Vineet Parmar

Major_Project_Blackbook__Copy_ Fin

Vineet

Document Details

Submission ID

trn:oid:::3618:105111339

Submission Date

Jul 20, 2025, 12:56 PM GMT+5:30

Download Date

Jul 20, 2025, 7:01 PM GMT+5:30

File Name

Major_Project_Blackbook__Copy_ Fin.pdf

File Size

820.1 KB

31 Pages

7,529 Words

46,025 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

60 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

2 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4%	Internet sources
4%	Publications
6%	Submitted works (Student Papers)

Match Groups

60 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

2 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4%	Internet sources
4%	Publications
6%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	arxiv.org	<1%
2	Internet	aiforsocialgood.ca	<1%
3	Publication	Vaibhav Kumar, Shubham Pasari, Vallabh Pravin Patil, Sumedha Seniaray. "Machi...	<1%
4	Internet	www.frontiersin.org	<1%
5	Internet	www.inderscience.com	<1%
6	Student papers	Australian National University on 2024-11-05	<1%
7	Student papers	University of Southern Queensland on 2023-10-20	<1%
8	Internet	ir.uitm.edu.my	<1%
9	Student papers	Addis Ababa University on 2025-05-28	<1%
10	Internet	thesai.org	<1%

11	Internet	www.teses.usp.br	<1%
12	Student papers	Liverpool John Moores University on 2024-06-18	<1%
13	Student papers	Ohio University on 2005-10-02	<1%
14	Student papers	University College Dublin (UCD) on 2024-06-27	<1%
15	Internet	knowledge.uchicago.edu	<1%
16	Student papers	Middlesex University on 2025-02-28	<1%
17	Student papers	University of Newcastle upon Tyne on 2019-05-30	<1%
18	Student papers	Wayne State University on 2024-10-29	<1%
19	Student papers	Capella University on 2024-02-23	<1%
20	Student papers	Liverpool John Moores University on 2020-08-10	<1%
21	Student papers	University of Colorado, Denver on 2024-11-29	<1%
22	Internet	www.ijraset.com	<1%
23	Internet	www.mdpi.com	<1%
24	Publication	Shaina Raza, Deepak John Reji, Femi Shajan, Syed Raza Bashir. "Large Scale Applic...	<1%

25	Student papers	University of Durham on 2023-05-18	<1%
26	Internet	provincia.rc.it	<1%
27	Internet	stax.strath.ac.uk	<1%
28	Internet	vdoc.pub	<1%
29	Internet	www.cs.uic.edu	<1%
30	Publication	"Advances in Intelligent Computing Techniques and Applications", Springer Scien...	<1%
31	Publication	"Text Mining Approaches for Biomedical Data", Springer Science and Business Me...	<1%
32	Publication	Abouzar Qorbani, Reza Ramezani, Ahmad Baraani, Arefeh Kazemi. "Multilingual n...	<1%
33	Publication	Anitha S. Pillai, Roberto Tedesco. "Machine Learning and Deep Learning in Natur...	<1%
34	Publication	D. Lakshmi, Ravi Shekhar Tiwari, Rajesh Kumar Dhanaraj, Seifedine Kadry. "Explai...	<1%
35	Publication	H.L. Gururaj, Francesco Flammini, S. Srividhya, M.L. Chayadevi, Sheba Selvam. "Co...	<1%
36	Student papers	Heriot-Watt University on 2024-08-14	<1%
37	Student papers	King's College on 2023-04-05	<1%
38	Student papers	Obudai Egyetem on 2025-05-16	<1%

39	Publication	Rishabha Malviya, Selcan Karakuş, Mukesh Roy. "Embedded Systems for Biomedi...	<1%
40	Student papers	University College Dublin (UCD) on 2024-06-25	<1%
41	Student papers	University of Nottingham on 2024-05-02	<1%
42	Student papers	University of Queensland on 2024-06-03	<1%
43	Internet	assets-eu.researchsquare.com	<1%
44	Internet	eprints.nottingham.ac.uk	<1%
45	Internet	huggingface.co	<1%
46	Internet	mzjournal.com	<1%
47	Internet	ns2.thinkmind.org	<1%
48	Internet	oxfordjournals.org	<1%
49	Internet	pmc.ncbi.nlm.nih.gov	<1%
50	Internet	scholar.archive.org	<1%
51	Publication	Abhishek Narayanan, Abijna Rao, Abhishek Prasad, Bhaskarjyoti Das. "Character ...	<1%
52	Publication	Sepp Hochreiter, Jürgen Schmidhuber. "Long Short-Term Memory", Neural Comp...	<1%

53 Publication

Xinyu Song, Ao Feng, Weikuan Wang, Zhengjie Gao. "Multidimensional Self-Attent... <1%

8

Contents

List of Figures

List of Tables

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 healthcare realm, 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 Natural Language Processing(NLP), we aim to use the extracted terms to suggest a differential diagnosis for potential diseases.

