A Systematic Review on Machine Translation and Transliteration Techniques for Code-Mixed Indo-Aryan Languages

Rukshan Dias
School of Computing
Informatics Institute of Technology
Colombo 06, Sri Lanka
rukshandias002@gmail.com

Deshan Sumanathilaka
Department of Computer Science
Swansea University
Wales, United Kingdom
t.g.d.sumanathilaka@swansea.ac.uk

Abstract

In multilingual societies, it is common to observe the blending of multiple languages in communication, a phenomenon known as **Code-mixing**. Globalization and the increasing influence of social media have further amplified multilingualism, resulting in a wider use of code-mixing. This systematic review analyzes existing translation and transliteration techniques for code-mixed Indo-Aryan languages, spanning rule-based and statistical approaches to neural machine translation and transformer-based architectures. It also examines publicly available code-mixed datasets designed for machine translation and transliteration tasks, along with the evaluation metrics commonly introduced and applied in prior studies. Finally, the paper discusses current challenges and limitations, highlighting future research directions for developing more tailored translation pipelines for code-mixed Indo-Aryan languages.

1 Introduction

Machine Translation and Transliteration have made progress in mapping barriers to cross-lingual communication, especially in Indo-Aryan languages (Perera and Sumanathilaka, 2025b). Indo-Aryan languages are spoken primarily in north and central India, as well as in a few neighbouring countries. The Indo-European language family includes the Indo-Aryan languages as a subfamily, comprising languages such as Hindi, Bengali, Marathi, Sinhala, and Urdu (Pal and Zampieri, 2020). With the advancement of Web 2.0, most digital platforms have become multilingual. During the past decade, the use of social networks and other digital platforms has increased significantly. With the web being multilingual and the increasing demand for social networks, users have begun to adopt their native language on these platforms. Code-mixed data, being more noisy and increasing in prevalence, makes developing a robust machine translation or transliteration system critical for cross-lingual communication and information access.

Several studies have been conducted on Indo-Aryan languages, focusing on monolingual translation or transliteration between Roman scripts to native scripts (Sumanathilaka et al., 2025a; Athukorala and Sumanathilaka, 2024; Herath and Sumanathilaka, 2024). However, several gaps exist, including a lack of code-mixed parallel corpora, insufficient benchmarking on real-world noisy text, a scarcity of research addressing both translation and transliteration combined, and limited systematic reviews that consolidate the state-of-the-art. Numerous studies have investigated MT and transliteration for Indo-Aryan languages, typically focusing on transliteration between native and Roman scripts or single translations (e.g., Hindi-English, Sinhala-English). Few recent studies have attempted to address code-mixed inputs for translation and transliteration tasks, often using NMT models, subword-level embeddings, word-level language identification, etc (Jadhav et al., 2022; Patel and Parikh, 2020; Gupta et al., 2024). However, it was identified that there are some challenges and gaps that have not been addressed. Table 1 presents a comparison between translation and transliteration across multiple languages.

In this paper, the authors have conducted a systematic review of existing studies on machine translation and transliteration for code-mixed Indo-Aryan languages. Previous studies on various datasets, preprocessing techniques, model designs, and evaluation approaches have been analyzed and reviewed. The challenges posed by informal Romanized writing and the limited scope of models evaluated in code-mixed contexts are highlighted in this review. Language identification as a preprocessing step, using multilingual

Proceedings of the Twelfth Workshop on Asian Translation (WAT 2025), pages 66–77

Language	Translation		Transliteration	
	Source	Target	Source	Target
Sinhala	My country	මගේ රට	mage rata	මගේ රට
Hindi	My country	मेरा देश	mera desh	मेरा देश
Tamil	My country	என் நாடு	en naadu	என் நாடு

Table 1: Examples of Translation and Transliteration

pre-trained models and collaborative modelling of transliteration and translation, are among the new avenues that have been explored.

From the reviewed literature, the author has identified that the target language of the translation is predominantly English, as seen in both studies and datasets. However, there are scenarios where the native indo-aryan language is the target language. Hence, it can be considered that translation tasks in this domain involve English-to-Indo-Aryan translation, Indo-Aryan-to-English translation, and Indo-Aryan-to-Indo-Aryan transliteration, depending on the dataset and the application.

This study makes the following key contributions:

- Provides a comprehensive comparative analysis of code-mixed translation and transliteration on Indo-Aryan languages.
- Presents one of the first systematic explorations of machine translation and transliteration applied to code-mixed Indo-Aryan languages.
- Discusses persistent challenges in code-mixed translation and transliteration and proposes future directions for research.

The remainder of this review paper is structured as follows. The second chapter outlines the methodology used to conduct this review, discussing the selection of studies, evaluation criteria, search strategies, and keywords. The third chapter would highlight the linguistic characteristics of code-mixed text and sociolinguistic motivations. Then it would analyse and review approaches, techniques, and architectures that have been explored to date in the domain. Followed by an overview of existing datasets and evaluation metrics employed in previous studies. Finally, the paper would discuss current limitations and gaps that have been identified as unaddressed and provide paths for future research.

