e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

HINGLISH TO ENGLISH TRANSLATION SYSTEM

Komal Potdar[*1], Namrata Gaikwad[*2], Meenakshi Sutar[*3], Aditya Kurapati[*4], Ronak Dabade[*5], Gaurang Khanderay[*6]

[*1,2,3,4,5,6]Department Of Multidisciplinary Engineering Vishwakarma Institute Of Technology, Pune, 411037, Maharashtra, India.

ABSTRACT

The use of code-mixing, a linguistic phenomenon where speakers blend multiple languages, is widespread among non-English speakers globally, particularly on social media. In the Indian context, individuals frequently engage in code-mixing, combining English and Hindi in their online conversations, a practice colloquially known as Hinglish. This linguistic fusion creates a vast amount of unstructured text on platforms such as social media, blogs, and reviews. As code-mixing becomes increasingly prevalent, it poses a significant challenge to machine translation systems. In this research paper, explore the algorithmic techniques developed to address the complexities of handling code-mixed messages, specifically focusing on the Hinglish context. It discuss the limitations of the system and its implications in the evolving field of text mining, shedding light on the modern yet localized way of expression prevalent in Indian online communication. Through this study, the aim is to contribute to the understanding of code-mixing challenges and advance the capabilities of natural language processing systems in accommodating the diverse linguistic practices observed in social media discourse.

Keywords: Code-Mixing, Devnagari, English, Hinglish, Natural Language Processing, NLP, Transliteration, Translation.

I. INTRODUCTION

In today's tech-driven world, it's becoming normal to mix different languages when people talk or post on social media. People aren't using just one language as much anymore. This mix of languages makes a lot of data that machines find tricky to understand. While there's been a bunch of work on translating pure languages, we now need to pay more attention to studying and figuring out content in mixed languages. This study focuses on translation model specifically for code-mixed language, like Hinglish, in Natural Language Processing (NLP). This model helps machines better handle and make sense of the mix of languages when used in everyday communication.

Hinglish is a language that mixes Hindi and English, blending them in conversations, individual sentences, and even words. For example, you might say, "nahi mei nahi aa sakta," which means "no, I cannot come." This style of speaking is becoming popular because it's a modern way of talking that still connects with local culture. In India, a diverse country, there are people on both ends of the spectrum. On one side, there are Hindi speakers who can read and understand the Devanagari script. On the other side, there are tourists from abroad who may or may not fully understand the language. Since many local Indian markets, a big draw for foreign tourists, have vendors who only understand Devanagari, there's a need to use Hinglish as a way for these two extremes to communicate effectively.

The idea of this design is to restate Hinglish(Hindi English) which is combination of Hindi and English language to pure English language. The proposed model operates by exercising Hinglish as a standalone language, performing as a direct translator of code-mixed language into a pure form. This facilitates the analysis of content expressed in mixed languages and serves to enhance the commerce between machines and humans, fostering a more authentic communication experience.

II. RELATED WORK

S. H. Attri, T. V. Prasad, and G. Ramakrishna in [5] first determined if the sentence contained an expression or idiom, then extracted it. The phrase was tokenized and classed as Hinglish, English, or Hindi, depending on its original language. They then used morphological and reverse morphological analysis on each term. POS Tagging sorted the words after analysis and Translation was carried out. "MujhE file send kar as soon as possible" and "asap" were translated into Hindi. The resulting phrase in Hindi was "mujhE yathA shIghra sanchikA bhEj" which translates to "Send me the file as soon as possible" in English. 12,000 Hinglish terms were

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

labelled with idioms and translations. Pure Hindi sentences were much more accurate than pure English sentences. Thus, Hinglish is Hindi with English words added, using Hindi syntactic and semantic components instead of English ones. Because of this, it was found that Hinglish sentences translated into pure Hindi were more accurate than those translate into English.