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 and medical expertise. 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 leveraging modern and evolving techniques and technologies, such as natural language processing (NLP) and machine learning (ML), we aim to provide a comprehensive healthcare ecosystem that breaks linguistic barriers.

At the heart of our endeavor lies the imperative to democratize access to healthcare, 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 reduce communication discrepancies between patients and healthcare professionals, facilitating more meaningful interactions and informed decision-making.

Through this paper, we present a detailed breakdown of our AI-driven healthcare platform, elucidating its methodology, experimental findings, and implications. Through a systematic analysis of our approach, we aim to highlight the massive potential of AI in healthcare delivery, catalyzing advancements that revolutionize 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 our experimentation and evaluation, we have validated the efficacy of our system in generating accurate diagnoses from language-mixed patient descriptions. The outcomes of our research further strengthen the use case 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.

Ultimately, this research offers a substantial contribution to leveraging technology for more effective healthcare provision. 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 using modern propositions and tools 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.

29

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 bring forth healthcare equity, enhance patient experience, 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 proposed healthcare project aims to revolutionize patient care through the fusion of advanced technologies. By leveraging hybrid ML algorithms and NLP techniques, our aim is to create a comprehensive system which can analyze 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.
- 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 process patient data and ensure timely responses.

- To comply with 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, such as Hinglish, which combines Hindi and English. These systems are either too generic, lacking medical context, or too rigid, failing to accommodate regional linguistic variations and informal expressions. This disconnect can result in misinterpretations, misdiagnoses, and delays in providing care, particularly in under-resourced or rural areas.

Moreover, the absence of a standardized framework to handle such linguistic diversity in healthcare communication adds to the complexity. Most AI models are trained predominantly on English-language datasets, rendering them less effective for multilingual or code-switched inputs common in real-world scenarios.

To address these limitations, our project aims to develop a comprehensive AI-powered language processing pipeline tailored specifically for Hinglish text. This system will perform tasks such as language detection, transliteration, translation, symptom extraction, and medical entity recognition. Ultimately, it will provide users with preliminary diagnostic insights and treatment suggestions, thereby supporting better healthcare access and communication in linguistically diverse populations.

1.6 Layout of the Report

A brief chapter-by-chapter overview of the report is presented below:

- **Chapter 2: Literature Review** – This chapter provides details about existing resources and research on technologies used to solve problems in multilingual healthcare communication, as well as the possible use of natural language processing, and machine learning in healthcare sector.
- **Chapter 3: Technical Support** – This section provides the essential mathematical foundations for the key algorithms employed in our Hinglish to English medical translation system. The theoretical underpinnings that support our implementation choices and evaluation methodologies.

- **Chapter 4: System Design** – The architecture of the proposed AI-driven healthcare platform is detailed module-wise, including the design of the language processing pipeline, keyword extraction components, and integration with language models.

- **Chapter 5: Implementation** – Implementation specifics are explained, including the techniques, algorithms, tools, and flow diagrams used in building the system.

- **Chapter 6: Results and Evaluation** – This chapter presents the experimental setup, dataset details, performance metrics, and evaluation results that demonstrate the system's effectiveness in processing Hinglish inputs and generating accurate outputs.

- **Chapter 7: Conclusions** – This chapter presents key findings about our research with a reflection on how the system contributes to addressing language barriers in healthcare.

- **Chapter 8: Future Work** – Directions for future research and potential system enhancements are outlined, including scalability improvements, multilingual support expansion, and integration with real-world healthcare platforms.

- **Chapter 9: Research Publications** – Any research outputs, conference papers, or journal publications derived from this project are listed and summarized.

35

Chapter 2

Literature Survey

A. Ghosh *et al.* [1] explore the summarization of code-mixed medical text, utilizing
45 the MM-CQs dataset which integrates Hindi-English medical queries with visual
1 information. Their proposed MedSumm framework employs large language models (LLMs) and visual language models (VLMs) to achieve effective summarization.

Ishali Jadhav *et al.* [2] present a language modeling approach for classifying
3 Hinglish text. They developed a Universal Language Model by fine-tuning an AWD-LSTM architecture on a Hindi-English code-switched corpus gathered from various blogging platforms. Their work demonstrates the efficacy of deep learning architectures in processing code-mixed language data for disease prediction.