2 Related Works

This section provides a review of the survey studies related to the domain of code-mixing and machine translation. The survey by Dabre et al. (2020) provides a comprehensive review of multilingual neural machine translation, with more focus on architectural paradigms, transfer learning strategies, parameter sharing mechanisms, and multilingual modelling techniques to improve translation quality. However, it has not reviewed the unique characteristics, challenges of code-mixed data, nor transliteration-related studies.

The survey by Thara and Poornachandran (2018) provides an introductory review and analysis of code-mixing and its impact on various NLP tasks, including POS tagging, NER, sentiment analysis, and machine translation. However, the work aims to provide a broad overview of code-mixing rather than an in-depth analysis of a specific task. It does not provide a comprehensive discussion of model architectures, datasets, or evaluation protocols relevant to these tasks. The survey covers code-mixing across multiple language families, but it does not specifically address the challenges posed by languages such as Sinhala, Hindi, Bengali, or other Indo-Aryan languages. The work by Hidayatullah et al. (2022b) is a systematic review that focuses on language identification in code-mixed text. It reviews existing language identification techniques, datasets and challenges in identifying language in multilingual social media content. This survey is valuable for understanding preprocessing and language segmentation, which is important in downstream tasks.

However, the survey does not discuss the topic of machine translation. Although accurate language identification can influence the quality of downstream machine translation, the survey does not explore the relationship between language identification and a machine translation system. Collectively, previous surveys do not provide a review focused on translation and transliteration techniques for Indo-Aryan code-mixed languages, nor on the linguistic and orthographic challenges associated with code-mixed, romanised Indo-Aryan texts. The present study fills this gap by connecting current methodologies, datasets, and existing challenges. Offering a focused and domain-specific perspective not available in prior literature.

3 Methodology

A comprehensive search was conducted to identify relevant papers in the domain. Academic databases, such as IEEE Xplore, Google Scholar, ACL Anthology, and ResearchGate, have been considered to identify relevant studies. Apart from searching the mentioned academic databases, several papers have been recognized from references cited in published papers, especially survey papers. A wide range of search keywords has been used to search relevant literature. In detail, search terms like "code-mixed translation/transliteration", "code-mixed indo-aryan languages", "code-mixed Romanized languages", "Hindi-English code-mixed translation", "Sinhala-English code-mixed translation", etc, have been followed.

Considering the limited research in this domain, this literature review task has focused on studies and papers published between 2018 and 2025. The first author has carefully labelled the papers for their relevance following a pre-structured extraction mechanism, and quality assessment was performed based on the Critical Appraisal Skills Programme checklist, examining the clarity, appropriateness of methodology, rigour of analysis, and relevance. The screening and selection process is presented in the Figure 1.

Every paper that was published before 2018 has been excluded. This review also includes studies that support the topic of code-mixed indo-aryan translation and transliteration. Papers that have introduced new algorithms and datasets related to MT, that follow the Neural Machine Translation approach, have been considered for this systematic review. Several papers have been excluded because they are not within the scope of Indo-Aryan languages. Authors have identified that there are different studies, apart from MT, on code-mixed text, such as sentiment analysis and Language identification, which have been excluded. Selected papers have been grouped by language type.

4 Code-Mixing background in Indic languages

Language mixing is a result of multilingual language usage across people. This behaviour is more common in multicultural and multilingual societies, such as most countries that use Indo-Aryan languages. Code-mixing is the practice of mixing multiple languages in a single instance

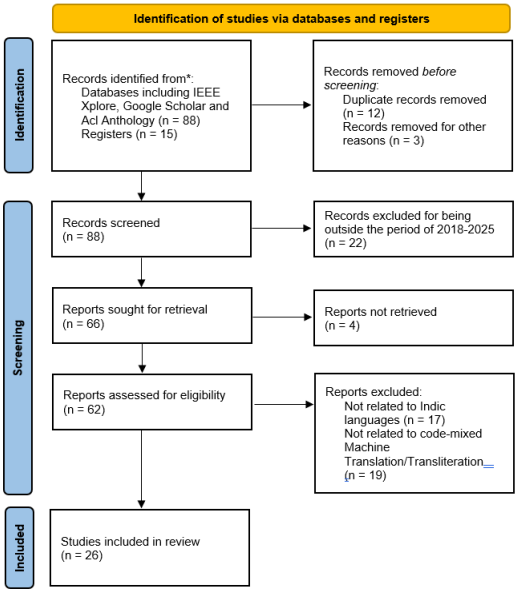

Figure 1: PRISMA flow of literature selection process

(Thara and Poornachandran, 2018). There are two main types of Code-mixing. The first one is Inter-Sentential Switching, which occurs at the boundaries of sentences. One sentence would have a single language type, but multiple sentences would have multiple language types. For example, "*Mujhe abhi jaana hai. I'll call you later*". '*Mujhe abhi jaana hai*' is Hindi and "*I'll call you later*" is English. The second one is Intra-Sentential Switching, which occurs within the same sentence. Therefore, borrowings from different languages can be found in a single sentence (Thara and Poornachandran, 2018). As an example in Hindi-English, "*Main kal office meeting attend karunga*". When considering the languages in code-mixed text Myers-Scotton (1997) has provided a theory called the Linguistic Matrix Language Frame (MLF) theory. In this concept, the dominant language is defined as the Matrix language, and the secondary language is defined as the Embedded language. This is applicable for both code-mixed and code-switched text (Iakovenko and Hain, 2024). The example on Sinhala can be found in Figure 2.

