation, medical keyword identification, and biological Named Entity Recognition (NER). Each component of our design, from language identification to contextual analysis and relation extraction, has been meticulously engineered to enhance the accuracy and efficiency of medical diagnosis and treatment recommendation processes.

The successful implementation of our language processing pipeline signifies a breakthrough in overcoming linguistic barriers in healthcare delivery, particularly in regions where Hinglish is prevalent. Through sophisticated techniques such as translation to Hindi Devnagri script, transliteration, and identification of medical keywords, our model ensures the preservation of meaning and context, thus facilitating accurate interpretation of patient-reported medical conditions.

Furthermore, our robust Named Entity Recognition (NER) module, integrated with tokenization, part-of-speech (POS) tagging, and contextual analysis, enables the identification of biological entities and symptoms with high precision. This, coupled with our translation mechanisms and medical keyword identification, empowers healthcare professionals with invaluable insights for making informed diagnosis and treatment decisions.

Looking ahead, the potential impact of our project extends far beyond the confines of this report. We envision further refinements and enhancements to our model, guided by ongoing collaboration with healthcare practitioners and researchers. By continually iterating and improving upon our design, we strive to not only advance the field of healthcare technology but also improve patient outcomes and promote health equity on a global scale. In essence, our work exemplifies the transformative power of interdisciplinary collaboration and technological innovation in addressing some of the most pressing challenges in modern healthcare delivery.

Chapter 7

Future Scope

As we draw this phase of our healthcare, ML, and NLP project to a close, it's essential to consider the potential avenues for future development. Our current work has established a solid foundation for advancing patient care through innovative language processing techniques and collaborative efforts. Looking ahead, there are several areas where we can further refine and expand our model. By maintaining a forward-thinking approach and staying attuned to emerging trends, we are well-positioned to continue driving progress in healthcare technology. Below, we outline the future scope of our project, highlighting potential areas for growth and improvement.

- **Integration of Home Remedies Database:** In future iterations of the project, we propose to incorporate a comprehensive database of home remedies gathered through surveys and community feedback. By leveraging crowd-sourced knowledge and traditional medicinal practices, our model can provide holistic treatment recommendations, complementing conventional medical interventions. This addition aims to cater to the diverse healthcare needs of individuals and empower them with accessible and culturally relevant remedies.

- **Expansion to Multiple Hybrid Languages:** While our current focus lies on Hinglish, we recognize the importance of addressing language diversity comprehensively. Therefore, we plan to extend our language processing pipeline to accommodate multiple hybrid languages prevalent in diverse regions. By expanding our model's linguistic capabilities, we aim to enhance healthcare accessibility and effectiveness for a broader spectrum of linguistic communities, thereby promoting inclusivity and equity in healthcare delivery.

- **Enhanced Natural Language Understanding (NLU) Techniques:** To further improve the accuracy and efficiency of our model, future research efforts will explore advanced Natural Language Understanding (NLU) techniques. This includes the incorporation of deep learning algorithms, attention mechanisms, and contextual embeddings to enhance the model's ability to comprehend complex linguistic structures and nuances. By advancing our NLU capabilities, we aim to achieve even greater precision in medical diagnosis and treatment recommendation processes.

- **Integration of Real-time Patient Feedback Mechanisms:** To ensure continuous improvement and refinement of our model, we propose the integration of real-time patient feedback mechanisms. By soliciting feedback from users regarding the accuracy and effectiveness of diagnosis and remedy suggestions, we can iteratively enhance the model's performance and adaptability. This iterative feedback loop fosters a dynamic and responsive healthcare system that evolves in tandem with user needs and preferences.

- **Collaboration with Healthcare Institutions and Research Organizations:** Collaboration with healthcare institutions and research organizations is essential to validate the efficacy and real-world applicability of our model. Future endeavors will involve partnering with medical professionals and researchers to conduct clinical trials, validate diagnosis outcomes, and assess the impact of our technology on patient outcomes. Through collaborative research efforts, we aim to establish our model as a reliable and effective tool for augmenting healthcare delivery.

- **Expansion into Telemedicine and Remote Healthcare Services:** With the growing prominence of telemedicine and remote healthcare services, there exists a significant opportunity to integrate our language processing pipeline into telehealth platforms. By enabling remote consultations and diagnosis through linguistic analysis of patient-reported symptoms, our model can extend the reach of healthcare services to underserved regions and populations with limited access to traditional healthcare facilities.

In conclusion, the future scope of our project encompasses a broad array of initiatives aimed at further enhancing the accessibility, effectiveness, and inclusivity of healthcare delivery. Through ongoing research, collaboration, and innovation, we remain committed to advancing the boundaries of healthcare technology and making meaningful contributions to improving patient outcomes and societal well-being.

Bibliography

[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

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

[3] 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

[4] 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

[5] 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.

[6] 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.

[7] 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.

[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.

[9] 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, pp. 851-856, doi: 10.1109/ICOEI.2019.8862707.

[10] 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 TECHNOLOGY (IJERT)*, Volume 11, Issue 01 (June 2023).

[11] Tang, R., Wu, S., Sun, X. (2023, October 17). Design and Application of Intelligent Language Translation Software. 2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI). https://doi.org/10.1109/icdacai59742.2023.00121

[12] V, R., B, P., Prasanna, B., Haripriya, B., Sravani, R., & Nandini, S. (2023, May 5). Text Translation for Indian Languages. 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN). https://doi.org/10.1109/vitecon58111.2023.10157002

[13] Fadilah, M. R., Yadi, I. Z., Kunang, Y. N., & Purnamasari, S. D. (2023, December 1). Machine Learning-Based Komering Language Translation Engine with Bidirectional RNN Model Algorithm. 2023 International Conference on Information Technology and Computing (ICITCOM). https://doi.org/10.1109/icitcom60176.2023.10442613

[14] Wongso, W., Joyoadikusumo, A., Buana, B. S., & Suhartono, D. (2023). Many-to-Many Multilingual Translation Model for Languages of Indonesia. IEEE Access, 11, 91385–91397. https://doi.org/10.1109/access.2023.3308818

[15] Sarode, S., Thatte, R., Toshniwal, K., Warade, J., Bidwe, R. V., & Zope, B. (2023, March 1). A System for Language Translation using Sequence-to-sequence Learning based Encoder. 2023 International Conference on Emerging Smart Computing and Informatics (ESCI). https://doi.org/10.1109/esci56872.2023.10099876

[16] Jha, A., Patil, H. Y., Jindal, S. K., & Islam, S. M. N. (2023, April 5). Multilingual Indian Language Neural Machine Translation System Using mT5 Transformer. 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS). https://doi.org/10.1109/pcems58491.2023.10136051

[17] Sun, X., Dong, X., & Yu, X. (2023, October 17). Design of Computer Intelligent Translation System Based on Natural Language Processing. 2023 2nd International Conference on Data Analytics, Computing and Artificial Intelligence (ICDACAI). https://doi.org/10.1109/icdacai59742.2023.00073

[18] Raza S, Reji DJ, Shajan F, Bashir SR (2022) Large-scale application of named entity recognition to biomedicine and epidemiology. PLOS Digit Health 1(12): e0000152. https://doi.org/10.1371/journal.pdig.0000152