53 Vaibhav Kumar *et al.* [3] introduce a supervised learning methodology for named entity recognition (NER) in specific domains. Their model incorporates a multi-
50 dimensional self-attention (MDSA) network and a CNN-based architecture. This
30 work underscores the significance of advanced machine learning methods for improving the precision of disease prediction and facilitating patient-doctor communication.

Y. Li and Qiao *et al.* [4] describe a cascaded approach for machine translation of bilingual text, specifically from Hinglish to English. Their study leverages Python libraries like NLTK and SpaCy, illustrating how BERT-based models can enhance the classification and translation of patient messages within healthcare contexts.

X. Song *et al.* [5] discusses the critical importance of ensuring translation accuracy and fidelity through quality assessment methodologies and evaluation metrics. Various studies propose robust evaluation frameworks to measure the performance
12 of translation systems, considering factors such as fluency, adequacy, and seman-
12 tic coherence, which are essential for benchmarking translation models and guiding improvements in system design.

16 X. Yu *et al.* [6] highlights the efficacy of techniques like Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and transformer architectures, in enhancing named entity recognition (NER) performance, particularly in code-mixed language texts. This research provides crucial guidance for developing robust NER systems capable of accurately parsing entities from multilingual data sources.

Y. Ren *et al.* [7] proposes a framework for translating code-mixed Hinglish to English which is grounded in machine translation and NLP principles, representing a practical application of theoretical insights from the literature. By leveraging advanced techniques such as neural machine translation and bilingual embeddings, this framework exemplifies the potential of contemporary methodologies in bridging linguistic divides and facilitating seamless communication across diverse language pairs.

N. Kosarkar *et al.* [8] explores the role of opinion mining systems in discerning sentiment and extracting opinions from multilingual textual data. The review underscores the diverse array of NLP techniques employed in sentiment analysis, providing guidance for developing effective frameworks tailored to code-mixed language data.

R. B. Mathew *et al.* [9] investigates the application of machine learning techniques for language modeling of code-switched data. They explore various algorithms and methodologies to enhance the understanding and processing of mixed-language inputs, which is crucial for developing effective NLP applications in multilingual contexts.

Joel Roy *et al.* [10] presents a comprehensive analysis of sentiment analysis techniques applied to code-mixed language data. They evaluate different approaches, including lexicon-based and machine learning methods, to determine their effectiveness in capturing sentiment nuances in multilingual texts, thereby contributing to the field of opinion mining.

Tang, R. *et al.* [11] focuses on the integration of advanced neural network architectures for improving translation systems. They discuss the potential of using transformer models and attention mechanisms to enhance the fluency and contextual accuracy of translations, particularly in code-mixed language scenarios.

51 Rajesh V. *et al.* [12] examines the challenges and solutions in named entity recognition for code-mixed languages. They highlight the limitations of traditional NER systems and propose novel approaches that leverage deep learning techniques to improve entity extraction from multilingual datasets.

Fadilah, M. R. *et al.* [13] explores user-centric design principles in the development of intelligent translation software. They emphasize the importance of creating intuitive interfaces that cater to user needs, enhancing the overall user experience in multilingual applications.

31 Wongso, W. *et al.* [14] discusses the role of evaluation metrics in assessing the performance of translation models. They propose a set of comprehensive metrics that account for various aspects of translation quality, including semantic coherence and contextual relevance, which are vital for benchmarking and improving translation systems.

Sarode, S. *et al.* [15] presents a proposed framework for enhancing healthcare communication through NLP techniques. They outline future directions for integrating diverse linguistic capabilities into healthcare applications, aiming to improve accessibility and effectiveness for a broader range of linguistic communities.

Chapter 3

Technical Foundations

This section provides the essential mathematical foundations for the key algorithms employed in our Hinglish to English medical translation system. We present the theoretical underpinnings that support our implementation choices and evaluation methodologies.

3.1 Logistic Regression for Language Identification

36 Our language identification module employs logistic regression to distinguish between English and Romanized Hindi words. The model estimates the probability that a given word belongs to English using:

10
$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_nx_n)}} \tag{3.1}$$

43 where $x = [x_1, x_2, \ldots, x_n]$ represents the feature vector extracted from the input word, and $\beta = [\beta_0, \beta_1, \ldots, \beta_n]$ are the learned model parameters.[19]

3.2 BERT-based Disease Classification

Our disease classification system uses fine-tuned BERT to generate probability distributions over disease classes. The softmax function computes probabilities as:

18 15
$$P(y_i|x) = \frac{e^{W_i \cdot h + b_i}}{\sum_{j=1}^{C} e^{W_j \cdot h + b_j}} \tag{3.2}$$

23 where h is the contextualized representation from BERT's final layer, W_i and b_i are the weight matrix and bias for class i, and C is the total number of disease classes.[20]