The main reason people tend to use code-mixed language in their day-to-day conversation is that it can express feelings easily and effectively (Sumanathilaka et al., 2023). Use of code-mixing is more common among younger demographics and urban populations (Senaratne, 2009). Due to globalization, most people have adapted to En-

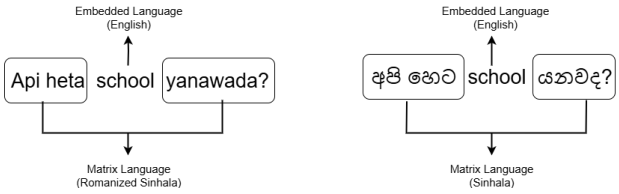

Figure 2: Matrix and Embedded language in code-mixed

glish as the universal language. This increases the number of people who are bilingual or multilingual, with some level of understanding in English. For this reason, it is more common to see people code-mix their native language with English. In the same way code-mixed language is used in verbal communication, it is also common to see this in written communication, especially on informal communication platforms like social media. In social media communication, it has been found that people tend to use their native language in a Romanized way. Despite the availability of native Unicode keyboards, many users prefer Romanized for its ease and convenience (De Silva, 2021).

5 Current state of Code-Mixed Indo-Aryan language translation and transliteration

This section contains a summary of the current state of Code-mixed translation and transliteration in Indo-Aryan languages. The language pairs presented have been carefully selected for discussion, with the choice driven by both the availability of resources and the identified importance of these pairs for advancing research in code-mixed translation and transliteration.

5.1 Hindi - English

Hindi, being one of the most popular languages in India, has several studies that have attempted to translate and transliterate code-mixed Hindi-English content. Bhowmick et al. (2023) have proposed a model to translate Roman Hindi codemixed sentences to monolingual English. First, to train the translation model, they have performed synthetic code-mixed sentence generation by training a mT5 model. The translation model training has been conducted on both augmented and manually scraped data, with 120000 and 4000 sentences. The overall translation pipeline would contain two models: the first is the correction model, which converts Roman Hindi to Devanagari and outputs a Mixed-Script sentence. The second one is the Translation model, which receives the output of the correction model and converts it to monolingual English. (Nair and Gupta, 2024) has proposed a study on exploring the capabilities of different LLMs in translating code-mixed Hindi-English data to English. While evaluating against state-of-the-art decorder-only models/LLMs like GPT-4, Gemini, GEMMA 2, BLOOMZ-3B, and Navarasa 2.0, the zero-shot prompting technique has been followed across all selected models. The Gemini model has outperformed other evaluated models on Roman Hindi code-mixed translation, achieving a BLEU score of 20.82%.

Another study(Gahoi et al., 2022) has been conducted on code-mixed Roman Hindi to English translation. It has utilised the PHINC dataset (Srivastava and Singh, 2020) to train the model, which contains 13,738 parallel code-mixed Hinglish and English sentences. The model has been trained by fine-tuning Salesken AIs pre-trained model mentioned in Huggingface Transformers. Upon evaluating the results, it was identified that the system struggled to translate long sentence inputs. This task returned a results score of 0.41493 for ROUGE-L and 0.80804 for WER Metrics. This paper Jadhav et al. (2022) also has suggested a novel solution to code-mixed Hinglish to English. It will first undergo language identification using an LSTM-based neural method with a dataset of 25,000 Hindi-English words. The identified English words would be tagged as 'en', and Hindi words would be tagged as 'hi'. The identified Hindi words would be transliterated using the Google Transliteration API. Words tagged as English would translate to Hindi using an NMT model. Finally, it would get concatenated to produce the final Hindi output. This has achieved a BLEU score of 0.737 and WER 0.238. This study has demonstrated that incorporating a language identification model into a code-mixed translation task enhances the accuracy of the output. However, it has been stated that the Language Identification model failed to identify language in ambiguous words and based on context.

5.2 Sinhala - English

Sinhala is the most used language in communication in Sri Lanka. Although there are keyboards for Unicode Sinhala, people tend to use the romanized version of Sinhala because of ease (De Silva,

2021; Perera et al., 2025). This creates the need for a translation system.

A recent study (Senanayaka et al., 2024) has been conducted on processing code-mixed Singlish text with RAG implementation for document retrieval. This model could translate English-Singlish code-mixed sentences to English by fine-tuning the LLaMA-2 7B parameter model. This is the first study to develop an LLaMA-based RAG framework tailored for code-mixed Singlish. A Synthetic parallel corpus has been generated using Claude-3-Sonnet. It has been stated that the transformer's attention mechanism enables it to process larger sentences more accurately than other models. This model achieves scores of BLEU 0.1347, ROUGE-1 0.3732, and METEOR 0.5923.

(Kugathasan and Sumathipala, 2022) has done a study on translating code-mixed Singlish to Sinhala. The authors have manually created a dataset comprising 5,000 code-mixed sentences with relevant Sinhala translations. The model is an LSTM Seq2Seq model with a Teacher Forcing mechanism. One of the main novelties of this code-mixed translation task is that, in preprocessing, it utilises the Levenshtein Edit Distance to address ambiguous words in Singlish to some extent. Overall, the study has received a BLEU score of 0.3389. Although a wide range of studies have been conducted on Romanized Sinhala transliteration (Sumanathilaka et al., 2025b; Dharmasiri and Sumanathilaka, 2024), these works exclusively focus on non code-mixed language data. Nevertheless, they underscore the significance of addressing code-mixed scenarios, given their prevalence and practical importance in real-world applications.

5.3 Gujarati - English

(Patel and Parikh, 2020) has proposed an approach to translate code-mixed romanized Gujarati sentences to Gujarati script. This approach would leverage a language identification model that tags the language before performing the translation or transliteration. Authors have identified that Gujarati-English code-mixed data can create ambiguity, for example, the word 'mate' in English, but in Gujarati, it means 'for' in a contextual sense. To overcome that problem, the Hidden Markov Model approach has been used to predict accurate language based on the context.

5.4 Bengali - English

A study has been conducted (Bhowmick et al., 2023) to translate Bengali roman codemixed sentences into monolingual English, utilising mT5 and integrating a correction model and a translation model into the pipeline.

(Shibli et al., 2023) addressed the task of automatic back-transliteration of code-mixed Romanized Bengali into native script Bengali. Their approach generates multiple candidate forms through rule-based phonetic mappings. It applies statistical language models for ranking, similar in spirit to similarity-based scoring and graph-ranking techniques. By resolving ambiguities in noisy romanized input, their system enables more accurate input for subsequent translation models.

5.5 Other Code-Mixed transliteration

Amin et al. (2023) focuses on generating Marathi-English code-mixed text, addressing the lack of code-mixed resources. The proposed method is based on Matrix Language Frame theory, which extracts English phrases identified and transliterated into the Devanagari script, ensuring that they phonetically blend with the surrounding text.

The paper (Wisal et al., 2022) proposes an approach to translating and transliterating code-mixed Roman Urdu-English into Urdu. It utilizes the "g2p-multilingual-byT5-small" deep learning pre-trained model and fine-tunes it with a corpus of code-mixed, romanized Urdu and Urdu translations, which was created by the authors. The system has achieved a BLEU score of 66.73%, providing a strong foundation for noisy low-resource language translations. It has been identified that the model struggles when handling rare vocabulary, culture-specific words, and short sentences.

6 Datasets and Evaluation metrics

6.1 Code-Mixed Datasets

To produce a robust machine translation and transliteration system for code-mixed Indic languages, the quality of the data on which the model was trained was heavily dependent. This section describes publicly available datasets that could be used for code-mixed translation and transliteration tasks on different Indic languages.

6.1.1 Hindi - English

The PHINC dataset (Srivastava and Singh, 2020) is a large parallel dataset with more than 13,000

Study	Languages	Task	Model Type	Dataset	Evaluation	Notes
Hindi						
Bhowmick et al. (2023)	Mixed Roman Hindi/Bengali to English	Translation	mT5 Seq2Seq (Two-step pipeline: Correction + Translation)	5,100 Hindi-English CM	BLEU, METEOR, ROUGE	Mixed Script Augmentation (Roman + native script) improves MT performance.
Nair and Gupta (2024)	Mixed Hindi to English	Translation	LLMs (BLOOMZ-3B) with LoRA and prompting	Not specified	BLEU, chrF++	Demonstrates LLM performance for Indic CM translation.
Gahoi et al. (2022)	Mixed roman Hindi to English	Translation (CM → Monolingual)	mBART (Transformer-based)	Train 8,060, Val 942, Test 960	ROUGE 0.41493; WER 0.80804	Transliteration to Devanagari + Hindi parallel text improves MT.
Jadhav et al. (2022)	Mixed Hinglish to English	Translation	LID layer + GNMT pipeline	25,000 Hindi-English word for LID	BLEU 0.737; TER 0.256; WER 0.238	Uses intermediate Hindi translation via LID for improved accuracy.
Sinhala						
Senanayaka et al. (2024)	Mixed Singlish to English	Bidirectional Translation (RAG)	LLaMA-2 7B + LoRA	100 annotated data + synthetic data	BLEU 0.1347; ROUGE-1 0.3732	Synthetic corpus generation; perplexity reduced to 11.95.
Kugathasan and Sumathipala (2022)	Mixed Roman Sinhala to Sinhala	Normalization; Transliteration; Translation	Seq2Seq LSTM with teacher forcing	5,000 SCM sentences	BLEU 0.3389	Handles OOV words, slang, and inconsistent Romanization.
Gujarati						
Patel and Parikh (2020)	Mixed Roman Gujarati to Gujarati	LID + Normalization + Translation	Naive Bayes + HMM + Dictionary methods	1200 manually created sentences	Manual / API comparison	Hidden Markov Model to predict language.
Bengali						
Shibli et al. (2023)	Roman Banglish to Bengali	Back-Transliteration	Nine transliteration models + BERT similarity	5,000 collected; 1,000 for evaluation	BLEU, ROUGE, WER, WIL	Addresses varied romanization; Google Translate performed best.
Marathi						
Amin et al. (2023)	Marathi-English (Minglish)	CM Text Generation	Linguistic code-mixed generation algorithm	Uses parallel EN-MR corpus	CMI = 0.2; DCM = 7.4	Generates realistic Marathi-English CM sentences.
Urdu						
Wisal et al. (2022)	Mixed Roman Urdu to Urdu	Translation	T5-based multilingual Transformer	17,689 manually created	BLEU 66.73	HuggingFace g2p_multilingual_byT5_small used; dataset created by volunteers.

Table 2: Summary of reviewed methods for Code-Mixed translation and transliteration

code-mixed Romanized Hindi paired with English translations. These sentences have been scraped from social media platforms and utilise the support of six existing corpora that were created for other NLP tasks. Different preprocessing tasks were conducted when creating the corpus to ensure its quality. This contains data from various domains, including Bollywood, sports, politics, and social events.

The Dakshina dataset (Roark et al., 2020) contains resources for 12 different South Indian languages, including Hindi, Bengali, Telugu, Tamil, and Sinhala. It has over 12 million pairs of Romanized to their native script forms. Unlike the PHINC, the Dakshina dataset is more generic and is used in a wider range of NLP tasks, including machine translation, which makes it more suitable for Indic languages.

The L3Cube-HingCorpus (Nayak and Joshi, 2022) is considered the largest code-mixed Hindi-English corpus when compared with other state-of-the-art datasets. It consists of 52.93 million sentences(1.04 billion tokens) collected from Twitter to ensure broader domain coverage and to address the lack of large scale code-mixed Hinglish resources. Unlike the previous two datasets mentioned, L3Cube-HingCorpus does not include parallel translations. Hence, it is not a dataset that could be directly used in translation tasks. However, several studies have outperformed the state-of-the-art results using this corpus.

The HinGe dataset (Srivastava and Singh, 2021) has been introduced to address the scarcity of quality Hindi-English code-mixed resources. The foundation of the HinGe dataset is sourced from the Hindi-English parallel corpus from IIT Bombay (Kunchukuttan et al., 2018). This dataset is structured into two components: Human-

generated and Machine-generated sentences. It contains a high-quality collection of 4,803 human-generated sentences, translations annotated with five expert annotators. On the machine-generated side, a total of 3,952 Hinglish sentences have been synthetically generated using two rule-based algorithms: Word-aligned Code-Mixing (WAC) and Phrase-aligned Code-Mixing (PAC). It has been mentioned that this corpus can be used for NLP tasks, such as language identification and POS tagging, in addition to machine translation.

This study (Sheth et al., 2025) has identified that synthetically generated data fails to capture the nuances of real language usage, and human annotation is crucial for creating a high-quality, natural code-mixed text resources. COMI-LINGUA (Sheth et al., 2025) has been developed to address this gap by providing the largest manually annotated dataset for code-mixed text. The translation annotation was performed by three expert annotators who are fluent in both Hindi and English. The dataset has been validated using Fleiss' Kappa measure. The COMI-LINGUA dataset is mainly structured for five different NLP tasks: word-level language identification, sentence-level language identification, part-of-speech tagging, name entity recognition, and machine translation with sentences in Romanized Hindi, Devanagari Hindi, and English.

6.1.2 Sinhala - English

Sinhala, being a low-resource language for Sinhala-English code-mixed, has very limited datasets available. Though large transliteration datasets exist (Sumanathilaka et al., 2024), the availability of code-mixed and properly annotated corpora is limited. This work by Kugathasan and Sumathipala (2022) has provided a corpus that could be used for translating code-mixed Singlish (Sinhala-English) to Sinhala. This corpus consists of over 5,000 parallel code-mixed sentences with their relevant Sinhala translations. There are some other datasets that contain code-mixed Romanized Sinhala for other NLP tasks like language identification, sentiment analysis, etc (Uthpala and Thirukumaran, 2024; Smith and Thayasivam, 2019). But for machine translation tasks, there are extremely limited datasets that could be used.

6.1.3 Bengali - English

The BNSENTMIX dataset (Alam et al., 2025) comprises diverse Bengali-English code-mixed sentences, totalling 20,000. The data has been collected from social media platforms and manually annotated the translations. While this is not a direct translation dataset, it could enhance machine translation pipelines for code-mixed Bengali-English.

6.1.4 Urdu - English

The work by Wisal et al. (2022) has attempted to translate romanised code-mixed Urdu-English text to monolingual Urdu. The authors have annotated 17689 code-mixed Roman Urdu sentences with their relevant translation, with the help of a few annotators.

6.2 Evaluation metrics for code-mixed

Evaluating code-mixed text has its own challenges due to the informal nature and diversity of language. There are several standard matrices for code-mix tasks that have been in use for decades, as well as other matrices that have evolved from these standards. In this section, different evaluation matrices could be used for code-mixed machine translation and transliteration tasks.

6.2.1 BLEU

This is considered the most widely used evaluation metric for machine translation tasks. It calculates the score by measuring the precision of n-grams in candidate translation against the reference translation, with Brevity Penalty to address translations that are short (Papineni et al., 2002). In this review, it has been identified that BLEU often correlates poorly when compared against human judgment.

6.2.2 METEOR

METEOR or Metric for Evaluation of Translation with Explicit Ordering is an improvement done on BLEU by calculating the score not just based on exact match, but stem and synonym (Banerjee and Lavie, 2005).

6.2.3 chrF++

The chrF++ is an enhanced version of chrF, which combines character-level matching with the lexical accuracy of word-level matching(Popovi, 2017). Since this benefits both word-level and character-level analysis, some recent code-mixed studies have utilised this approach for evaluation (Nair and Gupta, 2024).

6.2.4 Word and Character Error rate

Both of these evaluation metrics are logically similar and are based on the concept of Levenshtein distance, which measures the number of edits required to transform one string into another. WER compares the generated text with the reference text on the number of substitutions, deletions, and insertions to make them identical. Similar to WER, the CER would operate on a character level instead of a word level (Gohider and Basir, 2024). The equations for WER and CER would operate as follows:

$$\text{WER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}} \qquad \text{CER} = \frac{\text{S} + \text{D} + \text{I}}{\text{N}}$$

6.2.5 Translation Edit rate

The Translation Edit rate is an extended version of WER and CER. It would also consider the word shift when measuring the score. Word shift indicates the movement of the location of particular text. A lower TER score indicates a better translation (Snover et al., 2006).

$$\text{TER} = \frac{\text{S} + \text{D} + \text{I} + \text{H}}{\text{N}}$$

6.2.6 Human Evaluation

Because code-mixing admits many sentimentally correct forms that other metrics, like n-gram, fail to capture, human judgment would still be the most accurate method of evaluation. Recent studies have demonstrated that standard metrics, such as BLEU, can be misleading for code-mixed outputs, and that human assessments better reflect fluency and the preservation of code-mixing patterns (Gupta et al., 2024; Vavre et al., 2022). Case studies comparing automatic and human evaluations similarly show that human evaluations detect semantic faithfulness and nuanced phenomena introduced by code-mixing that automatic metrics would miss (Nguyen et al., 2023).

7 Gaps and Challenges in Machine Translation and Transliteration for Code-mixed Indo-Aryan Languages

Although recent advancements have been made in the domain of machine translation and transliteration for code-mixed Indo-Aryan languages, several gaps and challenges remain that can be identified. In this section of the review, we will discuss those identified gaps and challenges in this domain.

7.1 Limited Datasets

When it comes to machine translation and transliteration tasks in code-mixed Indo-Aryan languages, datasets play a significant role in the system's output. There are very limited datasets available for code-mixed Indo-Aryan languages, particularly those that can be utilised for machine translation and transliteration tasks. Through this review, it has been identified that there are more parallel corpus for code-mixed Hindi-English rather than other Indo-Aryan languages. Languages like Sinhala, Gujarati, and Bengali have an extremely limited number of datasets that can be used for translation and transliteration tasks. Hence, ensuring a gold standard parallel corpus is essential, especially for languages with limited datasets.

7.2 Transliteration Ambiguity

Transliteration ambiguity refers to a word that has multiple senses in the context of translation and transliteration (Perera and Sumanathilaka, 2025a). Identifying the correct meaning of the word is significantly important to process code-mixed language NLP tasks, including machine translations (Hidayatullah et al., 2022a). As an example in the Sinhala-English sentence *"Ape rate weather eka"*, the word *'rate'* in the romanised Sinhala format refers to the country. Hence, in this context, the word *'rate'* cannot be considered an English word which has the sense of a *"measure, quantity, or frequency"*. Most of the papers reviewed acknowledge addressing Transliteration ambiguity as a challenge, and only a limited number of studies have attempted to provide solutions for this issue in machine translation and transliteration tasks.

7.3 Non-Standard words

Since code-mixing is more commonly associated with social media or informal communication, it is more likely to contain non-standard words. (Hidayatullah et al., 2022a) has categorized the non-standard words into four main types: non-standard spelling, mixing words and numeric or special characters, word exaggeration or wordplay, and abbreviated words. Table 3 describes the types of non-standard words with examples (Barik et al., 2019).

7.4 Code-mixing lexical patterns

When communicating in code-mixed languages, people maintain a lexical pattern unique to each

Non-Standard word type	Example
Non-standard spelling	Prends(friends), plz(please)
word, numbers, special characters mixing	2morrow(tomorrow), 3wheel (Sinhala language meaning trishow in English)
Word exaggeration	gooood(good), woooow(woow), hellooo(hello)
Abbreviated words	TC(take care), tkt(ticket)

Table 3: Types of non-standard words

Pattern type	Sinhala Example	English translation
Present tense	'act krnawa', 'Drive karanawa'	'Acting', 'Driving'
Past tense	'act kara', 'Drive kara'	'Acted', 'Drove'
Indefinite article	'voice ekak', 'chapter ekak'	'A Voice', 'A Chapter'
Definite article	'voice eka', 'chapter eka'	'The Voice', 'The Chapter'
Suffixes	'studentsla', 'teacherla'	Plural form of 'student' and 'teacher'

Table 4: Types of code-mix lexical patterns in Sinhala-English

code-mixed language pair. This is not something that was agreed on formally, but something that could be identified when analyzing the code-mixed language patterns. For example, in Sinhala-English code-mixed language, the word "eka" would be used after most English words. Like "Computer eka" and "Vehicle eka" (Smith and Thayasivam, 2019). Table 4 describes more lexical patterns of Sinhala-English code-mixing. Although these patterns are unique to each language, some of the patterns remain unsolved when analysing the recent study on the domain.

7.5 Compatible evaluation metrics

Traditional machine translation metrics, such as BLEU, METEOR, and TER, are commonly used for compatibility; however, these metrics often fail or are insufficient for handling code-mixed translation and transliteration. The primary limitation of existing evaluation metrics is their inability to handle multiple valid translation outputs. When translating the embedding language to matrix language, the translation could have used an accurate synonym that matches the reference text. However, existing metrics lack the ability to understand contextual meaning.

7.6 Pre-processing and Language Identification Issues

Preprocessing is considered an important step, as code-mixed data tends to be noisier compared to standard text data. When code-mixing is involved with Romanized text, it becomes challenging to perform certain preprocessing tasks, such as spelling correction. A simple spelling correction system would not be able to succeed in a Romanized code-mixed setting since an 'incorrect' token may belong to the other mixed language. Applying a normal spelling correction model risks introducing further errors than normalizing them. Hence, it creates the need for a context-aware spelling correction system.

In this review, it has been identified that some proposed systems have implemented a Language Identification model in the pre-processing pipeline to address ambiguity. But state-of-the-art Language Identification models could only address word-level language identification. Sub-word level language identification is needed to address code-mixed words like 'studentsla'(plural form of Student), where 'Student' is English and 'la' is Sinhala. These challenges need to be addressed, as even small misclassifications propagate into major quality degradation.

8 Conclusion

This review paper has provided a comprehensive analysis of the current advancements, datasets, evaluation methods, and challenges in Machine Translation and transliteration techniques, with a specific focus on code-mixed Indo-Aryan languages. These studies are important to ensure effective communication across different code-mixed indo-aryan languages. In this review, it is evident that the code-mix translation and transliteration accuracies have significantly improved when combined with recent discoveries in the domain of NLP. This marks a promising direction for addressing future research gaps and producing products that solve real-world problems.

Limitation

This review contains several limitations that should be acknowledged. The review primarily focuses on literature published between 2018 and 2025. This ensures that recent advancements are reviewed, but it may have excluded earlier foundational studies. A few studies were excluded due to accessibility issues. Finally, this study focuses on academic studies rather than doing a systematic analysis of industrial or applied systems, which could offer additional insights into the practical difficulties of dealing with code-mixed Indo-Aryan text.

Acknowledgments

This paper has been conducted in compliance with the ethical standards of the Informatics Institute of Technology, Sri Lanka. Generative AI tools were used solely to enhance the clarity and readability of the manuscript.

References

S. Alam and 1 others. 2025. Bnsentmix: A diverse bengali-english code-mixed dataset for sentiment analysis. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 68–77, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dhiraj Amin, Sharvari Govilkar, Sagar Kulkarni, Yash Shashikant Lalit, Arshi Ajaz Khwaja, Daries Xavier, and Sahil Girijashankar Gupta. 2023. Marathi-english code-mixed text generation. *arXiv preprint arXiv:2309.16202*.

Maneesha U Athukorala and Deshan K Sumanathilaka. 2024. Swa bhasha: Message-based singlish to sinhala transliteration. *arXiv preprint arXiv:2404.13350*.

S. Banerjee and A. Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

A.M. Barik, R. Mahendra, and M. Adriani. 2019. Normalization of indonesian-english code-mixed twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

R.S. Bhowmick and 1 others. 2023. Improving indic code-mixed to monolingual translation using mixed script augmentation, generation & transfer learning. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

A.D. De Silva. 2021. Singlish to sinhala converter using machine learning. Master's thesis, University of Colombo School of Computing. [Accessed 25 August 2025].

Sachithya Dharmasiri and TGDK Sumanathilaka. 2024. Swa bhasha 2.0: Addressing ambiguities in romanized sinhala to native sinhala transliteration using neural machine translation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 241–246. IEEE.

A. Gahoi and 1 others. 2022. Gui at mixmt 2022: English-hinglish: An mt approach for translation of code mixed data. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1126–1130, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nada Gohider and Otman A. Basir. 2024. Recent advancements in automatic disordered speech recognition: A survey paper. *Natural Language Processing Journal*, 9:100110.

Ayushman Gupta, Akhil Bhogal, and Kripabandhu Ghosh. 2024. Multilingual controlled generation and gold-standard-agnostic evaluation of code-mixed sentences. *Preprint*, arXiv:2410.10580.

HM Anuja Dilrukshi Herath and TG Deshan K Sumanathilaka. 2024. Tamzhi: Shorthand romanized tamil to tamil reverse transliteration using novel hybrid approach. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 17(1).

A.F. Hidayatullah and 1 others. 2022a. A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development. *IEEE Access*, 10:122812–122831.

Ahmad Fathan Hidayatullah, Atika Qazi, Daphne Teck Ching Lai, and Rosyzie Anna Apong. 2022b. A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development. *IEEE Access*, 10:122812–122831.

O. Iakovenko and T. Hain. 2024. Methods of automatic matrix language determination for code-switched speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5791–5800, Miami, Florida, USA. Association for Computational Linguistics.

I. Jadhav and 1 others. 2022. Code-mixed hinglish to english language translation framework. In *2022 International Conference on Sustainable Computing*

and Data Communication Systems (ICSCDS), pages 684–688.

A. Kugathasan and S. Sumathipala. 2022. Neural machine translation for sinhala-english code-mixed text. *International Journal on Advances in ICT for Emerging Regions (ICTer)*, 15(3):60–71.

A. Kunchukuttan, P. Mehta, and P. Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

C. Myers-Scotton. 1997. *Duelling Languages: Grammatical Structure in Codeswitching*. Clarendon Press.

A.R. Nair and D. Gupta. 2024. Evaluating performance and accuracy of large language models in translating code-mixed hindi to english: A comparative study. In *2024 IEEE 21st India Council International Conference (INDICON)*, pages 1–6.

R. Nayak and R. Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Li Nguyen, Christopher Bryant, Oliver Mayeux, and Zheng Yuan. 2023. How effective is machine translation on low-resource code-switching? a case study comparing human and automatic metrics. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14186–14195, Toronto, Canada. Association for Computational Linguistics.

S. Pal and M. Zampieri. 2020. Neural machine translation for similar languages: The case of indo-aryan languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 424–429. Association for Computational Linguistics.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

D. Patel and R. Parikh. 2020. Language identification and translation of english and gujarati code-mixed data. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–4.

Sandun Sameera Perera, Lahiru Prabhath Jayakodi, Deshan Koshala Sumanathilaka, and Isuri Anuradha. 2025. Indonlp 2025 shared task: Romanized sinhala to sinhala reverse transliteration using bert. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 135–140.

Sandun Sameera Perera and Deshan Sumanathilaka. 2025a. Evaluating transliteration ambiguity in adhoc romanized sinhala: A dataset for transliteration disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2025)*.

Sandun Sameera Perera and Deshan Koshala Sumanathilaka. 2025b. Machine translation and transliteration for indo-aryan languages: A systematic review. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 11–21.

M. Popovi. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

B. Roark and 1 others. 2020. Processing south asian languages written in the latin script: the dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

S.M.M.R.J. Senanayaka, A.W.A.D.N.D. Abeysekara, and M.G.N. Premadasa. 2024. Singrag: A translation-augmented framework for code-mixed singlish processing. In *2024 9th International Conference on Information Technology Research (ICITR)*, pages 1–6.

C. Senaratne. 2009. *Sinhala-English code-mixing in Sri Lanka: A sociolinguistic study*. LOT Publications.

R. Sheth, H. Beniwal, and M. Singh. 2025. Comilingua: Expert annotated large-scale dataset for multitask nlp in hindi-english code-mixing. ArXiv preprint.

G.M.S. Shibli and 1 others. 2023. Automatic back transliteration of romanized bengali (banglish) to bengali. *Iran Journal of Computer Science*, 6(1):69–80.

I. Smith and U. Thayasivam. 2019. Language detection in sinhala-english code-mixed data. In *2019 International Conference on Asian Language Processing (IALP)*, pages 228–233.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

V. Srivastava and M. Singh. 2020. Phinc: A parallel hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*,

pages 41–49. Association for Computational Linguistics.

V. Srivastava and M. Singh. 2021. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208. Association for Computational Linguistics.

Deshan Sumanathilaka, Isuri Anuradha, Ruvan Weerasinghe, Nicholas Micallef, and Julian Hough. 2025a. Indonlp 2025: Shared task on real-time reverse transliteration for romanized indo-aryan languages. *arXiv preprint arXiv:2501.05816*.

Deshan Sumanathilaka, Nicholas Micallef, and Ruvan Weerasinghe. 2024. Swa-bhasha dataset: Romanized sinhala to sinhala adhoc transliteration corpus. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 189–194.

Deshan Sumanathilaka, Sameera Perera, Sachithya Dharmasiri, Maneesha Athukorala, Anuja Dilrukshi Herath, Rukshan Dias, Pasindu Gamage, Ruvan Weerasinghe, and YHPP Priyadarshana. 2025b. Swa-bhasha resource hub: Romanized sinhala to sinhala transliteration systems and data resources. *arXiv preprint arXiv:2507.09245*.

TGDK Sumanathilaka, Ruvan Weerasinghe, and YHPP Priyadarshana. 2023. Swa-bhasha: Romanized sinhala to sinhala reverse transliteration using a hybrid approach. In *2023 3rd International Conference on Advanced Research in Computing (ICARC)*, pages 136–141. IEEE.

S Thara and Prabaharan Poornachandran. 2018. Code-mixing: A brief survey. In *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2382–2388.

D. K. Uthpala and S. Thirukumaran. 2024. Sinhala-english code-mixed language dataset with sentiment annotation. In *2024 4th International Conference on Advanced Research in Computing (ICARC)*, pages 184–188.

Aditya Vavre, Abhirut Gupta, and Sunita Sarawagi. 2022. Adapting multilingual models for code-mixed translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7133–7141, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad Wisal, Abbas Mustafa, and Umair Arshad. 2022. Cmrutu: Code mixed roman urdu (roman urdu and english) to urdu translator. In *2022 24th International Multitopic Conference (INMIC)*, pages 1–5.