There is a lot of study being done on code-mixed material, especially on language tagging. The Jhamtani et al. model was an ensemble model, (2014)[4] which was combination of two classi-fiers to form a LID mixed with Hindi-English code. The first classifier employed features such as word frequency, modified edit distance, and character n-grams, while the second classifier used the output from the prior one for the current word as well as languages and pos tag for neighboring words to provide the final tag.

Authors in [2] propose a four-phase pipeline for automatic Hinglish-to-English translation. They also compare "code switching" and "code mixing" and examine contextual problems such as "chalega" which meaning both "moving" and "will It work?" This study used no comparable corpus. The language was tagged, transliterated into Devanagari, translated from English to Hindi, combined with Hindi, and translated back into English.

III. METHODOLOGY

Overview

The methodology provides a systematic approach to develop Hinglish to English Translation System. It focuses on accurate translation, replace short notation and user interactivity. It uses a dataset that contains Hindi idioms for linguistic enrichment. It has different feature like virtual keyboard, virtual assistant, voice module and file module.

Technological Stack

Frontend

The user interface is developed using HTML, CSS, and JavaScript, encompassing the design elements, virtual keyboard functionality, and the integration of voice features.

Backend

Python is employed for backend development to handle translation logic, short notation replacement, and interaction with the SQLite database.

Database

SQLite serves as the database, storing short notations, long notations, and a dataset of Hindi idioms.

Database and Dataset

The database is structured with two primary columns: "Short Notation" and "Long Notation". This design facilitates efficient retrieval of long notations associated with replaced short notations.

The dataset consists of Hindi idioms in Devnagari script paired with their corresponding English translations. This dataset enhances the linguistic component of the translation system, contributing to improved accuracy and coverage.

Proposed System

The system translates the user's Hinglish text to pure English. Long notations take the place of short notations in the text. It also provides the text box with a speech capability so that people may hear it. Finding and replacing all brief notations in Hinglish text and translating it into full English are the primary objectives of the current study. From a provided Hinglish text, the model recognizes the short notation and extracts it. Hindi translations of the remaining texts will be provided. After that, the module receives the Hindi text and translates it into English. We'll use the dataset to identify and replace all of the short notations with long ones.

The system has several input modes analogous as keyboard, voice, or in a train, as well as virtual keyboard as an early- stage creations of the design. Input type file is useful because voice type will be useful if the user can't write/ type properly or if the user chooses to submit the Hinglish text file. In addition, the virtual assistant is assigned to answer the user's questions.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