The model is optimized using cross-entropy loss:

6
$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}) \tag{3.3}$$

This formulation allows the model to express uncertainty across multiple potential diagnoses, which is crucial for medical differential diagnosis.[21]

3.3 LSTM Architecture

For comparison purposes, we implemented LSTM networks for language identification. The LSTM cell state update follows:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{3.4}$$

where f_t, i_t, and $\tilde{C}_t$ represent the forget gate, input gate, and candidate values respectively.[22]

3.4 Evaluation Metrics

Standard Accuracy measures the proportion of correct predictions:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{3.5}$$

Top-K Accuracy evaluates whether the correct diagnosis appears within the top K predictions:

$$\text{Top-}k\text{ Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \nVdash \left\{ y_i \in \hat{Y}_i^{(k)} \right\} \tag{3.6}$$

where $\hat{Y}_i^{(k)}$ represents the set of top-k predicted disease classes for sample i, and $\nVdash\{\cdot\}$ is the indicator function. This metric is particularly relevant for medical applications where differential diagnosis considers multiple potential conditions.

The mathematical formulations presented here provide the theoretical foundation for our experimental comparisons and system evaluation, supporting our findings that simpler models can be more effective for specific tasks within our healthcare translation model.[23]

37

Chapter 4

Design

In this section, we detail the methodology used for the diagnosis of a disease from Hinglish text entered by the patient. The proposed framework involves two main stages - Translation of Hinglish text to English and medical diagnosis from translated text. Fig. 4.1 illustrates the system design consisting of a translation module, followed by a fine-tuned BERT model for diagnosis. Each module is discussed in detail in the upcoming sections. 33

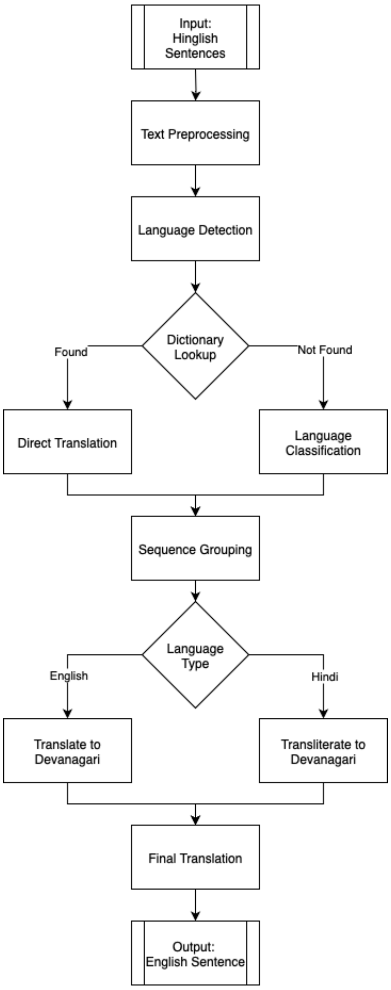


Figure 4.1: System Diagram of the proposed model

Algorithm 1 Hinglish to English Translation

```
Require: A list of Hinglish sentences (code-mixed Hindi-English)
Ensure: A list of fully translated English sentences
  Initialize empty list of translated sentences
  for each sentence in input sentences do
      Convert sentence to lowercase
      Split sentence into individual words
      Initialize empty list of processed words
      for each word in the sentence do
          if word exists in Hindi dictionary then
              Add dictionary translation with English language tag
          else
              Add word with identified language tag using language classifier
          end if
      end for
      Initialize empty output string
      Set start position to zero
      while start position is less than total number of words do
          Get current language from word at start position
          Set end position to start position plus one
          while end position is within bounds and current word has same language do
              Increment end position
          end while
          Extract sequence of words from start to end position
          if current language is English then
              Translate current sequence to Hindi
          else
              Transliterate current sequence to Devanagari script
          end if
          Append translated sequence to output string
          Set start position to end position
      end while
      Translate final output from Hindi to English
      Add translated sentence to result list
  end for
  return list of translated sentences
```

4.1 Translation from Hinglish to English

The Hinglish to English translation module in our project is vital in enabling effective communication between patients and healthcare professionals. 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 Devanagari script and transliterates Hindi words to Devanagari script to maintain linguistic integrity. Additionally, if we find common Hindi medical keywords and descriptions, we directly parse them to the corresponding English medical keyword to ensure that no important keywords are lost during translation. Finally, the entire text is translated into 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.

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

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.

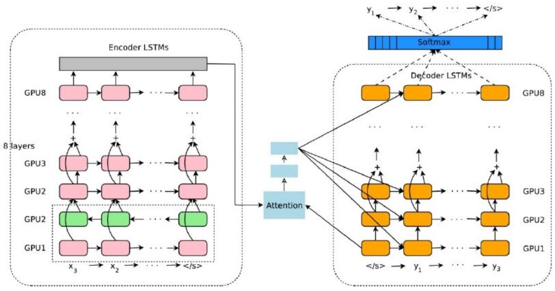


Figure 4.2: Block diagram of Google Neural Machine Translation[24]

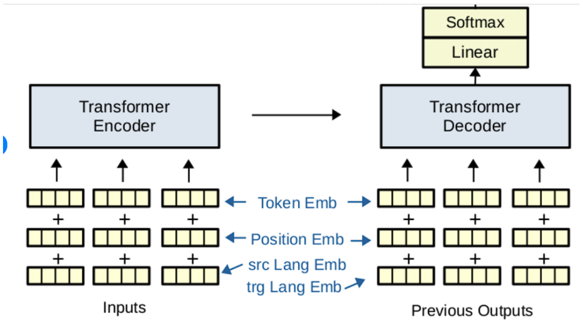


Figure 4.3: Block diagram of Indic Translation[25]

4. Medical Keyword Identification and Replacement: Medical keywords in both Hindi and English are identified within the translated text using keywords obtained

by surveying general medical practitioners for common descriptions of symptoms by patients. These keywords are replaced with corresponding medical symptoms or descriptors as provided by medical practitioners.

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.

4.2 Dataset

To facilitate Hinglish-to-English translation in medical contexts, we created a specialized dataset mapping common Hinglish medical terms to their English equivalents. A sample is shown in Tab. 4.1. This dataset includes 53 entries focusing on frequently used symptoms and body parts. For instance, ”sar” maps to ”head,” ”pet” to ”stomach,” and ”bukhar” to ”fever.” The dataset was developed with the assistance of a medical professional to ensure accuracy and relevance. It prioritizes high-frequency medical terms to optimize translation effectiveness. Each entry is carefully curated to retain essential medical terminology while enabling contextual understanding. The dataset serves as a benchmark for evaluating our translation model. By limiting the scope to common terms, we ensure a focused and practical approach to medical communication. This resource aids in bridging language barriers between patients and healthcare providers. Our dataset plays a crucial role in testing and refining Hinglish medical translation solutions.

Category	**Hindi Word**	**English Translation**
Anatomical Terms	pet	stomach
Anatomical Terms	sir	head
Anatomical Terms	aankh	eye
Common Symptoms	bhukhar	fever
Common Symptoms	khansi	cough
Common Symptoms	sardard	headache

Table 4.1: Dataset Description

4.3 Disease Identification

To enable accurate differential diagnosis based on patient-provided textual descriptions, a fine-tuned BERT-base-cased model has been integrated with the Hugging Face textclassification pipeline. The methodology consists of the following key steps:

Data Preparation: The input text, generated by the Hinglish-to-English translation module, is preprocessed for diagnosis. This involves:

Tokenization using the BERT-base-cased tokenizer to ensure compatibility with the pipeline. Padding or truncating sequences to a fixed length for batch processing. Label assignment based on annotated medical datasets containing symptom descriptions and corresponding diseases.

Model Fine-Tuning: The BERT-base-based model is fine-tuned on the prepared dataset. A classification head is added to the BERT model, enabling multi-class disease prediction. Training is performed with the cross-entropy loss function and the Adam optimizer for learning rate management.

Pipeline Integration: The fine-tuned model is deployed using the Hugging Face text-classification pipeline, which streamlines the classification process by abstracting tokenization, encoding, and inference. This pipeline simplifies integration into the broader diagnostic system.

Differential Diagnosis Generation: For each input text, the pipeline outputs probability scores for all potential diseases. The top-k diseases with the highest probability scores are selected as the differential diagnoses. By leveraging the BERT-base-cased model's capabilities for nuanced language understanding, the module achieves precise and efficient disease diagnosis, supporting accurate and timely medical interventions. Some examples of the above data are shown in Tab. 4.2

Hinglish Prompt	**Translated Text**	**Diagnosis**
Bathroom karte waqt jalan hoti hai aur bar bar bathroom jaana padta hai	After bathroom it hurts and sometimes there is frequent bathroom	urinary tract infection
Naak se pani beh raha hai aur chhikh aati rehti hai	Water is flowing from my nose and I keep sneezing	allergy

Table 4.2: Sample Input Output for k=1

Chapter 5

Implementation

5.1 Hinglish To English Translation

5.1.1 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.

19

47

5.1.2 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.

5.1.3 Models Explored

1. **Language Identification:** Two models were explored for language identification: Long Short-Term Memory (LSTM) networks and Logistic Regression. 5

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.

5.1.4 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 5.1: 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.

24

1

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 5.2: Overall Performance Table

44

Chapter 6

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.

6.1 Language Identification

5 For language identification, the performance of the two investigated models, Long Short-Term Memory (LSTM) networks and Logistic Regression, differed considerably. 1 As shown in Table 6.1, the Logistic Regression model achieved an accuracy of 88.84%, significantly outperforming the LSTM model's 50.57%. This notable difference highlights Logistic Regression's efficacy in classifying words as either English or Romanized Hindi, leading to its selection as the preferred method for this implementation task.

Model	Accuracy (%)
LSTM	50.57
Logistic Regression	88.84

Table 6.1: Comparison of Accuracy between LSTM and Logistic Regression

6.2 Translation and Transliteration

32 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.

6.3 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.

6.4 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.

6.5 Disease Diagnosis

The fine-tuned BERT model was trained and the quantitative results are shown in Fig. 6.1 and Fig. 6.2. Tab. 6.2 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.

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

Table 6.2: Comparison between old and new approach

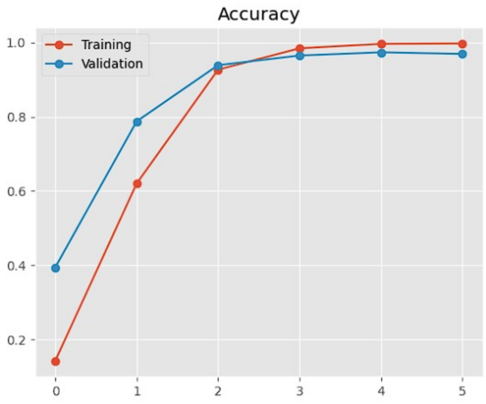


Figure 6.1: Accuracy of fine-tuned BERT Model

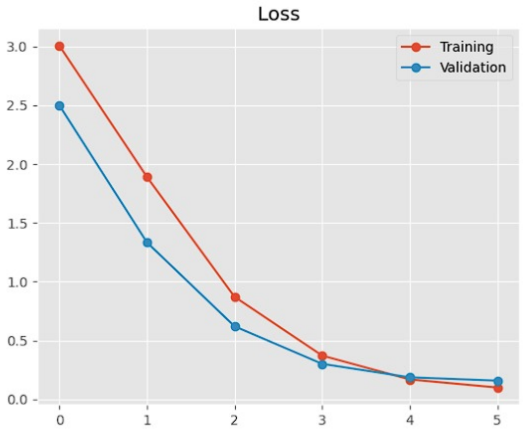


Figure 6.2: Loss of fine-tuned BERT Model

Chapter 7

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. Each component of our design, from language identification to contextual analysis and relation extraction, has been meticulously engineered to improve the efficiency and accuracy 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 Devanagri 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.

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 8

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 serve the various healthcare requirements 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 efficiency and accuracy 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 allow the model to comprehend complex structural linguistics 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 Research Organizations and Healthcare Institutions:** Collaboration with research organizations and healthcare institutions 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.

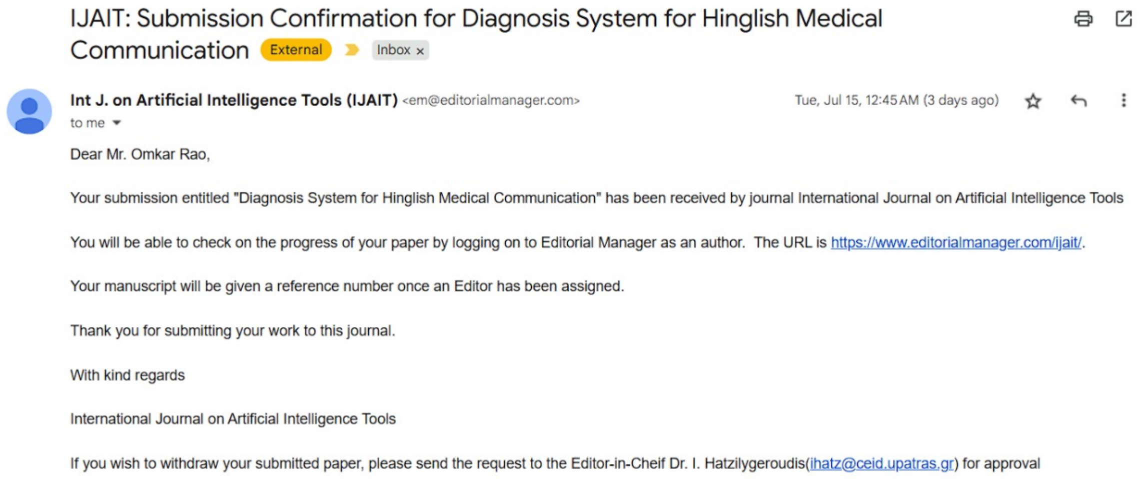
- **Expansion into Telemedicine and Remote Healthcare Services:** With the growing prominence of telemedicine and remote healthcare services, there exists a significant opportunity to integrate our language processing pipeline into telehealth platforms. By enabling remote consultations and diagnosis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

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.

Chapter 9

Research Publication

The paper titled ”Diagnosis System for Hinglish Medical Communication” has been submitted to the International Journal on Artificial Intelligence under submission ID: IJAIT-D-25-00121.



IJAIT: Submission Confirmation for Diagnosis System for Hinglish Medical Communication

Int J. on Artificial Intelligence Tools (IJAIT) <em@editorialmanager.com>

Tue, Jul 15, 12:45 AM (3 days ago)

to me

Dear Mr. Omkar Rao,

Your submission entitled "Diagnosis System for Hinglish Medical Communication" has been received by journal International Journal on Artificial Intelligence Tools

You will be able to check on the progress of your paper by logging on to Editorial Manager as an author. The URL is https://www.editorialmanager.com/ijait/.

Your manuscript will be given a reference number once an Editor has been assigned.

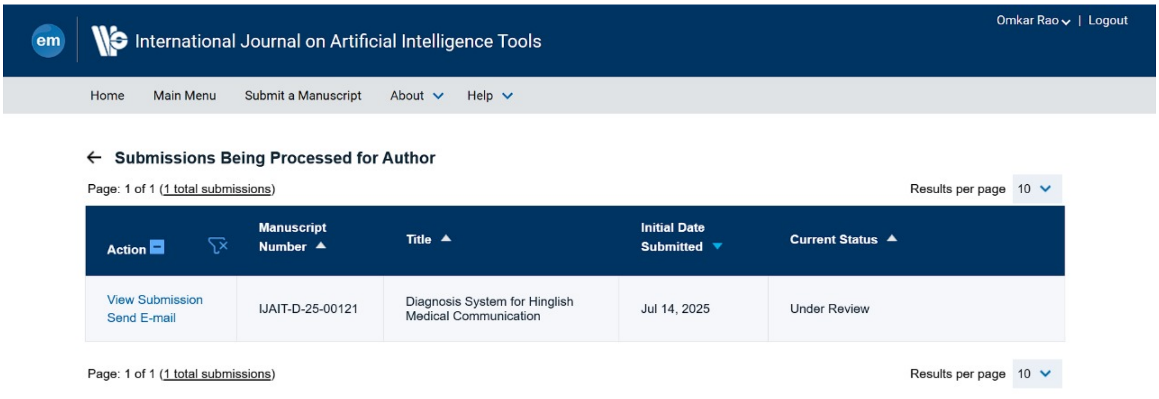
Thank you for submitting your work to this journal.

With kind regards

International Journal on Artificial Intelligence Tools

If you wish to withdraw your submitted paper, please send the request to the Editor-in-Cheif Dr. I. Hatzilygeroudis(ihatz@ceid.upatras.gr) for approval

Figure 9.1: Submission Confirmation of the Journal Paper



Submissions Being Processed for Author

Page: 1 of 1 (1 total submissions)

Results per page 10

Action	Manuscript Number	Title	Initial Date Submitted	Current Status
View Submission Send E-mail	IJAIT-D-25-00121	Diagnosis System for Hinglish Medical Communication	Jul 14, 2025	Under Review

Page: 1 of 1 (1 total submissions)

Results per page 10

Figure 9.2: Current Status of the Journal Paper Publication

Bibliography

[1] A. Ghosh, A. Acharya, P. Jha, S. Saha, A. Gaudgaul, R. Majumdar, A. Chadha, R. Jain, S. Sinha, and S. Agarwal, “MedSUMM: A multimodal approach to summarizing Code-Mixed Hindi-English clinical queries” in *Lecture Notes in Computer Science*, 2024, pp. 106–120. [Online]. Available: https://doi.org/10.1007/978-3-031-56069-9_8

[2] I. Jadhav, A. Kanade, V. Waghmare, S. S. Chandok, and A. Jarali, “Code-Mixed Hinglish to English Language Translation framework” in *Proc. 2022 Int. Conf. Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022, pp. 684–688. doi: 10.1109/icscds53736.2022.9760834.

[3] V. Kumar, S. Pasari, V. P. Patil, and S. Seniaray, “Machine Learning based Language Modelling of Code Switched Data” in *Proc. 2020 Int. Conf. Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 552–557. doi: 10.1109/icesc48915.2020.9155695.

[4] Y. Li and Qiao, “EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts” Jun. 2023. [Online]. Available: 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” *Math. Probl. Eng.*, vol. 2020, pp. 1–6, 2020. doi: 10.1155/2020/8604513.

[6] X. Yu, W. Hu, S. Lu, X. Sun, and Z. Yuan, “BioBERT Based Named Entity Recognition in Electronic Medical Record” in *Proc. 2019 10th Int. Conf. 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” in *Proc. 2022 IEEE 10th Int. Conf. Healthcare Informatics (ICHI)*, 2023, pp. 176–182. doi: 10.1109/ichi57859.2023.00033.

[8] N. Kosarkar et al., “Disease Prediction using Machine Learning” in *Proc. 2022 10th Int. Conf. Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)*, 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” in *Proc. 2019 3rd Int. Conf. Trends in Electronics and Informatics (ICOEI)*, 2019, pp. 851–856. doi: 10.1109/icoei.2019.8862707.

[10] J. Roy, R. Koshy, R. Roy, and A. Zachariah, “Human Disease Prediction and Doctor Booking System” *Int. J. Eng. Res. Technol. (IJERT)*, vol. 11, no. 1, pp. 1–6, Jun. 2023.

[11] R. Tang, S. Wu, and X. Sun, “Design and application of intelligent language translation software” in *Proc. Int. Conf. Data Analytics, Computing and Artificial Intelligence*, 2023, pp. 608–612. doi: 10.1109/icdacai59742.2023.00121.

[12] R. V, P. B, B. Prasanna, B. Haripriya, R. Sravani, and S. Nandini, “Text Translation for Indian Languages” in *Proc. Int. Conf. Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 2023, pp. 1–5. doi: 10.1109/vitecon58111.2023.10157002.

[13] M. R. Fadilah, I. Z. Yadi, Y. N. Kunang, and S. D. Purnamasari, “Machine Learning-Based Komering Language Translation Engine with bidirectional RNN model algorithm” in *Proc. 2023 Int. Conf. Information Technology and Computing (ICITCOM)*, 2023, pp. 62–66. doi: 10.1109/icitcom60176.2023.10442613.

[14] W. Wongso, A. Joyoadikusumo, B. S. Buana, and D. Suhartono, “Many-to-Many multilingual translation model for languages of Indonesia” *IEEE Access*, vol. 11, pp. 91385–91397, 2023. doi: 10.1109/access.2023.3308818.

[15] S. Sarode et al., “A System for Language Translation using Sequence-to-sequence Learning based Encoder” in *Proc. 2021 Int. Conf. Emerging Smart Computing and Informatics (ESCI)*, 2023, pp. 1–5. doi: 10.1109/esci56872.2023.10099876.

[16] A. Jha, H. Y. Patil, S. K. Jindal, and S. M. N. Islam, “Multilingual Indian Language Neural Machine Translation System Using mT5 Transformer” in *Proc. 2023 2nd Int. Conf. Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, Apr. 2023. doi: 10.1109/pcems58491.2023.10136051.

[17] X. Sun, X. Dong, and X. Yu, “Design of computer intelligent translation system based on natural language processing” in *Proc. 2023 2nd Int. Conf. Data Analytics, Computing and Artificial Intelligence (ICDACAI)*, 2023, pp. 352–356. doi: 10.1109/icdacai59742.2023.00073.

[18] S. Raza, D. J. Reji, F. Shajan, and S. R. Bashir, “Large-scale application of named entity recognition to biomedicine and epidemiology” *PLOS Digit. Health*, vol. 1, no. 12, p. e0000152, 2022. doi: 10.1371/journal.pdig.0000152.

[19] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*, 2nd ed. New York: John Wiley & Sons, 2000. Available: https://doi.org/10.1002/0471722146

[20] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN,

USA, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/N19-1423

[21] A. Mao, M. Mohri, and Y. Zhong, "Cross-Entropy Loss Functions: Theoretical Analysis and Applications" *arXiv preprint arXiv:2304.07288*, 2023. [Online]. Available: https://arxiv.org/abs/2304.07288

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory" *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: https://doi.org/10.1016/j.ipm.2009.03.002

[24] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation" *arXiv (Cornell University)*, 2016. doi: 10.48550/arxiv.1609.08144.

[25] Parida, Shantipriya & Panda, Subhadarshi & Kotwal, Ketan & Dash, Amulya & Dash, Satya & Sharma, Yashvardhan & Motlicek, Petr & Bojar, Ondřej. (2021). "NLPHut's Participation at WAT2021." 146-154.10.18653/v1/2021.wat-1.16.