Diagnosis System for Hinglish Medical Communication

Omkar Rao

Department of Computer Engineering,
Sardar Patel Institute of Technology, Andheri West
Mumbai, Maharashtra - 400058, India
omkar.rao@spit.ac.in

Vineet Parmar

Department of Computer Engineering,
Sardar Patel Institute of Technology, Andheri West
Mumbai, Maharashtra - 400058, India
vineet.parmar@spit.ac.in

Hatim Sawai

Department of Computer Engineering,
Sardar Patel Institute of Technology, Andheri West
Mumbai, Maharashtra - 400058, India
hatim.sawai@spit.ac.in

Prof. Varsha Hole

Department of Computer Science & Engineering,
Sardar Patel Institute of Technology, Andheri West
Mumbai, Maharashtra - 400058, India
varsha_hole@spit.ac.in

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 named entity recognition (NER), we aim to predict potential diseases based on the extracted symptoms and recommend appropriate doctor consultations.

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.

Keywords: Disease Prediction; Healthcare Technology;Hybrid Language Processing; Machine Learning (ML); Natural Language Processing (NLP); Communication Barriers in Healthcare

1. Introduction

Healthcare is rapidly evolving, with new technologies offering the promise of better patient care and improved outcomes. However, obtaining a timely and accurate diagnosis remains a challenge for many people, especially for those in remote areas or facing language barriers when trying to explain their symptoms. Traditional healthcare often depends on face-to-face visits, which can be difficult to arrange and may involve long waiting times. For patients who do not speak the same language as their healthcare providers, these challenges are compounded by communication gaps that can lead to misunderstandings or delayed treatment.

In this study, we introduce an AI-powered language translation system designed to make the diagnostic process more accessible and inclusive. Our goal was to help patients express their symptoms in everyday language regardless of whether language may be—and receive reliable diagnostic suggestions and treatment advice, all through an easy-to-use digital platform. By combining natural language processing and machine learning, our system offers a new way to bridge the linguistic and geographic divides in healthcare.

What sets this approach apart is its focus on enabling natural, two-way communication between patients and medical systems, regardless of language. We believe that improving our understanding is the first step toward improving care. When patients describe what they are experiencing in their own words, they are more likely to receive accurate and timely help.

In this study, we share the development and evaluation of our language translation system, including its architecture and performance with respect to healthcare delivery. Our findings suggest that AI can play a vital role in making healthcare more human-centered—by listening better, understanding more deeply, and reaching further than ever before.

2. Literature Review

A. Ghosh *et al.* [1] focuses on codemixed input text summarization in a medical context using the MM-CQs dataset, which combines Hindi-English codemixed medical queries with visual aids. The authors introduce a framework called MedSumm that leverages the power of large language models (LLMs) and visual language models (VLMs) for effective summarization.

Ishali Jadhav *et al.* [2] proposes a language modeling (LM) based approach for text classification of Hinglish text. They built a Universal Language Model Fine-tuned using the AWD-LSTM architecture on a Hindi-English code-switched corpus collected from various blogging sites, highlighting the effectiveness of deep learning architectures in handling code-mixed language data for disease prediction tasks.

Vaibhav Kumar *et al.* [3] proposes a supervised learning method for specialized domain named entity recognition (NER) tasks. The model consists of a multidimensional self-attention (MDSA) network and a Convolutional Neural Network (CNN) based model, emphasizing the importance of advanced machine learning techniques

to enhance the accuracy of disease prediction and facilitating doctor-patient interactions.

Y. Li and Qiao *et al.* [4] presents a pipelined mechanism for the machine translation of a bilingual language, specifically Hinglish to monolingual English. The study utilizes Python libraries such as NLTK and SpaCy, showcasing the application of BERT-based models to improve the classification and translation of patient messages in healthcare settings.

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

X. Yu *et al.* [6] highlights the efficacy of deep learning techniques, including 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 study 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 into English which is grounded in machine translation and Natural Language Processing (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 to improve the translation systems. They discussed the potential of using

4   *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

transformer models and attention mechanisms to enhance the fluency and contextual accuracy of translations, particularly in code-mixed language scenarios.