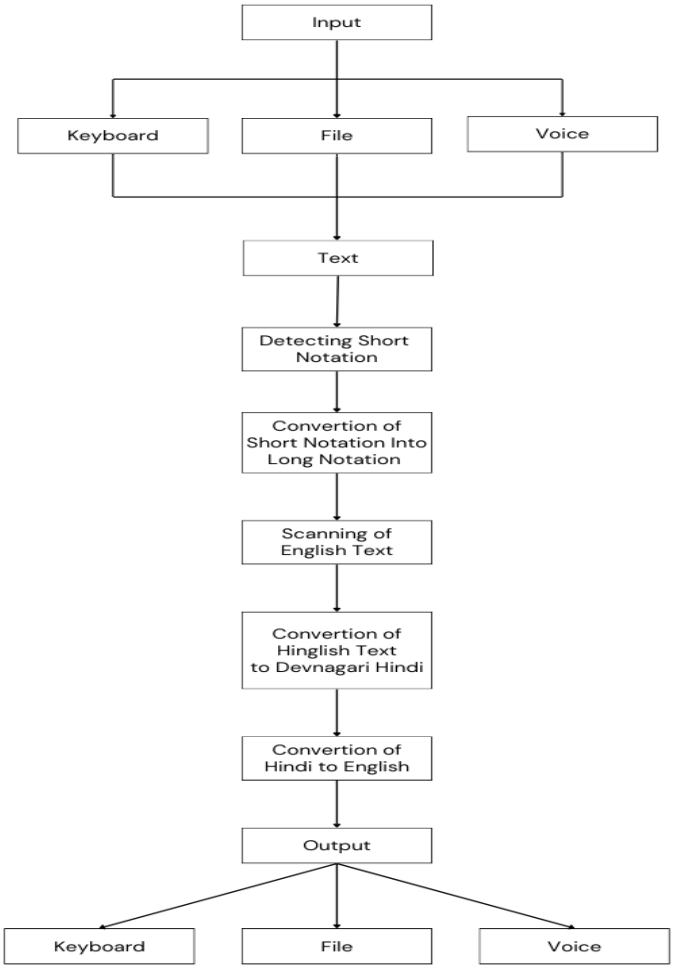

Figure 1: Proposed Architecture

IV. RESULTS

As shown in the given figures, a text is passed to the system which contains a short notation and Hinglish text. The system first checks for short notation in the given input. If any present then converts in the corresponding long notation from the database. Then it checks for English words in the given input and find its Hindi meaning. It performs transliteration and converts the whole text into devnagari script. Then system checks for idiom and fetch data from the dataset. Then translation is performed and final output is given.

```
(base) C:\Users\HP\CPP Project>python main.py
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\HP\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Enter Hinglish text:gm mera naam komal hai. me ek mahila abhiyantriki mahaavidyaalayaa ki chhaatraa hu.

**********I/P text***************
gm mera naam komal hai. me ek mahila abhiyantriki mahaavidyaalayaa ki chhaatraa hu.

**********Shortnotations available status***************
gm -> passing to database
Good morning : is available
mera -> passing to database
naam -> passing to database
komal -> passing to database
hai -> passing to database
. -> passing to database
me -> passing to database
ek -> passing to database
mahila -> passing to database
abhiyantriki -> passing to database
mahaavidyaalayaa -> passing to database
ki -> passing to database
chhaatraa -> passing to database
hu -> passing to database
. -> passing to database
**********Shortnotations removed text**************
Good morning mera naam komal hai . me ek mahila abhiyantriki mahaavidyaalayaa ki chhaatraa hu .
```

Figure 2: Output1

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:12/December-2023 Impact Factor- 7.868 www.irjmets.com

```
**********Identify English words**********
Good :  ['अ च्छा ']
morning :  ['अ च्छा ', 'सु बह']
mera
naam
komal
hai
. :  ['अ च्छा ', 'सु बह', 'mera', 'naam', 'komal', 'hai', '.']
me :  ['अ च्छा ', 'सु बह', 'mera', 'naam', 'komal', 'hai', '.', 'मु झे']
ek
mahila
abhiyantriki
mahaavidyaalayaa
ki :  ['अ च्छा ', 'सु बह', 'mera', 'naam', 'komal', 'hai', '.', 'मु झे', 'ek', 'mahila', 'abhiyantriki', 'mahaavidyaalayaa',
 'की']
chhaatraa
hu
. :  ['अ च्छा ', 'सु बह', 'mera', 'naam', 'komal', 'hai', '.', 'मु झे', 'ek', 'mahila', 'abhiyantriki', 'mahaavidyaalayaa',
'की', 'chhaatraa', 'hu', '.']

**********list items**********
['अ च्छा ', 'सु बह', 'mera', 'naam', 'komal', 'hai', '.', 'मु झे', 'ek', 'mahila', 'abhiyantriki', 'mahaavidyaalayaa', 'की',
 'chhaatraa', 'hu', '.']
**********Convert list -> sentence**********
अ च्छा  सु बह mera naam komal hai .  मु झे ek mahila abhiyantriki mahaavidyaalayaa की chhaatraa hu .

**********Final result**********
अ च्छा  सु बह  मे र  ना म्  कोमल्  है  ।  मु झे  एक्  महिल अभिय  न्त्रि  कि  महा वि  द्या लया  की  छा त्रा  हु  ।
21 चा र  दिन  की  चां दनी  फिर  अं ध रि  रा त
21 दरि द्र ता  कहल  की  जड़  है
Good morning, my name is Komal. I am a student of a women's engineering college.
```

Figure 3: Output2

For any translation to be successful, the system needs input in proper format without any spelling mistakes and grammatical errors otherwise it may get confused and will provide abnormal output. Fig.1 and Fig.2 outputs are from console where step by step explanation is given.