Rajesh V. *et al.* [12] examines the challenges and solutions in named entity recognition for code-mixed languages. They highlighted 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 for 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.

Wongso, W. *et al.* [14] discusses the role of evaluation metrics in assessing the performance of the 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 framework for enhancing healthcare communication through NLP techniques. They outlined future directions for integrating diverse linguistic capabilities into healthcare applications, aiming to improve accessibility and effectiveness for a broader range of linguistic communities.

A. Jha *et al.* [16] presents a comprehensive approach to multilingual Indian language neural machine translation using the mT5 transformer architecture. The study demonstrates the effectiveness of multilingual transformer models in handling diverse Indian languages, showcasing improved translation quality across multiple language pairs. The research emphasizes the importance of leveraging pre-trained multilingual models to address the challenges of low-resource Indian languages, providing a scalable solution for cross-lingual communication in the Indian subcontinent.

X. Sun *et al.* [17] focuses on the design and implementation of computer intelligent translation systems grounded in natural language processing principles. The authors explore the integration of advanced NLP techniques to enhance translation accuracy and contextual understanding. Their work emphasizes the importance of intelligent processing mechanisms that can handle semantic nuances and contextual variations in multilingual translation tasks, contributing to the development of more sophisticated translation frameworks.

S. Raza *et al.* [18] investigates the large-scale application of named entity recognition techniques in biomedicine and epidemiology domains. The study highlights the critical role of NER systems in extracting meaningful entities from biomedical texts, demonstrating their effectiveness in processing vast amounts of medical literature. The research underscores the importance of domain-specific NER applications and provides insights into the challenges and opportunities in applying NLP techniques to specialized medical and epidemiological datasets, which is particularly relevant for healthcare communication systems.

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 theoretical underpinnings that support our implementation choices and evaluation methodologies.

3.1. *Logistic Regression for Language Identification*

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 by using:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}} \tag{1}$$

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 the probabilities as:

$$P(y_i|x) = \frac{e^{W_i \cdot h + b_i}}{\sum_{j=1}^{C} e^{W_j \cdot h + b_j}} \tag{2}$$

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:

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{3}$$

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

3.3. *Long Short Term Memory (LSTM) Architecture*

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

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{4}$$

6 *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

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:

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{5}$$

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

$$\text{Top-}k \text{ Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \Vdash\left\{y_i \in \hat{Y}_i^{(k)}\right\} \tag{6}$$

where $\hat{Y}_i^{(k)}$ represents the set of top-k predicted disease classes for sample i, and $\Vdash\{\cdot\}$ 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 a 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]

4. Methodology

In this section, we detail the methodology used for the diagnosis of a disease based on the Hinglish text entered by the patient. The proposed framework involves two main stages - Translation of Hinglish text to English and medical diagnosis from the translated text. Fig. 1 illustrates the system design consisting of a translation module, followed by a fine-tuned BERT model for diagnosis. The modules are discussed in detail in the following sections.

4.1. *Translation from Hinglish to English*

The Hinglish to English translation module in our project is vital for facilitating seamless communication between patients and healthcare providers. By 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, thereby enabling efficient translation. For Hinglish text, the module translates English words to the Hindi Devanagari script and transliterates Hindi words to the Devanagari script to maintain linguistic integrity. Additionally, if we find common Hindi medical keywords and descriptions, we directly parse them into the corresponding English medical keyword to ensure that no important keywords are lost
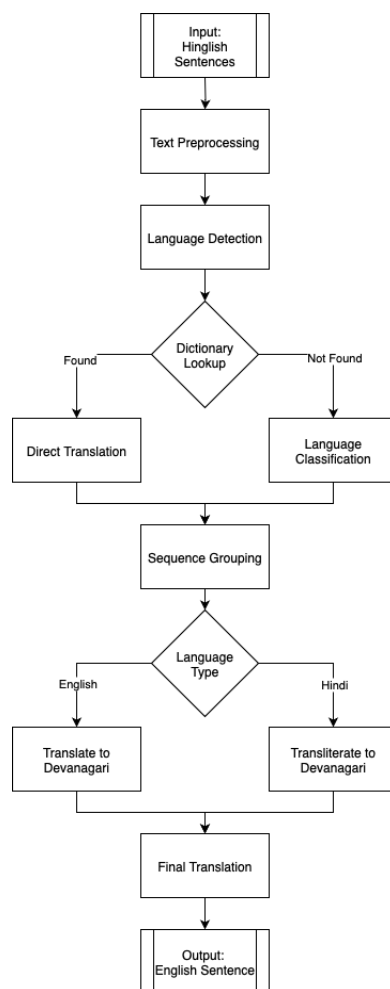
Fig. 1. System Diagram of the proposed model

during translation. Finally, the entire text was translated into English to ensure a 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 was used to categorize words as either English or Romanized Hindi.

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

3. Translation and Transliteration: English words were translated to Hindi Devnagri script using the Google Neural Machine Translation (GNMT) model, while

8   *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

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
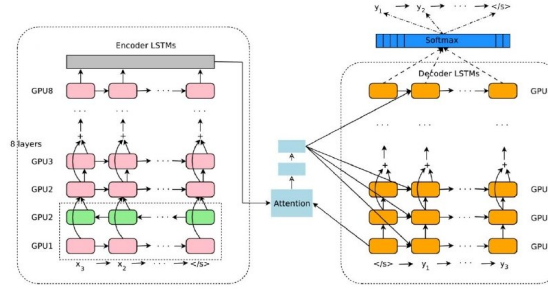   Add translated sentence to result list
  end for
  return list of translated sentences

Romanized Hindi words were transliterated to the Devanagri 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 keywords obtained

Fig. 2. Block diagram of Google Neural Machine Translation[24]
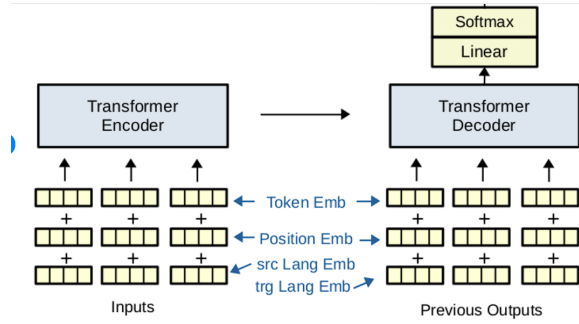
Fig. 3. Block diagram of Indic Translation[25]

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 that mapped common Hinglish medical terms to their English equivalents. The sample is shown in Tab. 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 the translation effectiveness. Each entry is carefully curated to retain essential medical terminology while enabling contextual understanding. The dataset serves as a benchmark for evaluating the translation model. By limiting the scope to common terms, we ensure a focused and practical approach to medical communication. This resource aids in bridging the language

barriers between patients and healthcare providers. Our dataset plays a crucial role in testing and refining the 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 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 was integrated with the Hugging Face textclassification pipeline. The methodology consisted of the following key steps:

Data Preparation: The input text, generated by the Hinglish-to-English translation module, was 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 was fine-tuned on the prepared dataset. A classification head was added to the BERT model, enabling multiclass disease prediction. Training was performed with the cross-entropy loss function and the Adam optimizer for learning rate management.

Pipeline Integration: The fine-tuned model was 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 capabilities

of BERT-base-cased model's nuanced language understanding, the module achieves precise and efficient disease diagnosis, supporting accurate and timely medical interventions. Some examples of the above data are listed in Tab. 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 2. Sample Input Output for k=1

5. Result and Analysis

The implementation of the Language Identification and Translation Module in our healthcare platform 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. As per Tab. 3, the Logistic Regression model outperformed the LSTM model with an accuracy of 88.84% compared with 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.

Model	Accuracy (%)
LSTM	50.57
Logistic Regression	88.84

Table 3. Comparison of Accuracy between LSTM and Logistic Regression

12   *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

5.2. *Translation and Transliteration*

The Google Neural Machine Translation (GNMT) model and IndicTrans model were employed for translation and transliteration tasks respectively. The GNMT model was used for translating English words to Hindi Devanagri 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 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 Devanagri sentence back into English using the GNMT model was executed effectively. This step is crucial for ensuring that healthcare providers can understand the patient's input, thereby facilitating seamless communication.

5.5. *Disease Diagnosis*

The fine-tuned BERT model was trained and the quantitative results are presented in Fig. 4 and Fig. 5. Tab. 4 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 4. Comparison between old and new approach

6. Conclusion

In conclusion, the Hinglish to English translation module significantly enhances communication between patients and healthcare providers by accurately translating and processing medical information, thereby facilitating better diagnosis and treatment recommendations. The integration of advanced natural language processing techniques and biological named entity recognition ensures that critical medical
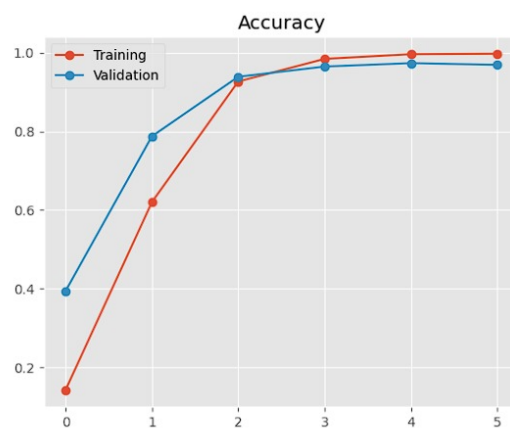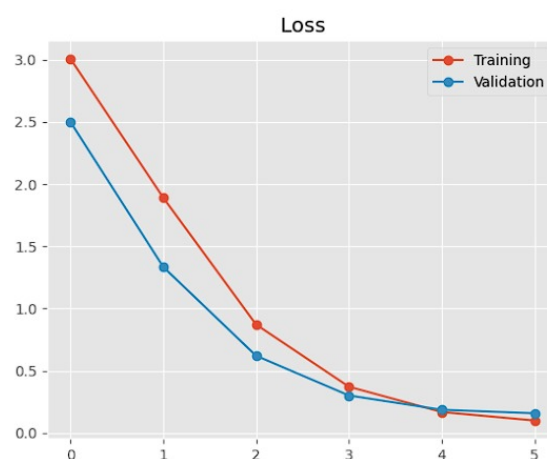
Fig. 4. Accuracy of fine-tuned BERT Model

Fig. 5. Loss of fine-tuned BERT Model

terms are preserved and understood in context. Future work will focus on refining the translation algorithms, expanding the domain-specific dictionaries for improved accuracy, and incorporating user feedback to enhance the adaptability and effectiveness of the system in diverse healthcare scenarios, ultimately aiming for a more comprehensive and user-friendly tool that can cater to a wider range of languages and dialects.

7. Declarations

Funding: This study received no external funding.
Conflict of Interest: The authors declare that they have no conflict of interest.

14 *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

Ethics Approval: I/We declare that the work submitted for publication is original, previously unpublished in English or any other languages), and not under consideration for publication elsewhere.

Consent to Participate/Informed Consent: Not Applicable.

Consent for Publication: I certify that all the authors have approved the paper for release and are in agreement with its content.

Data Availability: The datasets generated and/or analyzed during the current study are available in the following repositories:

1. Hinglish-English transliteration pairs:

`https: //github.com/anoopkunchukuttan/crowd-indic-transliteration-data/ blob/master/crowd_transliterations.hi-en.txt`

2. Dakshina Dataset for transliteration pairs:

`https://github.com/google-research-datasets/dakshina`

3. Natural Language Symptom Description:

`https://www.kaggle.com/datasets/niyarrbarman/symptom2disease`

4. Generated Dataset of Hinglish Symptom Description:

`https://github.com/Acdbb/Major_Project_Hinglish_Symptom_ Generated/blob/main/unique_symptoms_diagnosis.csv`

5. Translation Keywords:

`https://github.com/Acdbb/Major_Project_Hinglish_Symptom_ Generated/blob/main/hindi_dictionary.txt`

References

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 Tech-

nology 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. doi: `10.17577/ICCIDT2K23-221`.

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 (ICIT-COM)*, 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.

16   *Omkar Rao, Vineet Parmar, Hatim Sawai, Varsha Hole*

[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`.