V. LIMITATIONS

- Hinglish is most frequently encountered on informal platforms. Since informal writing rarely adheres to punctuation, correct spelling, and correct grammar, this adds an additional layer of complication to the translations.
- Several words used in Hinglish are also found in English [2], in such cases it creates an obstacle for the model to determine the source language and thus choose between transliterations and translations.
- There are no standard spellings in Hinglish; most users rely on the phonetics of the word to determine its romanised spelling [1], thus resulting in a variety of words with the same meaning but different spellings. For instance, “Nahi”, “Nai”, “Nhi” all mean “No” in English. This makes complication for the model.
- A large number of Hinglish terms have several meanings that can be determined from the sentence context alone [1]. An example of this could be the term “chalega” which in some sentences means “will work” and in some sentences means “walk”.

VI. FUTURE SCOPE

The identified limitations in the current Hinglish to English translation system pave the way for promising future developments. Future work could focus on refining the system to better handle the informal nature of Hinglish by incorporating advanced natural language processing (NLP) techniques that account for variations in punctuation, spelling, and grammar. Additionally, exploring context-aware machine learning models may contribute to improved disambiguation of meanings associated with Hinglish terms, addressing the challenge of multiple interpretations. Introducing dynamic spell-checking algorithms that adapt to the phonetic variations in Hinglish could enhance the accuracy of translations. Furthermore, the integration of sentiment analysis and cultural context recognition could add another layer of sophistication to the translation process, ensuring that nuances unique to Hinglish expressions are accurately captured. Overall, future advancements could involve a holistic approach, combining linguistic analysis, machine learning, and cultural context understanding to create a more robust and contextually aware Hinglish to English translation system.

VII. CONCLUSION

This paper addresses the complexities of translating Hinglish to English, recognizing challenges related to informal expressions and diverse linguistic variations. Users are advised to input accurate spellings for optimal results. While the system is robust, further advancements, such as dynamic spell-checking, remain avenues for

exploration. This project lays the groundwork for future research, aiming to create a more precise and culturally sensitive Hinglish to English translation experience.

VIII. REFERENCES

[1] Jadhav, I., Kanade, A., Waghmare, V., Chandok, S.S., Jarali, A.: Code-Mixed Hinglish to English Language Translation Framework. In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS). pp. 684–688. IEEE (2022). https://doi.org/10.1109/ICSCDS53736.2022.9760834.

[2] Dr. S.V. Kedar, Sakshi Bhangale, Kunal Deokar, "Translation: Code-Mixed Language (Hinglish) to English", IJARSCT Volume 2, Issue 3, May 2022.

[3] IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 5, No 2, September 2014 ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784www.IJCSI.org

[4] Attri, S.H., Prasad, T.V., Ramakrishna, G.: HiPHET: A Hybrid Approach to Translate Code Mixed Language (Hinglish) to Pure Languages (Hindi and English). Computer Science. 21, (2020). https://doi.org/10.7494/csci.2020.21.3.3624.

[5] Rao, Ashwini and DSouza, Nicole and Patel, Devarsh and Saravta, Jigyashu, Development & Study of Hinglish to English Translation and Classification Techniques. Available at SSRN: https://ssrn.com/abstract=4510958 or http://dx.doi.org/10.2139/ssrn.4510958.

[6] Chéragui, Mohamed Amine. Theoretical Overview of Machine Translation. Proceedings ICWIT. 2012.

[7] Hutchins W.J, Somers H L. An introduction to machine translation. London: Academic Press.1992: