Natural Language Processing Journal 11 (2025) 100158

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

Advances in machine transliteration methods, limitations, challenges, applications and future directions

A'la Syauqi [a,b,*], Aji Prasetya Wibawa [b]

[a] Department of Electrical Engineering and Informatics, Faculty of Engineering, Universitas Negeri Malang, Jl. Semarang no. 5, Malang 65145, Indonesia
[b] Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang 65144, Indonesia

ARTICLE INFO

Keywords:
Systematic literature review
Machine transliteration
Transliteration methods
Natural language processing
Script conversion
Low-resource languages
Cross-linguistic communication

ABSTRACT

Machine transliteration is critical in natural language processing (NLP), facilitating script conversion while preserving phonetic integrity across diverse languages. Using the PRISMA framework, this review analyzes 73 selected studies on machine transliteration, covering both methodological advancements and its role in NLP applications. Among these, 37 studies focus on transliteration methods (rule-based, statistical, machine learning, hybrid, and semantic), while 32 studies explore their application in NLP tasks such as machine translation, sentiment analysis, and text normalization. Rule-based methods provide structured frameworks but face challenges in adapting to linguistic variability. Statistical techniques demonstrate robustness yet depend heavily on the availability of parallel corpora. Machine learning models leverage neural architectures to achieve high accuracy but are constrained by data scarcity for low-resource languages. Hybrid approaches integrate multiple methodologies, while semantic knowledge-based models enhance accuracy by incorporating linguistic features. The review highlights transliteration's role in NLP applications such as machine translation, sentiment analysis, and text normalization, which are critical for improving multilingual language accessibility. Findings show that machine learning-based approaches dominate transliteration research (32 of 73 studies), followed by rule-based and hybrid methods. These approaches contribute to improving multilingual accessibility and NLP performance. This study provides actionable insights for researchers and practitioners by synthesizing advancements and identifying challenges. These insights enable the development more efficient and inclusive transliteration systems, ultimately supporting linguistic diversity and advancing multilingual NLP technologies. The review identifies gaps in addressing underrepresented languages like Javanese, where complex character sets, orthographic rules, and scriptio continua remain underexplored.

1. Introduction

Machine transliteration, converting text from one script to another while preserving phonetic properties, is a significant component of multilingual information retrieval and cross-linguistic communication. The field has evolved substantially, incorporating a range of methodologies from early rule-based systems to advanced machine-learning techniques. Seminal works have laid the foundation for understanding the progression and diversification of transliteration methods (Karimi et al., 2011; Knight and Graehl, 1998). This systematic literature review (SLR) aims to examine the various methods of machine transliteration, highlighting the limitations and challenges associated with each approach.

Several notable studies have addressed different aspects of machine transliteration. For instance, a comprehensive survey on machine transliteration systems discussing grapheme-based, phoneme-based, hybrid approaches, and correspondence-based models was provided. Various transliteration systems were also covered, including rule-based, n-gram approach, Support Vector Machines (SVM), statistical methods, and hybrid systems (Sen and Garg, 2015). However, the study primarily focused on specific languages or scripts and did not address broader limitations and challenges.

A review of machine transliteration, translation, evaluation metrics, and datasets, specifically within Indian languages, was conducted to address challenges and future work. While this study provided valuable insights into the field, it was limited to Indian languages and did not explore transliteration methods applicable to other language families (Jha, 2023). Similarly, a survey was conducted on machine transliteration and transliterated text retrieval, discussing approaches such

* Corresponding author at: Informatics Engineering, Faculty of Science and Technology, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Malang 65144, Indonesia.
E-mail addresses: ala.syauqi.2405349@students.um.ac.id, syauqi@ti.uin-malang.ac.id (A. Syauqi), aji.prasetya.ft@um.ac.id (A.P. Wibawa).

https://doi.org/10.1016/j.nlp.2025.100158
Received 19 December 2024; Received in revised form 21 May 2025; Accepted 28 May 2025

as syllable-based, phoneme-based, grapheme-based, hybrid, and combined methods (Prabhakar and Pal, 2018). Despite offering a broader perspective that included Indian and non-Indian language categories, the primary focus remained on Indian languages.

The scope of machine transliteration was extended by reviewing approaches across various languages, including Arabic, Japanese, Chinese, Farsi, and Slavic. This study also addressed the unique challenges associated with transliteration for each language group (Mammadzada, 2023). However, the discussion was limited to the mentioned languages and did not encompass diverse transliteration challenges.

While previous reviews primarily focus on specific languages, transliteration models, or evaluation metrics, this SLR provides a more extensive synthesis by examining multiple transliteration approaches and their applications across various NLP tasks. In addition to evaluating existing transliteration techniques, this study explores research gaps, particularly in low-resource languages, script variations, and the integration of transliteration in multilingual NLP systems. By offering a broader and more up-to-date perspective, this review is a critical reference for researchers and practitioners seeking to develop more effective and inclusive transliteration models.

This review seeks to build on the existing body of knowledge by providing a more inclusive overview of machine transliteration methods, their applications in NLP studies, and the resulting outcomes with the research questions (RQ) defined as follows:

RQ1: How are current methods of machine transliteration implemented across various languages?
RQ2: What are the current limitations and challenges associated with machine transliteration methods across various languages?
RQ3: How is transliteration utilized in NLP studies?
RQ4: What impact does transliteration have on the effectiveness and outcomes of these NLP studies?
RQ5: What are the future directions and potential advancements in machine transliteration?

By systematically identifying, assessing, and interpreting all available research evidence, this SLR aims to answer specific research questions and contribute to developing more effective transliteration methodologies. The findings of this review will underscore the critical role of accurate transliteration in enhancing communication across different languages and inform future advancements in the field.

This paper is structured as follows to ensure a clear and systematic discussion. The current section, Section 1 (Introduction), has provided an overview of machine transliteration, highlighted gaps in previous research, and outlined the research questions that guide this study. Section 2 (Research Methods) describes the methodology used in this SLR, adhering to the PRISMA framework. It outlines the research protocol to ensure transparency and rigor, including study selection criteria, data extraction, and synthesis methods. Section 3 (Research Results) examines research trends in transliteration, focusing on the distribution of studies across languages and scripts, key publication trends from 2014 to 2024, and a comparative analysis of transliteration methods. Section 4 (Discussion) is divided into two main subsections to address the research questions systematically. Section 4.1 (Machine Transliteration Methods, Limitations, and Challenges) directly addresses RQ1 and RQ2, providing an in-depth discussion of various transliteration methods and their associated challenges. Section 4.2 (Transliteration in Various NLP Studies and Their Impacts) responds to RQ3 and RQ4, discussing the applications of transliteration in different NLP tasks and evaluating its impact on model performance. Finally, Section 5 (Conclusion and Future Direction) addresses RQ5, summarizing key findings and outlining potential advancements in transliteration research.

2. Research methods

Conducting a systematic review requires robust and transparent methodologies to ensure the integrity and reliability of the study. The PRISMA framework provides standardized guidelines to enhance the clarity and completeness of systematic reviews and meta-analyses (Liberati, 2009). This systematic review, conducted from August 2023 to December 2024, follows PRISMA guidelines to guarantee a methodologically rigorous and transparent examination of the pertinent literature.

Given the significance of transliteration in languages with different writing systems, this study considers research addressing transliteration across diverse linguistic contexts. Studies focusing on transliteration challenges in non-Latin scripts, such as Arabic, Chinese, Japanese, and Indic scripts, were included to ensure a broad representation of transliteration approaches. However, only studies published in English were reviewed to maintain consistency and avoid potential biases introduced by machine translation. This decision ensures uniformity in data interpretation while acknowledging that transliteration challenges in some linguistic contexts may require further exploration.

This review focuses solely on peer-reviewed journal articles indexed in Scopus, ensuring that all included studies meet high scientific standards. Conference proceedings, preprints, and workshop papers were excluded to maintain methodological rigor. Each selected study underwent an assessment of its research design, dataset quality, and experimental validation to confirm its relevance and credibility. The assessment of study quality was based on research design, sample size, reproducibility of experiments, and methodological transparency, ensuring that only robust studies contributed to the review.

The study selection process followed a systematic approach based on predefined inclusion and exclusion criteria. Two independent reviewers evaluated each study, and disagreements were resolved through discussion until a consensus was reached. By applying the PRISMA framework, the selection process remained transparent and reproducible.

Formulating a meticulously designed research plan is essential, as it provides a comprehensive framework that directs the entire study from the initial data collection stages to analysis and interpretation. The preliminary phases involve the careful development of research questions and the establishment of inclusion and exclusion criteria. Extensive database searches are then performed, followed by thorough screening and detailed assessment stages to maintain high standards of relevance and methodological rigor. This systematic approach is critical for synthesizing literature that accurately addresses the research objectives. In preparing this systematic literature review, the stages were also influenced by Unterkalmsteiner et al. (2012), Radjenović et al. (2013). As shown in Fig. 1, the SLR process is delineated into three primary stages: planning, conducting, and reporting.

In the planning stage, the initial step involves identifying the need for a systematic review (Step 1). This step is crucial as it establishes the rationale and objectives for the review. Following this, a review protocol is developed (Step 2), which serves as a comprehensive plan outlining the methodology for the review. The protocol includes formulating research questions, search strategies, inclusion and exclusion criteria, quality assessment measures, and data extraction and synthesis procedures. Subsequently, the review protocol is evaluated (Step 3) to ensure its robustness and to minimize potential biases.

The conducting stage begins with a systematic search for primary studies (Step 4). This process involves identifying relevant literature through predefined search strategies. Once the studies are gathered, the next step is to select the primary studies (Step 5) based on the inclusion and exclusion criteria specified in the review protocol. Following selection, data is extracted from the primary studies (Step 6), which involves collecting pertinent information that addresses the research questions. The quality of the primary studies is then assessed (Step 7) to ensure

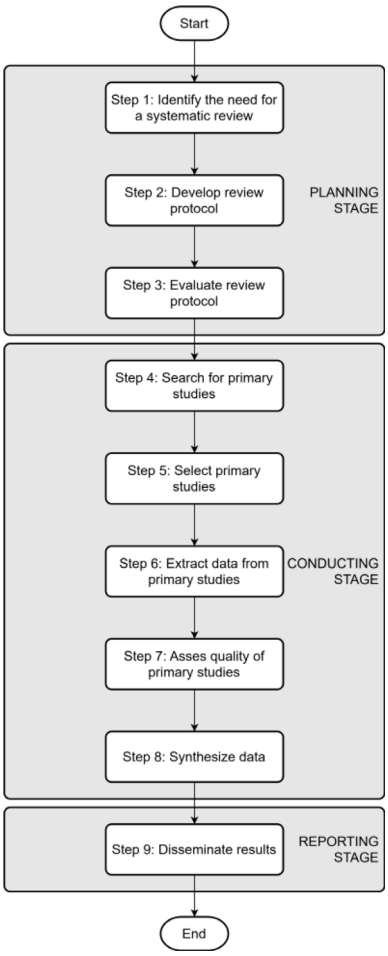

Fig. 1. Systematic literature review steps.

the reliability and validity of the findings. Finally, the extracted data is synthesized (Step 8) to understand the research topic comprehensively.

In the reporting stage, the synthesized results are disseminated (Step 9). This step involves compiling the findings into a coherent report that communicates the outcomes of the systematic review to the broader research community. The reporting stage ensures that the insights gained from the review are accessible and can inform future research and practice.

2.1. Review protocol

The review protocol is an essential component of the SLR, ensuring a structured and unbiased approach to the review process. This protocol is meticulously designed to guide the execution of the review and minimize researcher bias, thus enhancing the transparency and reproducibility of the study. It incorporates several key elements: defining research questions, outlining the search strategy, detailing the study selection process with specific inclusion and exclusion criteria, conducting quality assessments, and finally, the data extraction and synthesis process. These elements are elaborated upon in the following sections and illustrated in Fig. 2.

A. Research Question

The formulation of clear and precise research questions constitutes the foundation of the review protocol, as they define the scope and direction of this study. A well-structured set of questions ensures a systematic and objective investigation, guiding the selection, analysis, and synthesis of relevant literature. This systematic review examines how machine transliteration methods are applied across different languages, identifies their challenges and limitations, explores their role in NLP applications, and assesses their overall impact on research outcomes. Furthermore, this review considers potential advancements and future directions in the field. A detailed outline of these research questions is provided in the Introduction section, which serves as the fundamental basis for this study.

B. Search Strategy

The search strategy is meticulously defined to identify relevant studies from the existing literature, ensuring a comprehensive and unbiased selection process. The strategy involves carefully selecting appropriate and comprehensive databases, with Scopus chosen due to its extensive coverage and high-quality academic resources. Scopus is particularly esteemed for its extensive collection of peer-reviewed journals, conference proceedings, and other scholarly materials, making it an ideal source for identifying pertinent studies.

The development of effective search strings is a critical component of this strategy. Keywords such as ‘Transliteration’ are meticulously selected to retrieve relevant studies comprehensively. The keyword ‘Transliteration’ was used to retrieve preliminary studies. A total of 2.505 preliminary studies were retrieved. These search strings are crafted to capture a wide array of research articles pertinent to the topic, thus maximizing the inclusivity and relevance of the search results.

By adhering to this detailed and structured search strategy, the systematic literature review aims to compile a robust body of evidence, establishing a solid foundation for drawing accurate and meaningful conclusions about research in the transliteration field.

C. Study Selection Process

As illustrated in Fig. 1, the study selection process involves a series of meticulously planned stages designed to ensure a comprehensive and unbiased review. Initially, this process begins with identifying and retrieving relevant studies from selected databases. Following this, these studies are thoroughly filtered based on specific criteria, such as publication year, document type, language, and relevant keywords. Subsequent steps involve the removal of duplicate studies to prevent redundancy. Finally, an in-depth assessment of the remaining studies’ titles and abstracts is performed to ensure they align with the research questions. Each stage is crucial for compiling a robust body of evidence that will form the foundation for the systematic literature review.

(1) Filtering Studies

The filtering process is critical in the study selection process, ensuring that only the most relevant and high-quality studies are included in the systematic review. This process involves several key criteria:

a. By Year: Studies are filtered based on their publication year. Only studies published between 2014 and 2024 are considered, with those outside this range automatically excluded. From this filtering, a total of 1498 studies were obtained. This criterion ensures that the review encompasses the most recent and relevant research, reflecting the current state of knowledge in the field.
b. By Document Type: Only articles are included in the review to maintain a high standard of academic rigor. Non-article documents such as reviews, conference papers, and other publications are automatically excluded. This approach resulted in identifying 687 papers. This focus on peer-reviewed articles ensures that the included studies have undergone rigorous scrutiny by experts in the field.
c. By Language: The language of publication is another important filtering criterion. Only studies published in English are included, with non-English studies automatically excluded. As a result of this filtering, 499 papers were selected. This criterion ensures that the included studies are accessible to a wider audience and that the review can be conducted without language barriers.

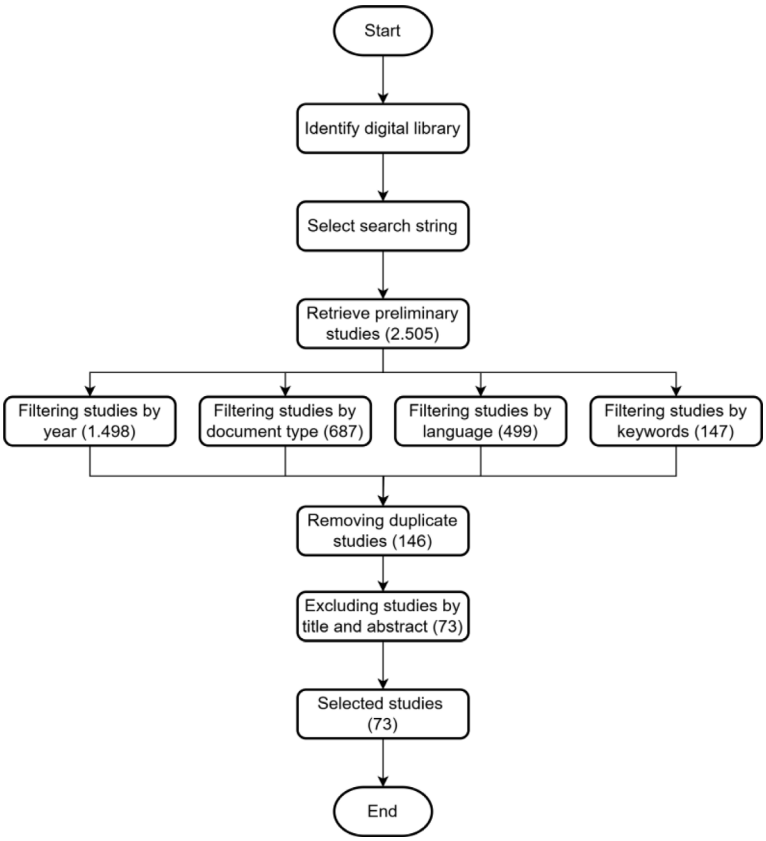

Fig. 2. Studies search and identification.

d. By Keywords: Keywords play a crucial role in identifying relevant studies. Studies containing specific keywords such as "Transliteration", "Machine Transliteration", "Transliteration Models", "Transliteration System", "Sanskrit", "Balinese Script". "Sanskrit Manuscripts", "Phonetic", "Transcription", and "Latin Script" are automatically included based on these keywords. Through this approach, 147 papers were selected. This approach ensures comprehensive coverage of the topic and captures various relevant studies.

By thoroughly applying these filtering criteria, the systematic review process ensures the inclusion of relevant and high-quality studies, providing a solid foundation for a comprehensive and unbiased literature review.

(2) Removing Duplicate Studies

Removing duplicate studies is an essential step in ensuring the accuracy and reliability of the systematic review. Duplicate studies can arise when multiple records of the same study are retrieved from different sources or databases. These duplicates can lead to redundant data, which can skew the results and analyses of the review.

A thorough and systematic approach is used to identify and eliminate duplicates to address this. Initially, a web-based application named Covidence (10) detects duplicate entries based on metadata such as titles, authors, publication years, and abstracts. This automated process significantly reduces the initial volume of duplicates, resulting in a refined set of 146 studies. The method helps ensure that unique and relevant studies are included in the subsequent stages of the review.

(3) Inclusion and Exclusion Criteria

The inclusion and exclusion criteria are pivotal in ensuring the relevance and quality of the studies selected for the systematic review. These criteria are meticulously defined to maintain a focused and rigorous review process.

a. Inclusion Criteria

i. Language: Only studies published in English are included. This criterion ensures that the selected studies are accessible to a broader audience and can be reviewed and understood without language barriers.
ii. Terminology: Studies must include specific terms such as "transliteration," "machine transliteration," "transliteration models," "transliteration system," or other related terms in the title, abstract, or keywords. This requirement helps identify studies that are directly relevant to transliteration.
iii. Document type: Only research articles are included. This criterion ensures that the studies selected are peer-reviewed and have undergone rigorous academic scrutiny, thereby maintaining the quality and reliability of the data.

b. Exclusion Criteria

i. Subject Area: Studies with subject areas not related to computer science are excluded. This criterion ensures that the review focuses specifically on the application and advancements in transliteration within the field of computer science.
ii. Language: Non-English studies are excluded. Excluding non-English studies helps to avoid potential translation errors and ensures consistency in the review process.

By applying these inclusion and exclusion criteria, the systematic review ensures that only the most relevant, high-quality studies are included. Applying these criteria resulted in the selection of 73 studies, providing a solid foundation for a comprehensive analysis of the current state of transliteration research.

D. Data Extraction and Synthesis Process

Data extraction involves the systematic collection of relevant information from the selected studies. This step is crucial as it forms the foundation for the synthesis and analysis phases of the systematic review. The process includes meticulously gathering the following details from each study:

a. Title: The full title of the study, which provides a summary of the research focus.
b. Journal Name: The name of the journal in which the study was published, ensuring the credibility and context of the publication.
c. Publication Year: The year the study was published helps assess the timeliness and relevance of the research.
d. Language/Script Studied: The specific language or script that is the subject of the study.
e. Research Objective: The primary aims and goals of the study, highlighting what the researchers intended to investigate or achieve.
f. Research Gap: The specific gaps in existing knowledge that the study aims to address, providing context for its contribution to the field.
g. Method: The research methodology includes experimental design, data collection techniques, and analytical approaches.
h. Result: The outcomes and findings of the study detailing the data and evidence collected.
i. Finding: The key discoveries or insights derived from the results, often including new knowledge or understanding generated by the research.
j. Strength: The strengths and advantages of the study, such as robust methodologies, significant findings, or unique contributions to the field.
k. Limitation: The limitations or weaknesses of the study, including potential biases, methodological constraints, or areas where further research is needed.
l. Challenge: The challenges encountered during the research process may include difficulties in data collection, analysis, or other aspects of the study.
m. Conclusion: The conclusions drawn by the researchers summarize the overall findings and implications of the study.

Synthesis involves integrating and analyzing the extracted data to identify patterns, draw meaningful conclusions, and answer the research questions. This step includes:

a. Identifying Patterns: Looking for common themes, trends, and relationships among the findings of different studies.
b. Drawing Conclusions: Based on the identified patterns and the synthesized data, conclusions are made to answer the research questions comprehensively.
c. Providing Recommendations: Suggest future research directions or practical applications based on the synthesis of findings.

This thorough and detailed approach ensures that the systematic review is comprehensive and unbiased and provides a solid foundation for understanding the current state of research in transliteration.

2.2. List of acronyms

A comprehensive list of acronyms is included to enhance the clarity and accessibility of this systematic literature review. It standardizes domain-specific language, ensuring clear and consistent references to key concepts in machine transliteration and NLP. Given the frequent use of specialized expressions across transliteration methods, machine learning models, and NLP applications, this section serves as a resource to support conceptual consistency. The full list, along with definitions, is presented in Table 1.

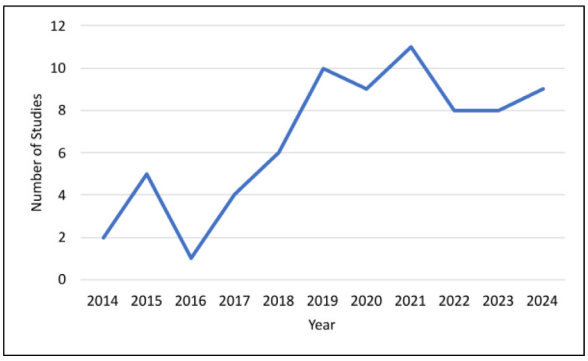

Fig. 3. Distribution of studies.

3. Research results

This section examines research results, including the languages or scripts used in the studies and their countries of origin. Each topic is detailed in the sub-sections to provide a comprehensive view of the development and distribution of transliteration research. The analysis identifies the most frequently studied languages and scripts and explains their countries of origin. Additionally, the types of datasets used in these studies are discussed to highlight their role in enabling the development and testing of transliteration models. This approach aims to emphasize the key contributions and trends in transliteration research and to provide a foundation for future research directions that are more effective and insightful.

In this systematic literature review, 73 research articles were identified and categorized into three main groups: 37 articles discussing transliteration methods or algorithms, 32 articles addressing other NLP research supported by transliteration methods, and four literature review papers. The discussion of transliteration methods is presented in Section 4.1, the application of transliteration in various NLP studies is covered in Section 4.2, and the literature review studies have been discussed in Section 1.

3.1. Trends in machine transliteration studies

The distribution of machine transliteration studies from 2014 to 2024 indicates that these topics have changed. These dynamics are illustrated in Fig. 3, demonstrating that machine transliteration remains a relevant area of research. The analysis encompasses the top six journals that have published the most research papers on transliteration from 2014 to 2024.

As illustrated in Fig. 4, the data highlights these journals' prominence in advancing the transliteration field. The journals identified are the Indian Journal of Science and Technology, Computación *y* Sistemas, Journal of King Saud University - Computer and Information Sciences, International Journal of Advanced Computer Science and Applications (IJACSA), Indonesian Journal of Electrical Engineering and Computer Science, and ACM Transactions on Asian and Low-Resource Language Information Processing. Notably, the ACM Transactions on Asian and Low-Resource Language Information Processing stands out with the most publications, totaling nine papers, underscoring its pivotal role in disseminating advancements in machine transliteration. Following this, the Indonesian Journal of Electrical Engineering and Computer Science and IJACSA contribute significantly to the body of knowledge with four publications each. This substantial contribution highlights these journals' commitment to fostering research in this area. Additionally, the Journal of King Saud University - Computer and Information Sciences, Computación *y* Sistemas, and the Indian Journal of Science and Technology have each published three papers on transliteration,

Table 1
List of acronyms.

Acronyms	Full Form	Definition
BAB	Belajar Aksara Bali	A system for learning Balinese script
BiLSTM	Bidirectional Long Short-Term Memory	A type of neural network for sequence prediction
BLEU	Bilingual Evaluation Understudy	An evaluation metric for machine translation quality
CNN	Convolutional Neural Network	A deep learning model used in image processing and pattern recognition
CRF	Conditional Random Field	A probabilistic model used for sequence labeling tasks
CRNN	Convolutional Recurrent Neural Network	A hybrid deep learning model that combines convolutional layers for spatial feature extraction and recurrent layers for sequential data modeling is used in tasks such as handwritten transliteration.
DNN	Deep Neural Networks	A class of artificial neural networks with multiple layers is used to model complex patterns in data for tasks such as classification and prediction.
GRU	Gated Recurrent Units	A type of RNN architecture that captures long-range dependencies with fewer parameters than LSTM, making it more computationally efficient.
HLSTM	Hierarchical Long Short-Term Memory	An advanced model combining an LSTM for hierarchical data processing
HMM	Hidden Markov Model	A statistical model used in sequential data prediction
HNTSumm	Hybrid Neural Text Summarization	A model combining neural networks for summarizing text
HOG	Histogram of Oriented Gradients	A feature descriptor used in computer vision and image processing to detect object edges and shapes by counting gradient orientations in localized portions of an image.
HRL	High-Resource Language	Languages with abundant resources for NLP tasks
HTR	Handwritten Text Recognition	The process of converting handwritten text into machine-readable text
ITRANS	Indian Transliteration Standard	A standard system for transliterating Indian languages
LBtrans-Bot	Latin-to-Balinese Script Transliteration Bot	A robotic system designed to transliterate Latin characters into Balinese script, utilizing the Noto Sans Balinese font for accurate script rendering.
LRL	Low-Resource Language	Languages with limited resources for NLP tasks
LSTM	Long Short-Term Memory	A recurrent neural network architecture designed to model long-range dependencies in sequence data.
MLP	Multi-Layer Perceptron	A type of neural network model with multiple layers of processing
MNMT	Multilingual Neural Machine Translation	A machine learning model for translating between multiple languages
MODI	MODI Script	A historical script formerly used for writing Marathi and other languages in western India, often used in administrative and handwritten documents.
MRTL	Multi-Round Transfer Learning	A transfer learning approach that iteratively fine-tunes neural MT models by leveraging transliteration to reduce cross-linguistic character-level differences and enhance knowledge transfer between language families.
MT	Machine Translation	The process of automatically translating text from one language to another
NEM	Neural Embedding Model	A machine learning model that represents words in a vector space
NLP	Natural Language Processing	The field of study focuses on the interaction between computers and human language
NMT	Neural Machine Translation	A deep learning-based model for automatic translation
NOOR-TR	NOOR Transliteration and Reversibility System	A system for transliterating text between languages with different scripts
OCR	Optical Character Recognition	A technology for recognizing text in images or scanned documents
OCRopy	OCR Python	An OCR framework implemented using Python for text recognition
OOV	Out-of-Vocabulary	Refers to words not present in the model's training vocabulary, often causing errors or low accuracy in tasks like translation or transliteration.
PBSMT	Phrase-Based Statistical Machine Translation	A model that uses phrase alignment for machine translation
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	A framework for conducting systematic reviews
RBLatAm	Rule-Based Latin-to-Amharic	A transliteration system that applies user-defined rules to convert Latin characters into Amharic script for multilingual digital communication.
RNN	Recurrent Neural Network	A type of neural network designed to process sequential data by maintaining internal memory of previous inputs.
RuTUT	Roman Urdu to Urdu Translator	A rule-based transliteration system that converts Roman Urdu into Urdu script using character substitution and Unicode mapping.
SoEo	Social Eagle Optimization Algorithm	A metaheuristic optimization algorithm enhances model learning performance, integrated into deep BiLSTM networks for improved sentiment classification in multilingual transliteration contexts.
SLR	Systematic Literature Review	A method of reviewing and synthesizing research studies
SMS	Short Message Service	A text messaging service component of mobile communication systems, often characterized by informal and non-standard language use in transliteration research.
SMT	Statistical Machine Translation	A machine translation approach based on statistical models derived from bilingual text corpora.
SNN	Siamese Neural Network	A type of neural network used for comparing and identifying similarities
SVM	Support Vector Machine	A supervised learning model for classification tasks
TAB	Transliteracy Aksara Bali	A transliteration system for Balinese script
TTS	Text-to-Speech	Technology for converting written text into spoken words
WFST	Weighted Finite State Transducer	A probabilistic model used in transliteration to map input sequences to output sequences based on weighted transitions.

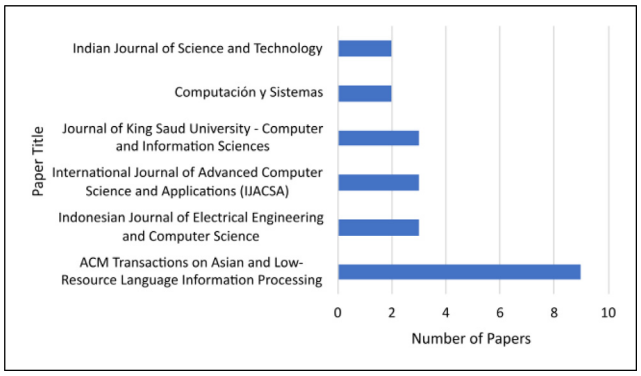

Fig. 4. Prominent journals in selected studies.

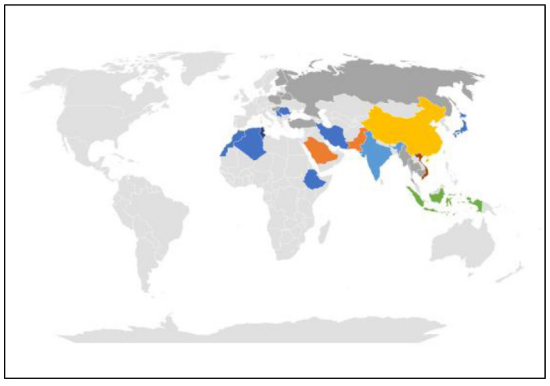

Fig. 5. Language/script origin in studies.

further emphasizing the scholarly interest and concentrated efforts in this domain.

Analyzing these publication trends provides valuable insights into the research landscape and identifies key journals that serve as essential platforms for disseminating knowledge and advancements in machine transliteration.

3.2. Language and script

The analysis of machine transliteration studies from 2014 to 2024 reveals significant trends in the focus on various languages and scripts. As shown in Table 2, the data indicate that Indian languages/scripts dominate the field with a substantial 50 studies, reflecting significant interest and possibly the complexity and diversity of scripts in the Indian subcontinent. Following India, Indonesian languages/scripts have been the focus of 10 studies, highlighting the growing interest in the transliteration challenges posed by the diverse linguistic landscape of Indonesia. Chinese languages/scripts are the subject of 8 studies, underscoring the importance of transliteration in a language with a unique logographic writing system. With six studies, Tunisian Arabic highlights the research interest in Arabic script transliteration, reflected in the five studies dedicated to Saudi Arabian Arabic. These findings suggest a broader regional focus on Arabic script transliteration. Lastly, Vietnamese languages/scripts have been explored in 4 studies, pointing to the transliteration challenges associated with its Latin-based script and tonal nature. This distribution of studies underscores the global interest in addressing the transliteration needs of diverse languages and scripts, with a notable emphasis on regions with complex linguistic and script systems.

The geographical distribution of the languages/scripts studied in the papers is visually represented, as illustrated in Fig. 5. This visual representation complements the data in Table 2, emphasizing the global

Table 2
Distribution of studies by language/script origin.

Language/script's country of origin	Number of studies
India	50
Indonesia	10
China	8
Tunisia	6
Arab Saudi	5
Vietnam	4
Pakistan	3
Algeria	2
Ethiopia	2
Iran	2
Japan	2
Korea	2
Morocco	2
Romania	2
Russia	2
Serbia	2
Rusia	2
Turkey	2
Belarus	1
Cambodia	1
English	1
Finlandia	1
French	1
Hungaria	1
Kurdish	1
Lithuania	1
Myanmar	1
Polandia	1
Thailand	1

scope of transliteration research and the significant focus on regions with complex linguistic and script systems. The map serves as a visual aid to better understand the regional emphasis and the diverse linguistic landscapes that are the subject of transliteration studies. Based on the languages/scripts studied in the papers, Table 3 shows the distribution of research interest across various languages and scripts. Hindi is the most researched language, with 12 studies dedicated to it. The next most studied languages are Arabic, with ten studies, and Balinese, with eight, indicating their prominence in transliteration research. Both Arabizi and Bengali have five studies each, underscoring their importance. These languages represent the top five most studied in the context of transliteration, reflecting the academic focus and the complexities involved in their transliteration processes. The distribution of research studies by language/script is depicted in Fig. 6. Hindi is the most frequently studied language/script, accounting for 22% of the research. It is followed by Arabic at 19%, Balinese at 15%, and Arabizi, Bengali, Punjabi, and Tamil, each representing 9%. Chinese constitute 8% of the research. Due to their extensive use and varied applications, this distribution underscores the significant interest in Hindi and Arabic transliteration. Including Balinese, Arabizi, Bengali, Punjabi, and Tamil signifies a growing interest in less commonly studied scripts, highlighting the expanding scope of machine transliteration research.

Notably, the data also indicate that the Javanese script has been the subject of only one study, suggesting that it remains an underexplored area with significant potential for further research. This disparity in the number of studies highlights the need for more focused investigations into lesser-studied scripts like Javanese to enrich knowledge in machine transliteration. By addressing these gaps, future research can contribute to developing more comprehensive and effective transliteration models, facilitating better cross-linguistic communication and information access.

3.3. Comparative analysis of transliteration methods

Transliteration research is heavily influenced by linguistic complexity, script diversity, and technological challenges associated with

Table 3
Distribution of studies by language/script.

Region	Language/script	Number of studies	Language/script	Number of studies
South Asia	Hindi	12	Bengali	5
	Punjabi	5	Tamil	5
	Devanagari	4	Gurmukhi	3
	Urdu	3	Assamese	2
	Gujarati	2	Kannada	2
	Marathi	3	Bodo	1
	Devnagari	1	Malayalam	1
	MODI	1	Odia	1
	Sanskrit	1	Telugu	1
Middle East	Arabic	10	Arabizi	5
	Amharic	2	Farsi	1
	Kurdish	1	Ottoman	1
	Persian	1	Turkish	1
East Asia	Chinese	5	Japanese	2
	Korean	2	Cantonese	1
	Uyghur	2		
Southeast Asia	Balinese	8	Javanese	1
	Khmer	1	Myanmar	1
	Sundanese	1	Thai	1
	Vietnamese	2	Nôm	1
	Vietnamese national	1		
Central Asia	Cuneiform	1		
Europe	Russian	3	Cyrillic	2
	Belarusian	1	Finnish	1
	French	1	Hungarian	1
	Kurrentschrift	1	Lithuanian	1
	Polish	1	Romanian	1
	Serbian	1		

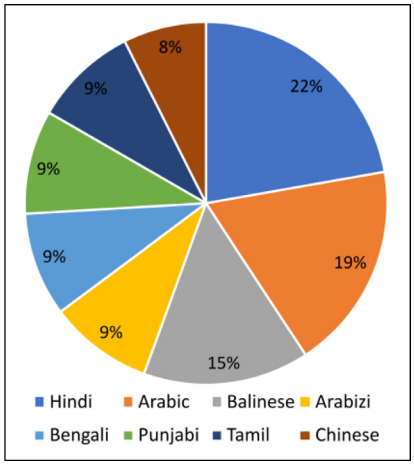

Fig. 6. Distribution of studies by language and script.

different writing systems. The predominance of Indian scripts in transliteration studies suggests that languages with syllabic and abugida scripts pose greater computational challenges, necessitating specialized methods. Similarly, the focus on Chinese and Arabic transliteration highlights the need for models capable of handling logographic structures and diacritic-based orthographies, which often pose challenges for traditional statistical and rule-based methods.

Given these challenges, various transliteration techniques have been developed to address the specific requirements of different language families. Table 4 presents a comparative analysis of these methods, evaluating their effectiveness based on key performance aspects such as accuracy, complexity, efficiency, strengths, and limitations. This comparison provides insights into how different approaches have evolved to accommodate the unique characteristics of each script, revealing trade-offs in interpretability, computational demands, and adaptability.

As shown in Table 4, rule-based approaches offer transparency and interpretability, making them suitable for languages with well-defined phonetic rules. However, their rigidity limits their ability to generalize across diverse linguistic structures, necessitating extensive manual rule crafting. Statistical approaches, such as Statistical Machine Translation (SMT) and Hidden Markov Models (HMM), provide greater flexibility through probabilistic modeling. However, they rely heavily on large parallel corpora, posing challenges in resource-limited scenarios.

Machine learning-based methods, particularly deep learning architectures like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), demonstrate superior accuracy in learning complex transliteration patterns. Nonetheless, their high computational costs and dependence on vast amounts of annotated data hinder scalability, especially for low-resource languages. Although hybrid approaches effectively combine rule-based, statistical, and machine-learning techniques to balance accuracy and efficiency, their increased model complexity often complicates optimization and interpretability.

Meanwhile, semantic knowledge-based methods leverage linguistic context to enhance transliteration accuracy and preserve semantic integrity across transliterated texts. Despite their potential, their effectiveness is constrained by the availability of structured semantic resources, making them less feasible for low-resource languages.

These findings suggest that the selection of a transliteration method should be guided by task-specific requirements, balancing interpretability, computational efficiency, and linguistic adaptability. Future research should explore integrating semantic knowledge with machine learning models to optimize transliteration performance while addressing the limitations of existing approaches. Additionally, evaluating transliteration methods through standardized benchmarks and real-world datasets would further enhance the understanding of their practical applicability.

4. Discussion

4.1. Machine transliteration methods, limitations, and challenges

Based on the analysis of selected studies, transliteration is a complex process influenced by various linguistic factors such as phonetic structures, orthographic conventions, and script differences. Different languages pose unique challenges in transliteration, requiring specialized methods to handle these variations effectively.

Table 4
Comparison of transliteration methods.

Method	Accuracy	Complexity	Efficiency	Strengths	Limitations
Rule-Based	Moderate	High	Low	Clearly defined linguistic rules, transparent and interpretable	Struggles with linguistic variations and requires extensive manual rule creation
Statistical (SMT, HMM)	Moderate-High	Moderate	Moderate	Uses probabilistic models to handle variations, more flexible than rule-based methods	Requires a large parallel corpus for effective training and may suffer from domain-specific biases
Machine Learning (LSTM, GRU)	High	High	Moderate	Learns complex transliteration patterns, generalizes well to unseen data	Computationally expensive, requires extensive labeled data for optimal performance
Hybrid (SVM+HMM)	Very High	Very High	Moderate	Combining the advantages of statistical and machine learning models for enhanced accuracy	High computational cost and requires careful hyperparameter tuning
Semantic Knowledge-Based	High	High	Moderate	Incorporates contextual and linguistic semantics, improving accuracy in ambiguous cases	Relies on structured semantic resources, which may be unavailable for low-resource languages

Table 5 summarizes the key linguistic characteristics of different languages and the primary transliteration methods used to address these challenges. The table presents a comprehensive overview of the various languages and scripts examined in this study. It highlights their key phonetic features, transliteration challenges, and the corresponding methods to address these complexities. Each language exhibits distinct linguistic characteristics that influence the most effective transliteration approach. For instance, Punjabi (Gurmukhi) and Hindi present the challenge of schwa deletion, where certain vowels appear in written text but are omitted in pronunciation. Rule-based and HMM methods are utilized to handle this phenomenon, as they can effectively learn phonetic patterns and dynamically apply transformation rules.

Similarly, Arabic and its Romanized variant, Arabizi, pose significant challenges due to the absence of diacritics, leading to potential ambiguities in transliteration. Weighted Finite State Transducer (WFST) models combined with lexicon-based approaches are implemented to mitigate this issue, allowing for probabilistic reconstruction of missing vowels. The primary challenge for tonal languages such as Vietnamese and Thai is tone preservation, as tonal variations significantly alter meaning. Statistical Machine Translation (SMT) and neural network-based methods have been highly effective in capturing and preserving tonal distinctions.

Furthermore, the transliteration of complex scripts such as Balinese, Javanese, and Nôm encounters difficulties in digital representation and character recognition, particularly due to the intricate ligatures and historical variations in these scripts. Rule-based methods, Optical Character Recognition (OCR), and unicode mapping techniques are widely adopted to overcome these challenges, facilitating accurate conversion from traditional scripts to modern digital formats.

As presented in Table 5, transliteration methods are carefully designed to address the linguistic challenges of different languages. For instance, the probabilistic nature of HMM makes them particularly effective for handling vowel omission in Punjabi and Hindi. Similarly, Weighted Finite State Transducers (WFST) reconstruct missing diacritics in Arabic transliteration by leveraging contextual probability. SMT methods ensure tonal accuracy in Vietnamese, highlighting the role of data-driven approaches in phonetic preservation. Meanwhile, neural network-based models effectively handle morpho-phonetic complexities in languages such as Korean and Japanese, demonstrating the advantage of deep learning in transliteration tasks.

Drawing from the selected studies, the transliteration methods are categorized as shown in Fig. 7. This figure illustrates the diverse approaches to machine transliteration, including statistical methods, rule-based methods, machine learning, and hybrid methods. Each category encompasses techniques such as Conditional Random Fields (CRF), HMM, SVM, and Deep Neural Networks (DNN). Detailed explanations of these methods are provided in the subsequent sub-sections, offering a comprehensive understanding of the development and application of each method in machine transliteration.

A. Rule-Based

Rule-based transliteration methods have been widely applied due to their explicit linguistic rules and structured approach. These methods rely on predefined mappings and phonetic rules to convert one script to another, ensuring consistency and predictability in transliteration output. One of the primary advantages of rule-based methods is their high level of control and transparency, as each rule is explicitly defined based on linguistic principles. This characteristic makes rule-based transliteration particularly effective for languages with well-established phoneme–grapheme correspondences, enabling accurate and reversible script conversion. These methods depend less on large annotated corpora, making them suitable for low-resource languages where training data may be scarce.

As shown in Table 6, rule-based transliteration methods, while providing structured and transparent script conversion, also face notable challenges such as phonetic variability, backward compatibility issues, and the extensive manual effort required to refine linguistic rules. The table compares different rule-based approaches, illustrating their advantages and limitations across multiple languages and scripts. The following section elaborates on these methods, their limitations, and challenges.

Several studies have demonstrated the effectiveness of rule-based transliteration. A system for Punjabi successfully incorporated phonetic rectification techniques to manage the inherent short vowel schwa, utilizing bigram models for improved accuracy (Abbas, 2020). This approach was particularly useful in addressing inconsistencies in pronunciation, which often led to challenges in script conversion. Similarly, a transliteration approach for Ottoman Turkish to Latin script employed character mapping and normalization techniques to enhance phonetic representation (Jaf and Kayhan, 2021). The effectiveness of this method was evident in its ability to accommodate historical

Table 5
Linguistic features and transliteration methods.

Language/Script	Region	Key phonetic features	Challenges in transliteration	Method sed
Amharic	Africa	Ge'ez script, bidirectional text	Phonetic syllabary structure	Rule-based (maps syllables to phonetic equivalents)
Uyghur (Arabic & Latin script)	Central Asia	Phoneme-based script mapping	Adaptation to different writing conventions	Semantic Knowledge-Based Model (maps phonemes across scripts)
Chinese (Simplified & Traditional)	East Asia	Logographic writing system	Character disambiguation	Neural Networks (LSTM/Transformer) (learns contextual character relationships)
Japanese (Kana & Kanji)	East Asia	Mixed phonetic and logographic script	Handling furigana annotations	SMT + Neural Machine Transliteration (NMT) (handles phonetic-logographic transitions)
Korean (Hangul & Hanja)	East Asia	Syllable-based script, phoneme blocks	Consonant assimilation, vowel harmony	Hybrid (Statistical + Neural) (captures phonetic context in syllable-based scripts)
Cyrillic (Russian, Serbian, Belarusian, etc.)	Europe	Phonetic variations	Consistency in character mapping	Rule-based (ensures consistency in phonetic mappings)
Tatar (Cyrillic & Latin)	Europe	Vowel harmony, consonant shifts	Conversion between alphabets	Hybrid (Bigram + Rule-based) (models vowel harmony statistically)
Arabic	Middle East	Diacritic-dependent, consonant-heavy	Lack of vowel markings	WFST + Lexicon (reconstructs missing vowels probabilistically)
Arabizi (Romanized Arabic)	Middle East	Informal orthography, code-mixed text	Non-standard spelling variations	Hybrid (Rule-based + CRF) (handles informal spelling variations)
Kurdish (Latin & Arabic script)	Middle East	Double consonants, vowel inconsistencies	Inconsistent mapping across scripts	Rule-based (applies phonetic transformation rules)
Ottoman Turkish	Middle East	Script mixing (Arabic, Persian, Turkish)	Historical text complexity	OCR + Rule-based (extracts characters and applies mapping rules)
Hindi	South Asia	Schwa deletion, compound consonants	Handling short vowels, multiple transliteration conventions	Rule-based + HMM (models probability of vowel deletion)
Punjabi (Gurmukhi)	South Asia	Schwa deletion, retroflex sounds	Maintaining phonetic integrity	Rule-based + HMM (handles vowel omission dynamically)
Bengali	South Asia	Nasalized vowels, inherent vowels	Disambiguation of homophones	SMT (learns patterns from bilingual corpora)
Tamil	South Asia	Agglutinative morphology	Handling long vowel variations	Rule-based + Neural Network (captures morphological changes)
Roman Urdu (Latin script for Urdu)	South Asia	No standard orthography	Multiple spelling variations	Unicode-based mapping (ensures consistent representation)
Balinese (Aksara Bali)	Southeast Asia	Syllabic script, scriptio continua	Digital representation and preservation	Rule-based (OCR & Unicode mapping) (ensures preservation of traditional characters)
Javanese (Aksara Jawa)	Southeast Asia	Complex ligatures, conjuncts	Lack of digital resources	Rule-based + OCR (recognizes and standardizes ligatures)
Vietnamese (Latin-based with tones)	Southeast Asia	Tonal variations, diacritic dependency	Tone preservation	SMT-based (predicts tonal changes using parallel corpora)
Nôm (Vietnamese ancient script)	Southeast Asia	Ideographic script	Transliteration to modern Vietnamese	SMT-based (maps ancient script to modern equivalents)
Thai	Southeast Asia	Tonal system, complex syllable structures	Maintaining phonetic accuracy	Hybrid (Statistical + Rule-based) (captures syllable and tone structures)

script variations, ensuring that transliterated text remained faithful to its original phonetic form. Another study developed a rule-based system for English and Hindi to Punjabi transliteration, emphasizing phonetic analysis and orthographic rules to enhance accuracy (Kaur and Singh, 2015a). This approach effectively handled the phonetic variability inherent in source and target scripts, ensuring high transliteration accuracy.

A phonetic-based transliteration system was introduced for Balinese script conversion, ensuring accurate representation of phonemes while maintaining compatibility with traditional script conventions (Abebaw et al., 2023). Balinese, a language with a unique and intricate script, poses significant challenges for transliteration. The rule-based approach successfully addressed these challenges by implementing well-defined phonetic conversion rules, ensuring the transliterated text remained readable and accurate. Similarly, the NOOR-TR system proposed a bidirectional transliteration approach for Arabic and English, leveraging diacritic symbols to enhance reversibility and transliteration consistency (Alginahi et al., 2018). This method was particularly beneficial for languages like Arabic, where the absence of diacritics in written text often leads to ambiguities in pronunciation and meaning.

In the transliteration of Balinese palm-leaf manuscripts, knowledge representation and phonological rules were applied to preserve phonetic accuracy (Kesiman et al., 2017). These rules ensured that ancient manuscripts could be digitized and accessible to modern readers while maintaining linguistic and phonetic integrity. A Kurdish transliteration system was also developed to handle double consonants and phonological aspects, ensuring high precision in script conversion (Ahmadi, 2019). As a language with multiple script variants, Kurdish required

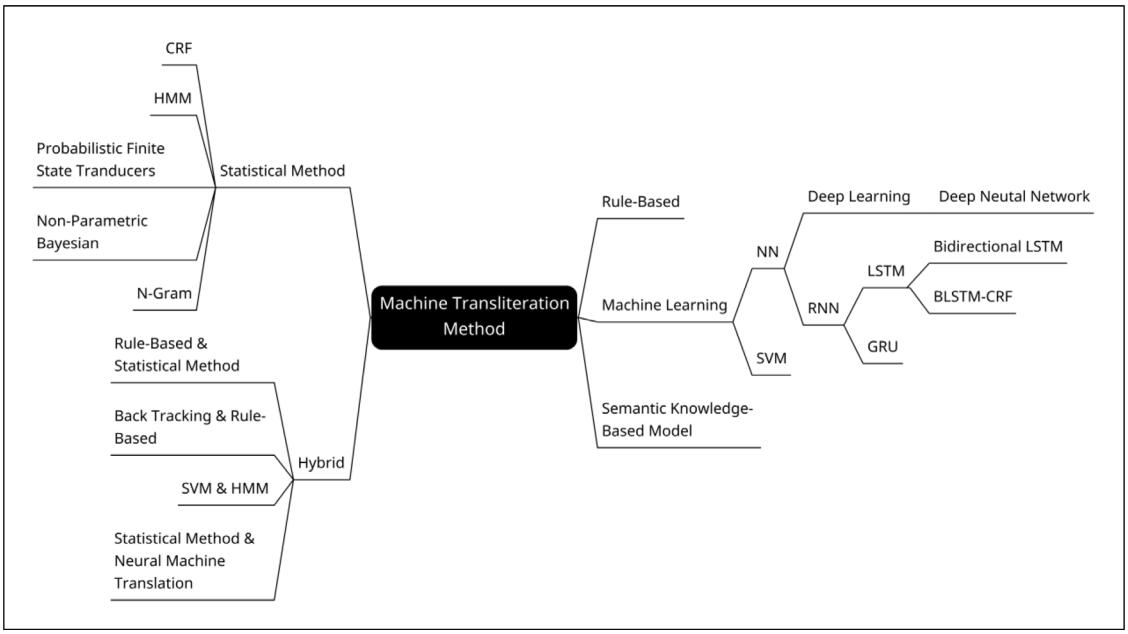

Fig. 7. Transliteration methods.

Table 6
Comparison of rule-based transliteration methods.

Language/Script	Method	Advantages	Disadvantages	Reference
Polish, English–Lithuanian	Rule-based transliteration with exception lists	Provides structured and consistent transliteration	Requires extensive manual rule definition and maintenance due to variations in transliteration rules over time	Kasparaitis (2023)
Balinese	Rule-based transliteration with backward compatibility	Ensures traditional transliteration rules are maintained	Struggles with maintaining compatibility with older transliteration standards	Indrawan (2021)
Ottoman Turkish	Character mapping and normalization	Enhances phonetic representation and preserves historical script variations	Faces challenges due to historical script inconsistencies and character mappings	Jaf and Kayhan (2021)
Balinese	Phonology-based transliteration	Maintains compatibility with traditional script conventions	Struggles with backward compatibility due to inconsistencies between older and modern standards	Indrawan (2020)
Punjabi	Grapheme-based with phonetic rectification	Effectively manages short vowel schwa using Bigram models	Phonetic information is crucial, making rule-based approaches insufficient in certain cases	Kaur and Singh (2015a)
Latin–Amharic	Rule-based transliteration with Unicode mappings	Improves transliteration accuracy for digital communication	Difficulties handling numeric representations and double consonants	Indrawan et al. (2018b)
Kurdish	Structured rule-based character mapping	Ensures high precision in detecting double consonants	Inconsistencies between Arabic-based and Latin-based orthographies, complicating normalization	Ahmadi (2019)

a robust transliteration system that could accurately map phonetic structures across different scripts.

Further research focused on addressing historical and modern transliteration challenges. A Balinese transliteration system incorporated backward compatibility to maintain traditional linguistic characteristics while adapting to modern standards (Indrawan, 2020). Such compatibility was crucial for ensuring historical transliterations remained intelligible and applicable in contemporary linguistic contexts. Meanwhile, an English-to-Tamil phonetics-based transliteration tool demonstrated improved phoneme alignment for accurate script conversion (Anbukkarasi, 2023). With its distinct phonetic structure, Tamil benefited greatly from a rule-based approach that carefully mapped English phonemes to their Tamil counterparts. The development of LBtrans-Bot highlighted the role of transliteration in preserving the Balinese script through robotic applications (Danilov et al., 2016). This development demonstrated the growing importance of transliteration in digital applications, ensuring that traditional scripts could be integrated into modern technological environments.

Rule-based transliteration has also been adapted for social media text and informal writing. The RBLatAm system integrated user-defined rules to enhance Latin-to-Amharic transliteration in digital communication (Indrawan et al., 2018b). Integrating user-defined rules was crucial for improving communication in multilingual digital spaces, where transliteration often serves as a bridge between languages. Similarly, Nyastra, a rule-based transliteration application, addressed pronunciation-based inconsistencies in Balinese transliteration observed in public signage (Laksminy, 2024). Another system, RuTUT, successfully handled Roman Urdu-to-Urdu transliteration using character substitution rules and Unicode-based mappings (Shahroz et al., 2020). These examples highlight the versatility of rule-based transliteration in handling a wide range of linguistic challenges.

Despite their advantages, rule-based approaches face several challenges. One major limitation is the extensive manual effort to

define and refine linguistic rules. Translating Polish and English proper nouns into Lithuanian encountered difficulties due to variations in transliteration rules over time, requiring a comprehensive exception list (Kasparaitis, 2023). This difficulty highlighted the challenge of maintaining consistency in rule-based systems, particularly when dealing with languages that have undergone significant orthographic changes. Similarly, the method developed to accommodate backward compatibility on the learning application-based transliteration to the Balinese script struggles with maintaining compatibility with older transliteration rules, complicating the process (Indrawan, 2021).

Another significant challenge is handling phonetic variability and script inconsistencies. Ottoman Turkish transliteration faced historical script variations that complicated character mappings (Jaf and Kayhan, 2021). A Balinese transliteration system struggled with backward compatibility due to inconsistencies between older and modern transliteration standards (Indrawan, 2020). In multi-script transliteration to Punjabi, phonetic information played a crucial role beyond simple grapheme conversion, making rule-based approaches insufficient in certain cases (Kaur and Singh, 2015a).

Rule-based methods are also constrained by their reliance on static mappings, which may not fully capture linguistic ambiguities. The Latin-to-Amharic transliteration system encountered difficulties handling numeric representations, script mapping conventions, and double consonants (Indrawan et al., 2018b). Likewise, a Kurdish transliteration system faced inconsistencies in character representation between Arabic-based and Latin-based orthographies, complicating normalization efforts (Ahmadi, 2019).

From a broader perspective, transliteration methods have evolved significantly over the past decade, shifting from traditional rule-based approaches to more adaptive, data-driven techniques. Early transliteration systems heavily relied on manually crafted linguistic rules, which provided high accuracy in structured environments but struggled with linguistic variability. The introduction of statistical and hybrid models enhanced transliteration by incorporating probabilistic learning, reducing dependency on rigid predefined rules. More recently, deep learning architectures such as Transformer-based models have further advanced transliteration, enabling context-aware predictions and handling low-resource language pairs more effectively.

This shift has been particularly evident in applications that require real-time and scalable transliteration, such as machine translation, OCR, and speech processing. For instance, recent studies highlight how neural machine transliteration improves cross-lingual text processing by learning from multilingual corpora. Additionally, phoneme-based modeling and adaptive rule-learning advancements have bridged the gap between traditional rule-based approaches and modern machine-learning techniques, providing more robust transliteration solutions across diverse linguistic landscapes.

Recent advancements in hybrid transliteration models suggest that integrating rule-based accuracy with the adaptability of statistical and machine-learning techniques can yield more robust solutions. Optimizing rule-based frameworks through dynamic rule learning and adaptive phonetic modeling has been explored as a potential approach to enhance transliteration performance.

Ultimately, the shift from purely rule-based transliteration to more flexible, data-driven methodologies reflects a broader trend in computational linguistics. While rule-based methods will remain relevant, especially for low-resource languages with limited training data, their role will likely evolve into a complementary function within more advanced AI-driven transliteration frameworks. This transition highlights the importance of interdisciplinary research in bridging linguistic knowledge with computational techniques to develop scalable and effective transliteration systems.

B. Statistical Method

Statistical methods have significantly contributed to the advancement of machine transliteration over the past decade. These approaches rely on probabilistic models trained on large parallel corpora, allowing them to capture phonetic relationships between different scripts. Statistical transliteration offers greater flexibility and adaptability than rule-based methods, making it suitable for handling phonetic variations and script inconsistencies. Over time, statistical approaches have evolved, integrating more sophisticated techniques to address the limitations of early models and improve transliteration accuracy, particularly for low-resource languages.

As presented in Table 7, statistical transliteration techniques utilize structured probabilistic modeling to support effective script conversion across various linguistic contexts. These methods have been instrumental in expanding transliteration capabilities across diverse languages. However, they continue to face challenges such as data sparsity, phonetic inconsistencies, and computational complexity, which sometimes limit their applicability. The following sections provide an in-depth discussion of these methods, highlighting their advantages and the challenges they seek to overcome.

One of the key statistical transliteration techniques developed in recent years is SMT. An SMT-based system converted Nôm scripts into Vietnamese national scripts, demonstrating improved accuracy over rule-based methods (Dinh et al., 2021). The statistical framework provided a data-driven mechanism for learning transliteration mappings from bilingual corpora, reducing the need for manually defined linguistic rules. Nevertheless, polyphonic characters and limited annotated data posed substantial obstacles, affecting accuracy and robustness, especially for languages with intricate phonetic structures.

Phonology-augmented statistical frameworks have also been introduced to refine transliteration accuracy, particularly in low-resource languages. These frameworks incorporate phonetic features into statistical models, emphasizing the role of pseudo-syllables in improving phonetic alignment (Ngo, 2019). By integrating linguistic knowledge with statistical learning, these models enhance transliteration consistency. Despite this, the dependency on predefined phonetic representations limits scalability to languages with highly variable phonetic structures.

Another statistical approach that has gained prominence is the HMM-based method. A study on Punjabi-to-English transliteration demonstrated that HMM effectively captured sequential dependencies within transliteration tasks, improving phonetic alignment between source and target languages (Garg, 2019). The probabilistic structure of HMM enabled robust handling of phonological variations. However, these models strongly depended on high-quality parallel corpora, which restricted their effectiveness in low-resource environments where sufficient training data is unavailable.

WFST has also been applied in transliteration tasks, particularly for converting the Tunisian Arabic chat alphabet into its formal script. By incorporating linguistic features into probabilistic transducers, WFST-based transliteration demonstrated its capability to process informal and non-standard orthographies prevalent in digital communication and social media (Karmani, 2019). Despite its effectiveness in structured scenarios, this approach encountered challenges in dealing with missing phonetic markers and inconsistencies in informal spellings, which affected transliteration accuracy.

A non-parametric Bayesian approach has been explored to facilitate automatic romanization from bilingual corpora. This method demonstrated notable improvements in transliteration mappings for languages with extensive graphemic variations, such as Myanmar and Russian (Taguchi et al., 2015). Bayesian models leverage probabilistic inference to estimate transliteration patterns, making them particularly valuable for under-resourced languages. Nevertheless, transliteration accuracy remained inconsistent for languages exhibiting high phonetic variability, indicating the need for further refinement in capturing contextual phonetic dependencies.

Another advancement in statistical transliteration involved using an n-gram language model-based forward–backward transliteration approach. This technique was effectively employed for the Punjabi

Table 7
Comparison of statistical transliteration methods.

Language/Script	Method	Advantages	Disadvantages	Reference
Nôm–Vietnamese	(SMT)	Improved accuracy over rule-based methods	Struggles with polyphonic characters and data sparsity	Dinh et al. (2021)
Punjabi–English	HMM	Captures sequential dependencies and phonetic alignment	Strongly depends on high-quality parallel corpora	Garg (2019)
Tunisian Arabic chat alphabet	WFST	Effective for informal and non-standard orthographies	Reduced accuracy due to missing phonetic markers and orthographic variations	Karmani (2019)
Myanmar, Russian	Non-parametric Bayesian approach	Handles extensive graphemic variations	Struggles with high linguistic variability and a lack of robust parallel datasets	Taguchi et al. (2015)
Punjabi Gurmukhi	N-gram language model (forward–backward)	High accuracy when bilingual data is available	Performance drops significantly in low-resource settings due to data sparsity.	Goyal (2022)

Gurmukhi script, demonstrating that corpus-based statistical models achieve high transliteration accuracy when sufficient bilingual data is available (Goyal, 2022). The reliance on frequency-based probabilistic modeling allowed for improved transliteration predictions. However, as with other statistical methods, performance declined significantly in low-resource scenarios where data sparsity hindered model training.

Despite substantial progress, statistical transliteration methods still face critical limitations that affect their general applicability. One of the primary concerns is data sparsity, which restricts the effectiveness of statistical models that rely on large bilingual corpora for training. For instance, the transliteration of Nôm to Vietnamese encountered accuracy issues due to the scarcity of annotated data, making it difficult to establish reliable transliteration mappings (Dinh et al., 2021). Similarly, low-resource languages such as Balinese and Myanmar often lack sufficient training corpora, leading to degraded model performance in statistical approaches.

Handling phonetic variability and linguistic inconsistencies remains another significant challenge. While HMM provides a structured approach to learning phonetic correspondences, it struggles to generalize well across different transliteration contexts. The HMM-based Punjabi-to-English transliteration model degraded performance when faced with irregular spellings and ambiguous phonemes, emphasizing the difficulty of applying statistical models to languages with high phonetic complexity (Garg, 2019).

Similarly, WFST-based transliteration methods, while effective in processing structured text, encountered substantial challenges in informal settings, such as the transliteration of the Tunisian Arabic chat alphabet. Missing phonetic markers and orthographic variations reduced transliteration accuracy, requiring additional morphological analysis to mitigate inconsistencies (Karmani, 2019). The necessity for such post-processing steps increases computational complexity, reducing the efficiency of WFST-based transliteration models for real-time applications.

The historical evolution of scripts also presents complications in transliteration tasks. Due to extensive graphemic variations and the absence of robust parallel datasets, the automatic induction of romanization systems from bilingual corpora has been particularly difficult for languages like Myanmar and Russian. While effective in probabilistic inference, the non-parametric Bayesian transliteration model struggled to maintain accuracy across languages with highly inconsistent phonetic structures (Taguchi et al., 2015).

Additionally, n-gram-based forward–backward transliteration models suffered significant performance degradation in low-resource environments despite their ability to generate high accuracy scores with well-annotated datasets. This limitation underscores the importance of data quality and availability in ensuring the reliability of statistical transliteration methods (Goyal, 2022). The challenges faced by statistical models have led to increasing interest in hybrid approaches that integrate statistical and deep learning techniques to enhance transliteration robustness.

The role of statistical transliteration methods in language processing remains significant, particularly in structured environments where large bilingual corpora are available. These models have enabled efficient transliteration by leveraging probabilistic learning techniques to capture phonetic relationships across different scripts. Their ability to generalize transliteration patterns without requiring explicit linguistic rules has contributed to their widespread adoption in various applications, including machine translation and digital text processing. However, their reliance on high-quality training data remains a major constraint, particularly for low-resource languages where annotated datasets are scarce.

Furthermore, statistical methods often struggle with phonetic inconsistencies and ambiguous spellings, limiting their effectiveness in dynamic linguistic contexts such as social media text and informal communication. While probabilistic models provide a structured framework for transliteration, they lack the adaptability of more advanced deep learning techniques, which can dynamically adjust to phonetic and contextual variations. The increasing demand for multilingual transliteration across diverse scripts further emphasizes the need for enhancements in statistical methodologies, including incorporating adaptive modeling techniques and hybrid approaches that integrate rule-based precision with statistical inference.

As transliteration research progresses, ongoing refinements in probabilistic modeling and data augmentation strategies will be crucial to improving the robustness of statistical transliteration methods. Integrating statistical models with emerging neural architectures may provide a more balanced approach, combining statistical frameworks' interpretability with machine learning techniques' adaptability. Addressing these challenges will play a key role in advancing the reliability and scalability of transliteration systems, ensuring their effectiveness in a broader range of linguistic and computational applications.

C. Machine Learning

Machine transliteration has greatly benefited from advancements in machine learning. The application of various machine learning techniques in transliteration tasks has demonstrated their effectiveness in managing the complexities of converting text between different scripts.

As summarized in Table 8, machine learning-based transliteration methods leverage various neural architectures to improve transliteration accuracy across different languages and scripts. While offering flexibility and enhanced performance, these approaches also present challenges such as data scarcity, inference complexity, and contextual adaptability. The following sections provide a detailed discussion of these methods, their advantages, and the challenges they address.

Recurrent Neural Network (RNN) and its variant, Long Short-Term Memory (LSTM) networks, have been widely utilized due to their ability to model sequential data. A bidirectional RNN agreement model was proposed to optimize sequence-to-sequence learning, significantly improving transliteration accuracy in Japanese-to-English tasks (Liu et al., 2020). The introduction of such models marked a shift from purely

Table 8
Comparison of machine learning-based transliteration methods.

Language/Script	Method	Advantages	Disadvantages	Reference
MODI–Devanagari	CRNN model	Effectively captures spatial and sequential features for handwritten script transliteration	Lack of the MODI dataset and missing word demarcation symbols	Joseph (2024)
Hindi–English	GRU-based sequence-to-sequence model	Captures long-range dependencies with lower computational cost than LSTM	Struggles with morphological complexity and syntactic differences	Ansari (2022)
Romanized Tunisian	Hybrid BiLSTM-CRF sequence labeling model	Improves transliteration accuracy by integrating sequence labeling and contextual dependencies	The ambiguity of Latin characters complicates transliteration	Younes (2022)
French–Vietnamese	RNN with alignment representation and pretrained embeddings	Enhances transliteration accuracy using alignment-based phonetic modeling	Graphemic variations complicate phonetic alignment, affecting transliteration consistency	Le (2019)
Japanese–English	Bidirectional RNN agreement model	Optimizes sequence-to-sequence learning with improved alignment accuracy	Inference complexity and noise in hidden states over long sequences	Liu et al. (2020)
Low-Resource Languages	3-way transfer learning approach	Reduces data scarcity issues by leveraging pivot languages	High dependency on related language datasets for effectiveness, particularly in languages such as Chinese and Korean	Wu (2022)
Roman-Urdu–Urdu	Bidirectional attention-based model	Focuses on relevant segments of input text, improving transliteration accuracy	Requires better contextual adaptation for transliteration of homophones	Alam (2021)

rule-based methods to data-driven approaches that can better capture phonetic correspondences across scripts. Despite these advancements, these models introduced computational complexity and challenges related to noise accumulation over long sequences, requiring further refinement in inference techniques.

Similarly, LSTM networks were employed for transliterating and translating Twitter data, effectively handling informal language and social media slang (Vathsala and Holi, 2020). The success of LSTM in handling transliteration within unstructured digital text environments highlights the adaptability of ML-based methods. However, reliance on large annotated datasets remains a limitation, particularly for low-resource languages.

Multilayer GRU has also been explored for Hindi-to-English transliteration, demonstrating improved accuracy through gating mechanisms that capture long-range dependencies in text (Ansari, 2022). The GRU architecture reduces computational complexity while maintaining comparable performance to LSTMs, making it a more efficient alternative in transliteration tasks requiring real-time processing. Nevertheless, the complexity of morphologically rich languages like Hindi still poses challenges, as transliteration must account for syntactic and semantic variations that cannot always be captured by sequence-based models alone.

Deep learning techniques incorporating bidirectional models and attention mechanisms have further advanced machine transliteration. A Roman-Urdu-to-Urdu transliteration model using attention-based bidirectional networks achieved the highest BLEU scores, demonstrating superior performance over unidirectional models (Alam, 2021). Attention mechanisms represent a key advancement in transliteration, allowing models to focus on relevant parts of input sequences. Nevertheless, this approach is still sensitive to the availability of high-quality training data, and handling homophones in transliteration remains an ongoing challenge.

The convolutional recurrent neural network (CRNN) model has also been applied to transliterate the handwritten MODI script to Devanagari. This approach improved character-level transliteration accuracy by leveraging spatial and sequential features, creating a valuable handwritten MODI dataset (Joseph, 2024). The combination of convolutional and sequential processing highlights a growing trend in transliteration research: integrating different deep learning architectures to enhance transliteration performance for complex scripts. Despite its effectiveness, the lack of large datasets for handwritten historical scripts remains a major barrier to further advancements.

Sequence labeling techniques using CRF and Bidirectional Long Short-Term Memory (BiLSTM) have been investigated for Romanized Tunisian dialect transliteration. A hybrid BiLSTM-CRF model successfully integrated sequence labeling strategies, achieving higher transliteration accuracy by capturing contextual dependencies (Younes, 2022). These techniques enhance transliteration by leveraging probabilistic sequence prediction, which improves the handling of ambiguous phonetic mappings. However, transliteration models for languages with highly variable orthographies, such as Romanized dialects, often require additional linguistic preprocessing steps to ensure consistency.

In addressing low-resource language transliteration, alignment representation sequences and pre-trained embeddings were utilized in an RNN-based model for the French-Vietnamese language pair, demonstrating notable BLEU score improvements (Le, 2019). This approach enables models to generalize transliteration rules from a small amount of bilingual data, making them highly effective for underrepresented languages. However, phonetic variations within a language pair can introduce inconsistencies in transliteration, requiring further refinement in how embeddings are trained and applied.

Furthermore, a three-way transfer learning approach was introduced to enhance transliteration accuracy in low-resource settings. This method optimized dual training objectives by leveraging source-pivot-target datasets and reducing the embedding distance between source and pivot languages (Wu, 2022). The advantage of this method is its ability to utilize resources from related languages to improve performance for less-resourced scripts. On the other hand, dependence on pivot languages introduces potential biases that can affect generalization across unrelated language pairs.

A machine learning approach has also been employed to translate handwritten Hindi text into Braille using Histogram of Oriented Gradients (HOG) features and an SVM classifier. This method demonstrated high accuracy in recognizing Hindi characters and converting them into Braille, significantly improving accessibility for visually impaired individuals (Jha, 2019). While this advancement contributes to inclusive technology, challenges remain in extending such methods to languages with more complex script variations.

Several challenges have been identified in machine learning-based transliteration. The lack of data resources poses a significant limitation, particularly for low-resource languages. Translating the handwritten MODI script to Devanagari struggles due to the absence of a MODI script data repository and word demarcation symbols (Joseph, 2024). Similarly, Hindi transliteration using GRUs faces difficulties due to

the morphological complexity and syntactic differences between Hindi and English (Ansari, 2022). These challenges highlight the necessity of developing more adaptive ML-based methods capable of handling diverse linguistic structures.

Data scarcity remains a critical limitation for deep learning models. Translating the handwritten MODI script to Devanagari highlights the lack of a substantial dataset for effective deep-learning model deployment. Data preparation is lengthy and costly, requiring expert verification (Joseph, 2024). Additionally, the transliteration of Romanized Tunisian dialects requires context consideration at both the word and sentence levels due to the ambiguity of Latin characters, which adds complexity to the transliteration process (Younes, 2022).

The challenge of inference complexity and noise in hidden states is another significant issue in machine transliteration. In bidirectional RNN-based sequence-to-sequence models, exact inference is often intractable due to the agreement mechanism, leading to inconsistencies in transliteration results. Additionally, noise accumulation in hidden states over long sequences reduces model reliability, impacting transliteration accuracy (Liu et al., 2020). Future advancements must improve inference efficiency to ensure more stable transliteration results.

The growing reliance on deep learning models in transliteration highlights the increasing demand for more context-aware and linguistically informed approaches. The increasing sophistication of machine learning architectures has led to significant improvements in transliteration accuracy, particularly in handling phonetic complexities and diverse script systems. Deep learning techniques, such as bidirectional models and attention mechanisms, have enhanced transliteration consistency by allowing models to focus on key linguistic elements. Moreover, integrating pre-trained embeddings has improved phonetic alignment, enabling better adaptation across different language pairs. However, challenges remain in ensuring the stability and interpretability of these models, particularly when dealing with languages that exhibit high morphological complexity. While statistical and hybrid methods have historically played an essential role, the growing reliance on neural models underscores the need for further refinement in balancing computational efficiency with transliteration accuracy. The interplay between linguistic knowledge and data-driven techniques continues to shape transliteration research, reinforcing the importance of structured rules and adaptive learning in achieving reliable script conversion.

D. Hybrid Method

Hybrid methods in machine transliteration have garnered significant attention due to their ability to combine the strengths of various techniques to improve accuracy and efficiency. These approaches integrate rule-based, statistical, and machine-learning techniques to enhance transliteration performance, making them particularly effective in handling diverse linguistic structures and informal text variations. Unlike purely rule-based or statistical models, hybrid approaches leverage multiple methodologies to overcome specific transliteration challenges, ensuring greater flexibility and adaptability.

As presented in Table 9, hybrid transliteration methods offer considerable improvements over standalone approaches by addressing phonetic inconsistencies, informal script variations, and language mixing in transliteration tasks. However, they also face notable challenges, including data sparsity, lack of standardization, and the complexity of integrating multiple methodologies within a single framework. The following discussion elaborates on implementing hybrid transliteration techniques and their associated challenges.

One study explores two models for transliterating Arabizi into Arabic script for the Tunisian dialect. A rule-based approach and a discriminative model based on CRF are used. These models address the transliteration task at both word and character levels, highlighting the complexity of converting informal, non-standard social media and SMS language into formal script. The rule-based model applies predefined transliteration rules, while the discriminative model treats transliteration as a sequence classification task. These models demonstrate their effectiveness by significantly reducing error rates (Masmoudi, 2019). The study highlights the intricate nature of Tunisian Arabizi, which incorporates elements from multiple languages, including Berber, French, and Italian, thereby complicating transliteration consistency. The hybrid approach allows for greater accuracy by leveraging linguistic knowledge and statistical learning. However, the infrequency of certain Arabic characters in the training corpus leads to errors, particularly in handling ambiguous Latin characters and cases like the "ta-marbouta" at the end of words. Such character sparsity in the training corpus underscores the challenge of transliterating informal digital text, where linguistic variations are highly prevalent.

Another study discusses various techniques for processing code-mixed Indian regional languages, focusing on language identification, transliteration, and named entity recognition. The study reviews statistical, rule-based, and neural-based approaches, emphasizing the challenges of informal and non-standard spellings in mixed-code text (Kumbhar, 2024). Accurate language identification is crucial for processing code-mixed text, as each word in a sentence must be correctly labeled. The hybrid approach integrates rule-based segmentation and machine learning-based classification, significantly improving transliteration performance. However, challenges persist due to the variability of Romanized strings across different languages and the difficulty in developing robust language identification models to handle informal text. These challenges illustrate the necessity for extensive training data and adaptable transliteration models that can process evolving linguistic trends in informal digital communication.

A method combining rule-based transliteration with the weighted Levenshtein algorithm was introduced to convert Moroccan Arabizi to Arabic script. This sequential combination leverages the strengths of both approaches, significantly improving transliteration accuracy (Hajbi, 2024). The study highlights the transliteration challenges associated with non-Latin script languages and suggests that the methodology can be applied to other Arabic dialects facing similar issues. However, due to the lack of standardized transliteration rules, Moroccan Arabizi presents multiple variations of the same Arabic word, creating inconsistencies. This one-to-many mapping issue complicates the transliteration process, as users frequently incorporate code-switching and borrow foreign words from languages such as French and Spanish.

Another study explores a backtracking algorithm for generating a Romanian Cyrillic lexicon, combining Cyrillization and de-Cyrillization algorithms with structured transliteration rules (Ciubotaru, 2021). This iterative approach benefits from expert validation, ensuring high accuracy by refining ambiguous mappings. The expert-guided refinement process highlights the necessity of continuous human intervention in certain transliteration tasks, particularly when dealing with historical texts and language digitization efforts. However, the iterative nature of the process emphasizes the labor-intensive nature of hybrid transliteration methodologies, as linguistic inconsistencies require multiple refinement cycles before achieving satisfactory accuracy.

A hybrid approach combining SVM and HMM has also been proposed. In this approach, transliteration is viewed as a multi-class pattern classification problem, where the hybrid model integrates SVM for classification accuracy and HMM for sequence learning (Chatterjee, 2021). The results indicate that SVM outperforms other methods, while HMM provides robust sequential dependencies. This combination suggests that hybrid models can optimize transliteration by leveraging different machine learning techniques. However, the requirement for large annotated corpora and computational resources remains a limiting factor, particularly for low-resource languages.

Another hybrid transliteration approach integrates image processing, OCR, and rule-based conversion techniques for translating Hindi script into English. This method follows three main steps: image processing to extract text from images, OCR to recognize Hindi text, and transliteration to convert the recognized text into English script (Yadav, 2023). The findings highlight the potential of combining visual recognition with transliteration techniques, particularly in cases where

Table 9
Comparison of hybrid transliteration methods.

Language/Script	Method	Advantages	Disadvantages	Reference
Tunisian Arabizi	Rule-based + CRF	Captures linguistic features; reduces error	Struggles with rare characters; ambiguous mappings	Masmoudi (2019)
Indian Code-Mixed	Rule-based + Machine Learning	Improved accuracy in identifying mixed texts	Requires extensive labeled training data	Kumbhar (2024)
Moroccan Arabizi	Rule-based + Weighted Levenshtein Algorithm	Improves transliteration accuracy	Lack of standardization causes inconsistencies	Hajbi (2024)
Romanian Cyrillic	Rule-based + Backtracking Algorithm	Ensures accuracy through iterative refinement	Requires expert involvement for accuracy	Ciubotaru (2021)
Hindi–English	SVM + HMM Hybrid	Strong classification and sequence modeling	Needs large annotated corpora	Chatterjee (2021)
Hindi OCR	Image Processing + OCR + Rule-Based	Effective for digitizing printed text	Challenges in OCR accuracy for handwriting	Yadav (2023)

human transcription is not feasible. Nevertheless, challenges remain in ensuring the accuracy of OCR-based text recognition, especially for handwritten documents where inconsistencies in character shapes may reduce accuracy.

Hybrid methods in machine transliteration have evolved significantly, demonstrating their ability to merge rule-based precision with statistical and machine learning adaptability. Integrating multiple methodologies enhances transliteration quality, making these approaches particularly effective in addressing complex linguistic scenarios. However, the effectiveness of hybrid approaches depends heavily on the availability of large, well-annotated training corpora and computational resources. The growing demand for real-time transliteration in digital communication, multilingual text processing, and historical text digitization underscores the need for hybrid models that dynamically adapt to linguistic variations.

Despite their advantages, hybrid transliteration models face significant standardization, scalability, and processing efficiency challenges. Integrating multiple techniques requires extensive optimization to ensure seamless performance across various scripts and languages. Developing more adaptable hybrid models also necessitates continuous advancements in linguistic data annotation, machine learning optimization, and phonetic modeling. By refining hybrid methodologies and enhancing their ability to learn from diverse linguistic contexts, transliteration systems can achieve greater accuracy and robustness, bridging linguistic gaps in an increasingly digitalized world.

E. Semantic Knowledge-Based Model

Semantic knowledge integration in machine transliteration has demonstrated significant potential, particularly in enhancing accuracy and adaptability. Unlike traditional statistical or rule-based approaches, semantic knowledge-based models incorporate linguistic attributes such as language origin and gender detection to refine transliteration accuracy. These models generate contextually aware transliterations by leveraging semantic information, making them especially effective for named entity transliteration across diverse language pairs. This approach has been particularly useful in transliterating Uyghur-Chinese person names, where semantic scores enhance performance beyond phonetic-based models (Murat et al., 2017).

Experiments incorporating semantic knowledge have shown promising results. A study applied semantic features, including gender identity and language origin, to optimize a general transliteration model, significantly improving accuracy for Uyghur-Chinese names (Murat et al., 2017). The gender attribute proved especially effective, as identifying a name's masculinity or femininity aided in generating more accurate transliterations. While the language origin feature also contributed, its impact was less pronounced, suggesting that different semantic attributes influence transliteration performance to varying degrees.

Further testing with the UC-C Data Experiment highlighted a key limitation: models lacking semantic integration struggled with Uyghur person names of Chinese origin, reducing overall accuracy. However, integrating semantic knowledge mitigated these issues, improving the system's handling of names from mixed linguistic backgrounds. These results underscore the adaptability of semantic models in addressing complex transliteration challenges beyond simple phonetic transformations. Table 10 illustrates how a semantic knowledge-based transliteration method incorporates contextual attributes to enhance accuracy. These approaches enable more refined transliterations, particularly for named entities and culturally specific terms, reducing errors caused by phonetic ambiguities.

Despite these advantages, semantic knowledge-based models face notable challenges. One primary limitation involves the complexity of automatic transliteration across languages with distinct alphabets and phonological structures, such as Uyghur and Chinese. Unlike Latin-based scripts, where phonetic mappings are more predictable, transliterating between orthographically distant languages requires sophisticated linguistic modeling. Additionally, research on Uyghur-Chinese name transliteration remains limited, leading to gaps in developing effective techniques tailored for these languages (Murat et al., 2017). Addressing this challenge requires further investigation into how named entities transform across linguistic systems.

Another major challenge lies in character selection and variation. Unlike statistical machine translation, which emphasizes alignment and reordering, person name transliteration demands deeper consideration of semantic factors such as gender and language origin. Incorporating these attributes makes semantic transliteration distinct from conventional machine translation, requiring an interdisciplinary approach that blends linguistics, computational modeling, and AI-driven learning algorithms (Murat et al., 2017).

Semantic knowledge-based models have redefined machine transliteration by introducing linguistic context. Semantic integration enables more nuanced outputs than traditional models that rely primarily on phonetic rules or statistical probabilities. This approach is particularly valuable for proper nouns, named entities, and culturally significant terms where accuracy is crucial.

The continued advancement of semantic-based transliteration depends on the availability of rich linguistic datasets, improved semantic annotation techniques, and deeper integration with modern neural architectures. While current implementations demonstrate notable accuracy gains, future research should focus on refining language-specific semantic modeling to enhance adaptability across multilingual transliteration tasks. Moreover, integrating semantic-driven transliteration into AI-powered NLP systems will further improve cross-lingual text processing, reinforcing the role of semantic knowledge in computational linguistics.

The discussion on various transliteration methods highlights that each approach has distinct advantages and limitations. Rule-based methods offer clear and interpretable transformations but struggle with linguistic irregularities that cannot be fully addressed through predefined rules. In contrast, statistical and probabilistic models provide greater flexibility by learning phonetic variations from data. However, their effectiveness is often constrained by the availability of large, high-quality training corpora.

Table 10
Semantic knowledge-based transliteration methods.

Language/Script	Method	Advantages	Disadvantages	Reference
Uyghur-Chinese	Semantic Knowledge + Phonetic Model Semantic Knowledge + Named Entity Recognition	Enhances accuracy with gender & origin attributes Improved contextual understanding of names	Struggles with transliterating rare name variations Limited research on Uyghur-Chinese transliteration	Murat et al. (2017)

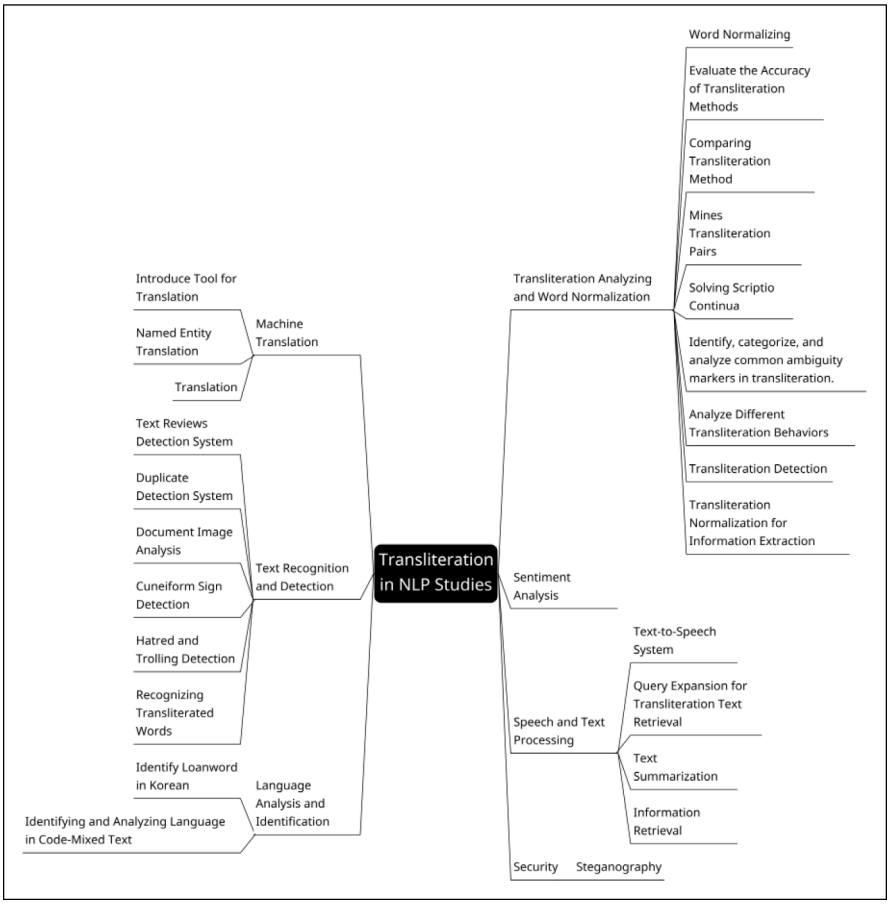

Fig. 8. Transliteration in various NLP studies.

An alternative to these approaches is machine learning, particularly deep learning techniques, which have advanced transliteration by capturing complex phonetic and orthographic patterns. These models handle non-linear linguistic transformations and adapt to diverse language structures. Nevertheless, their reliance on extensive computational resources and large labeled datasets poses practical limitations, especially in resource-constrained environments.

Semantic knowledge-based models provide an alternative by incorporating linguistic and contextual knowledge to address transliteration challenges beyond statistical inference and data-driven learning. These models leverage structured semantic information, such as lexical databases and ontologies, to enhance transliteration accuracy, particularly when phonetic transformations are insufficient. While semantic models improve contextual relevance, their effectiveness depends on the availability of high-quality linguistic resources and domain-specific knowledge bases.

Considering the strengths and limitations of these methods, hybrid transliteration approaches have gained attention as a way to integrate rule-based precision with the adaptability of statistical, neural, and semantic knowledge-based models. Combining linguistic rules with data-driven learning techniques and structured semantic information, hybrid methods aim to enhance accuracy and interpretability, making them suitable for addressing a wider range of transliteration challenges.

Choosing the most appropriate transliteration method depends on several factors, including linguistic complexity, data availability, and computational efficiency. Rule-based systems ensure consistency and interpretability, while statistical and machine-learning approaches offer greater adaptability to language variations. By leveraging the strengths of multiple techniques, hybrid methods provide a balanced and effective solution. Further refinement of hybrid architectures and integration of explainable AI techniques can improve accuracy and transparency in transliteration systems.

4.2. Transliteration in various NLP studies and their impacts

Transliteration in NLP has significantly supported and enhanced research in various applications. Transliterating text from one script to another is vital in numerous NLP tasks, providing foundational support for further studies and advancements. Based on the analysis of selected studies, Fig. 8 illustrates the application of transliteration in several research areas. These include Machine Translation, Text Recognition and Detection, Language Analysis and Identification, Transliteration Analyzing and Word Normalization, Sentiment Analysis, Speech and Text Processing, and Security. Each area addresses specific challenges and advancements, highlighting transliteration's multifaceted role. This analysis underscores the importance of transliteration in improving

NLP, enabling the development of robust and versatile systems for effective language processing.

A. Machine Translation

Machine translation (MT) uses computer software to translate text or speech from one language to another. This process involves employing algorithms and models to facilitate the translation, aiming to improve accuracy and efficiency in linguistic conversions (Chen et al., 2023). The advancement of MT has been driven by integrating sophisticated computational techniques, particularly transliteration, which serves as a crucial intermediary for improving translation quality across diverse linguistic contexts. While conventional MT systems rely heavily on direct word-to-word or phrase-based translations, transliteration offers a unique mechanism to address phonetic and orthographic discrepancies, especially for languages with different scripts. By mapping phonetic similarities across languages, transliteration not only aids in better word alignment but also enhances named entity recognition, low-resource language translation, and historical text digitization. This study explores the various implementations of transliteration in MT, highlighting its contributions and limitations in different applications.

A comparative analysis of various transliteration approaches is presented in Table 11 to better understand the role of transliteration in MT. This table outlines key methodologies, their advantages, and the challenges of different transliteration techniques across multiple language pairs. The comparison highlights the effectiveness of machine learning-based transliteration models in improving translation accuracy, especially for low-resource languages and complex scripts. Additionally, it underscores the trade-offs between rule-based, statistical, and neural approaches in handling phonetic variations and script inconsistencies. The insights from this comparison provide a foundational understanding of how transliteration enhances MT systems while identifying areas for further refinement. Building on this comparative analysis, the following sections delve deeper into specific transliteration methods and their applications in different MT scenarios.

One approach involves bilingual named entity conversion through a chunk symmetry strategy combined with an English–Chinese transliteration model (Li et al., 2021). This method generates candidate entity translation pairs by aligning bilingual named entities and reordering and correcting using a transliteration model. By leveraging machine learning, the system refines candidate translations, improving the identification of person, place, and institution names. This hybrid approach effectively mitigates inconsistencies in entity translation, which often arise due to phonetic variations between languages. However, while it enhances named entity recognition, the reliance on predefined alignment strategies limits its adaptability to highly dynamic language pairs. Refining this approach requires exploring deep learning models capable of learning entity mappings from larger multilingual datasets without rigid alignment constraints.

The Transkribus tool significantly advances historical document translation, integrating Handwritten Text Recognition (HTR) with Document Understanding functionalities to process and translate manuscripts (Vukčević, 2020). Transliteration plays a crucial role in this context by accurately mapping handwritten characters into machine-readable text, ensuring that the digital output preserves historical documents' semantic and contextual integrity. A German-Serbian case study highlights its effectiveness in improving translation quality, particularly for complex scripts that pose challenges for traditional OCR methods. However, the variability in handwriting styles and the degradation of historical texts remain obstacles. These challenges underscore the necessity of adaptive transliteration models that can learn from diverse handwriting datasets and incorporate linguistic heuristics to improve text restoration.

In the context of low-resource language (LRL) translation, transliteration serves as a bridge between high-resource languages (HRL) and LRL by transforming HRL text into LRL-like structures (Karakanta et al., 2018). This approach implemented through a sequence-to-sequence LSTM model with a bidirectional encoder and global attention, facilitates the creation of parallel corpora for underrepresented languages. By leveraging phonetic and orthographic similarities, transliteration enhances word alignment between HRL and LRL, improving MT accuracy. However, the quality of translation still depends on the availability of high-quality monolingual corpora, which remains a challenge for many LRLs. To address this, further research should focus on unsupervised and semi-supervised learning techniques that can generate synthetic parallel data to augment training corpora.

A transliteration-based Bodo-English MT system provides further evidence of transliteration's impact on translation accuracy, particularly for languages with limited digital resources (Islam, 2019). Employing a phrase-based statistical MT model effectively handles general and news domain translations while addressing issues related to out-of-vocabulary (OOV) words. The conversion of Bodo script into Romanized text enables a more consistent mapping between language pairs, reducing translation ambiguities. While the model demonstrates improvements in translation accuracy, its effectiveness is constrained by the availability of extensive parallel corpora. The model's dependency on extensive parallel corpora highlights the need for continuous corpus expansion and integration of neural MT techniques to enhance robustness.

For morphologically rich languages, a Sanskrit-Gujarati MT system exemplifies how transliteration can facilitate accurate translation despite complex linguistic structures (Raulji et al., 2022). This system employs a grammatical transfer approach, incorporating tokenization, lemmatization, and morphological analysis to handle Sanskrit's rich inflectional morphology. Including transliteration ensures that word structures are preserved, maintaining the linguistic integrity of the translated text. The model achieved a satisfactory BLEU score, demonstrating its effectiveness in reducing surface-level translation errors. However, given the high degree of morphological variation of Sanskrit and Gujarati, incorporating a hybrid deep learning framework that integrates rule-based and statistical models could enhance translation performance.

Multilingual Neural Machine Translation (MNMT) has also benefited from transliteration, particularly for Indic languages such as Malayalam and Tamil (Bala Das et al., 2024). Adopting a standardized transliteration script (ITRANS) ensures consistency across different Indic languages, facilitating improved translation accuracy. Using English as a pivot, this model enables more effective translations between Indic languages, reducing the complexity of direct translation between low-resource language pairs. This approach highlights the advantage of transliteration in language standardization and exposes the limitation of pivot-based MT models, which can introduce additional translation errors. Further exploration of end-to-end MNMT models that incorporate transliteration as an intrinsic feature, rather than relying on intermediary pivot languages, is essential for improving translation accuracy.

In English-Amharic MT, transliteration has been leveraged to enhance translation performance through corpus-based methods (Biadgligne, 2023). This study demonstrated substantial improvements in BLEU scores using Google's transliteration tool to convert Amharic sentences into English. Creating a high-quality Amharic-English transliteration corpus helped overcome the challenge of script differences, enabling more effective MT for this low-resource language pair. However, while automated transliteration enhances translation accuracy, it struggles with dialectal variations and script ambiguities. These limitations highlight the need for dynamic transliteration models that adapt to regional language variations and user-specific preferences.

Finally, a Multi-Round Transfer Learning (MRTL) approach has introduced a novel method for improving neural MT by unifying transliteration across language families (Maimaiti et al., 2019). This approach reduces character-level differences through a shared transliteration

Table 11
Comparison of transliteration approaches in machine translation.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Machine Translation
Hybrid Named Entity Transliteration Li et al. (2021)	Hybrid approach combining bilingual named entity alignment with ML-based transliteration	English–Chinese Named Entities	Combines bilingual named entity alignment with machine learning-based transliteration	Requires optimization to enhance accuracy	Improves translation of person, place, and institution names
Transkribus HTR & Document Understanding Vukčević (2020)	Automated tool integrating HTR and Document Understanding	Historical Document Translation (German-Serbian)	Recognizes, transcribes, and translates handwritten texts	Limited accuracy due to handwriting variations	Enhances historical document digitization and translation
Transliteration for LRL MT Karakanta et al. (2018)	LSTM model leveraging bilingual glossaries and phonetic transformation	High-Resource to Low-Resource Language Pairs	Uses LSTM model and bilingual glossaries for phonetic transformation	Requires high-quality monolingual corpora	Generates parallel corpora, improving MT for LRL
Bodo-English Transliteration Islam (2019)	Phrase-Based Statistical MT	Bodo-English Translation	Converts Bodo script to Roman script before translation	Out-of-vocabulary words impact accuracy	Reduces translation errors for unknown words
Sanskrit-Gujarati Symbolic MT Raulji et al. (2022)	Grammatical Transfer Model	Sanskrit-Gujarati Translation	Performs morphological analysis before translation	Complexity of morphologically rich languages	Ensures accurate representation of Sanskrit text in Gujarati
MNMT with ITRANS Bala Das et al. (2024)	MNMT	Indic-to-Indic Language Translation	Uses a standardized script for better cross-lingual transfer	Requires high-quality English-Indic corpora	Improves consistency in MNMT models
English-Amharic MT Biadgligne (2023)	Transliteration corpus with Google's online tool	English-Amharic Machine Translation	Creates a high-quality Amharic-English corpus	Low-resource language challenges	Boosts BLEU scores and MT model performance
Multi-Round Transfer Learning Maimaiti et al. (2019)	Low-Resource NMT using transliteration for cross-language consistency	Low-Resource NMT	Applies transliteration for cross-language character consistency	Needs shared vocabulary across training languages	Enhances translation accuracy via character-level mapping

framework, enabling better transfer learning from high-resource to low-resource languages. The experiments demonstrated that increasing the size of parent models during fine-tuning significantly enhances translation quality. However, the effectiveness of this approach depends on the degree of lexical similarity between the parent and child languages, indicating that further optimization is required for languages with minimal phonetic overlap.

Across these applications, transliteration emerges as a key strategy in MT, enhancing named entity recognition, low-resource language processing, and historical text translation. Its integration with machine learning and deep learning models continues to refine MT systems, demonstrating its potential to bridge linguistic gaps and improve translation accuracy. Further exploration is needed to address challenges such as reliance on pre-aligned datasets, difficulty processing highly inflected languages, and dependence on existing corpora. Optimizing transliteration models remains crucial to addressing linguistic diversity in MT, particularly in handling morphologically rich languages and low-resource translation scenarios. Advances in multilingual neural architectures with built-in transliteration mechanisms could enhance MT system adaptability and reduce dependency on extensive external corpora, making translations more precise and scalable. Additionally, incorporating multilingual neural architectures with built-in transliteration mechanisms could reduce the dependency on external corpora, making MT systems more autonomous and scalable.

The growing role of transliteration in MT signifies a paradigm shift in how languages are processed computationally. By leveraging phonetic and orthographic similarities across linguistic families, transliteration improves word alignment and fosters the development of more inclusive MT systems that accommodate linguistic diversity. As MT continues to evolve, transliteration will remain a cornerstone of multilingual NLP, driving innovations in cross-linguistic communication and digital language preservation.

B. Sentiment Analysis

Transliteration is essential in sentiment analysis, particularly for processing multilingual and dialectal data where script inconsistencies hinder accurate classification. By aligning textual representations across different linguistic forms, transliteration ensures that sentiment-bearing expressions are preserved, improving the robustness of classification models.

Various studies have explored different transliteration techniques, integrating them with sentiment analysis frameworks to enhance text representation and classification accuracy. Table 12 compares these approaches, outlining their methods, applications, key features, challenges, and impact on sentiment analysis. Examining these variations provides insight into the role of transliteration in optimizing sentiment classification and addressing linguistic challenges. The following discussion explores specific methods that illustrate these principles in practical applications.

A notable approach integrating transliteration is the Social Eagle Optimization (SoEo) algorithm-based deep BiLSTM model, which applies transliteration after text normalization and language identification (Londhe and Kumari, 2022). This step is crucial as it restores text to its original script, reducing inconsistencies that often arise in multilingual sentiment analysis. Unlike conventional methods that struggle with script variations, this approach ensures a more stable representation of sentiment-laden words, leading to higher classification accuracy. The model's significant improvement over state-of-the-art techniques underscores how transliteration enhances text processing and complements deep learning architectures such as BiLSTM networks by providing

Table 12
Comparison of transliteration approaches in sentiment analysis.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Sentiment Analysis
SoEo Algorithm-based deep BiLSTM for Sentiment Analysis Londhe and Kumari (2022)	SoEo algorithm-based deep BiLSTM model	Sentiment classification of multilingual texts	- The transliteration phase converts text to its original script - Improved classification accuracy over state-of-the-art methods	- Handling multilingual data variations - Lack of accessible language dictionaries	- Enhances sentiment prediction across multiple languages - Addresses linguistic resource limitations
Semi-supervised Approach for Arabizi Sentiment Analysis Guellil et al. (2021)	Semi-supervised learning with deep learning models (RF, LR, CNN, LSTM)	Sentiment analysis of Arabizi (Algerian dialect) messages	- Uses transliteration to convert Arabizi into Arabic - Applies word embedding models (Word2Vec, fastText)	- Selecting the best transliteration candidate - Errors in transliteration impacting sentiment classification	- Improves sentiment classification for dialectal Arabic - Enables better handling of Arabizi in digital communication

a more coherent linguistic input. By bridging gaps in language resources, this method demonstrates that transliteration is not merely a preprocessing step but an integral component in refining sentiment prediction.

Similarly, a semi-supervised approach for sentiment analysis of Arabizi messages highlights the advantages of transliteration over translation, particularly for dialectal variations in the Algerian context. Arabizi is a hybrid Arabic form incorporating Latin characters (Guellil et al., 2021). It is commonly used in digital communication but lacks standardized orthography. This study constructs a sentiment corpus automatically and manually, ensuring validation by native speakers of the Algerian dialect. By integrating transliteration with deep learning models such as random forest, logistic regression, CNN, and LSTM networks, this study improves the accuracy of sentiment mapping for dialectal expressions. While translation often distorts the contextual meaning of dialect-specific words, transliteration preserves linguistic nuances, making it a more reliable strategy. However, challenges remain in selecting the optimal transliteration candidate, as incorrect mappings can reduce classification accuracy. The study mitigates this issue by incorporating additional parameters, such as contextual distance, to refine candidate selection. This approach strengthens sentiment classification for dialectal languages and highlights the necessity of expanding parallel corpora for more effective transliteration models.

These studies reinforce transliteration as a fundamental tool in sentiment analysis, particularly for languages with limited digital resources. By preserving linguistic integrity and enhancing model adaptability, transliteration proves indispensable in addressing script discrepancies and dialectal complexities. Advancements in this area should focus on refining transliteration techniques through adaptive learning mechanisms, enabling models to dynamically adjust to diverse linguistic patterns and improve sentiment classification across varied language contexts.

C. Text Recognition and Detection

Transliteration methods have gained considerable attention in text recognition and detection due to their ability to process complex and multilingual datasets. Various approaches have been explored to improve recognition accuracy and detection efficiency by integrating transliteration techniques. Transliteration plays a crucial role in enhancing text processing capabilities across different languages and orthographic systems by facilitating the conversion of diverse scripts into more recognizable forms.

In this context, transliteration is a bridge for better representation and interpretation of multilingual text, ensuring more accurate recognition and detection in computational applications. Table 13 analyzes various transliteration techniques, emphasizing their applications, distinct characteristics, challenges, and contributions to text recognition and detection. This summary offers valuable insights into how these methods address linguistic variations and script complexities, forming the foundation for the following in-depth discussions.

An HLSTM model has been implemented to detect hate speech and trolling content in code-mixed social media text. This approach incorporates word embedding and character embedding features to enhance the representation of code-mixed words within sentences. The transliteration phase is fundamental in improving the model's ability to recognize contextual meaning, as converting words into a standardized script mitigates ambiguity in sentiment classification (Shekhar, 2023). The model demonstrates superior performance compared to BiLSTM, CRF, and SVM, emphasizing the significance of transliteration in refining sentiment analysis models for multilingual digital content. Nonetheless, the informal and dynamic nature of language in social media presents challenges in establishing adaptable transliteration rules, necessitating further advancements in rule-based and data-driven transliteration frameworks.

A neural network-based approach incorporating transliteration tools has been proposed for recognizing transliterated English words in Persian text. This model integrates Google Translate and Behnevis for transliteration while employing neural networks for word detection. Transliteration addresses discrepancies in word representation, aligning English and Persian orthographies to improve the identification of low-frequency words using contextual and non-contextual information (Hoseinmardy and Momtazi, 2021). The study demonstrates that transliteration substantially enhances word detection accuracy in bilingual texts. Nevertheless, the reliance on external transliteration tools introduces inconsistencies, highlighting the necessity of developing more robust and internally optimized transliteration frameworks to ensure consistency and accuracy.

A deep learning approach utilizing weak supervision and transliteration alignment has been introduced in detecting cuneiform signs. This method employs iterative training and weak annotation generation from transliterated texts, effectively enhancing the detection of ancient scripts without requiring extensive manual labeling. This approach bridges historical linguistic gaps by aligning cuneiform characters with their modern transliterations, improving model training efficiency. The methodology effectively reduces dependence on manually annotated datasets, making it a scalable solution for historical text recognition. However, the inherent variability in cuneiform script structures presents significant challenges, necessitating adaptive strategies to ensure consistent alignment across different periods and regional variations (Dencker, 2020).

Table 13
Comparative overview of transliteration methods in text recognition and detection.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Text Recognition & Detection
HLSTM for Hate Speech and Trolling Detection Shekhar (2023)	HLSTM	Hate speech and trolling detection in code-mixed social media text	- Utilizes word and character embeddings for code-mixed text representation - Transliteration enhances context recognition for sentiment classification	- Ambiguous word identification in code-mixed language - Need for improved detection of the intent behind hate speech	- Achieves superior accuracy compared to other ML models - Effectively detects language and intent behind hateful words
Neural Network and Tool-Based English-Persian Transliteration Hoseinmardy and Momtazi (2021)	Neural network and tool-based method	Recognition of transliterated English words in Persian text	- Integrates Google Translate and Behnevis for transliteration - Uses neural networks for word detection	- Word mismatch issues - Low-frequency word identification	- Enhances precision in detecting transliterated English words in Persian texts
Deep Learning with Weak Supervision for Cuneiform Sign Detection Dencker (2020)	Deep learning with weak supervision and transliteration alignment	Cuneiform sign detection	- Uses iterative training and weak supervision - Generates weak annotations from text transliterations	- Lack of manual sign annotations - Challenges in aligning ancient scripts with modern text	- Improves cuneiform sign detection accuracy - Reduces dependence on manual annotations
LSTM-based OCRopy for Palm Leaf Manuscript Recognition Kesiman et al. (2018)	LSTM-based OCRopy architecture	Text recognition in palm leaf manuscripts (Balinese, Khmer, Sundanese)	- Evaluate binarization, layout analysis, segmentation, and writer identification - Uses transliteration for OCR processing	- Struggles with non-Balinese scripts - Requires more synthetic data for training	- Improves OCR processing for Balinese scripts - Highlights the need for improved synthetic training data
Hybrid Deep Neural Model for Duplicate Question Detection Rani et al. (2021)	Hybrid deep neural model (SNN + MLP)	Duplicate question detection in transliterated bi-lingual data	- Converts bilingual transliterated questions into monolingual English - Uses Google Transliteration for preprocessing	- Ensuring semantic similarity across languages - Dependence on transliteration accuracy	- Improves duplicate question detection across languages - Enhances performance in multilingual query systems
Deep Learning and Transformers for Fake Review Detection in Bengali Shahariar et al. (2024)	Deep learning and pre-trained transformers	Fake Review Detection in Bengali	- Translates English to Bengali and transliterates Romanized Bengali - Applies augmentation and preprocessing for better detection	- Handling variations in transliteration accuracy - Ensuring reliable text augmentation	- Enhances fake review detection in multilingual contexts - Strengthens robustness of sentiment classification models

A study focusing on document image analysis for Southeast Asian palm leaf manuscripts examines using an LSTM-based OCRopy architecture for transliteration. The research evaluates binarization, layout analysis, and segmentation techniques for Balinese, Khmer, and Sundanese scripts. Transliteration facilitates OCR processing by converting complex scripts into more standardized formats that can be processed more effectively (Kesiman et al., 2018). While this method enhances OCR performance for Balinese manuscripts, it exhibits limitations when applied to non-Balinese scripts, necessitating the generation of synthetic training data to enhance model generalization across diverse Southeast Asian languages. These findings underscore the crucial role of transliteration in historical manuscript digitization, where script preservation and recognition remain critical challenges.

A hybrid deep neural model combining a Siamese Neural Network (SNN) and a Multi-Layer Perceptron (MLP) has been introduced to detect duplicate questions in transliterated bilingual data. The model first transliterates bilingual questions into monolingual English using Google Transliteration before applying deep learning techniques to identify semantic similarity (Rani et al., 2021). This approach improves the identification of duplicated queries across languages, demonstrating the effectiveness of transliteration in enhancing cross-linguistic text normalization. Nonetheless, the accuracy of transliteration remains a determining factor in overall system performance, as incorrect transliteration mappings can lead to misclassification, highlighting the importance of precise transliteration in multilingual NLP applications.

Fake review detection in Bengali utilizes a deep learning-based approach integrating transliteration and text augmentation. English words are first translated into Bengali, and Romanized Bengali words are transliterated back into Bengali script to ensure linguistic fidelity. Augmented datasets are subsequently processed using pre-trained transformers to improve classification accuracy (Shahariar et al., 2024). The incorporation of transliteration preserves sentiment-related textual cues, which is particularly relevant for dialect-rich languages. However, selecting optimal transliteration variants remains a critical issue, as inconsistencies in transliterated forms may impact classification reliability, necessitating further refinements in transliteration standardization for sentiment analysis applications.

Transliteration methods have proven instrumental in enhancing text recognition and detection capabilities across various NLP applications. They facilitate the accurate representation and classification of multilingual and code-mixed texts, which is crucial for improving the performance of NLP models. These studies collectively highlight the importance of transliteration in addressing challenges associated with language variation, low-resource languages, and complex script recognition. Future research should continue to explore and optimize these methods to further advance the text recognition and detection field, particularly in the context of diverse and multilingual datasets.

D. Speech and Text Processing

Transliteration plays a vital role in speech and text processing within NLP, enabling computational models to handle diverse scripts more efficiently. As multilingual data expands, transliteration enhances speech synthesis, text summarization, and information retrieval by preserving phonetic structures across orthographic systems. Unlike direct translation, transliteration retains phonetic integrity, accurately representing linguistic nuances. Such phonetic integrity is crucial in domains where phonetic consistency influences system performance, such as text-to-speech (TTS) processing, search retrieval, and automatic text summarization. Understanding the diverse applications of transliteration is essential for optimizing NLP systems to handle linguistic variability effectively.

To provide a clearer perspective on the role of transliteration in NLP, Table 14 presents a structured comparative overview of various transliteration techniques, outlining their methodologies, applications, key features, challenges, and impact. These techniques support a range of NLP tasks, from improving phonetic accuracy in speech synthesis to enhancing retrieval efficiency in mixed-script environments. While some approaches prioritize preserving linguistic fidelity, others prioritize optimizing computational scalability, highlighting the trade-offs involved in their implementation. Analyzing these methods reveals their strengths and limitations, offering valuable insights into their effectiveness in addressing speech synthesis, text recognition, and information retrieval challenges. The following sections examine specific transliteration strategies, illustrating their contributions to advancing multilingual text processing.

One notable advancement in TTS systems is the development of a Tamil TTS system employing a phonemic concatenation approach, where transliteration plays a pivotal role in bridging non-Latin scripts with speech synthesis engines. Instead of generating speech directly from Tamil script, which presents challenges due to its complex orthographic structure, the system converts Tamil text into Latin characters (English), reducing computational complexity while ensuring phonetic accuracy. This transliteration step facilitates better text recognition and enhances the intelligibility of synthesized speech, particularly for vocally impaired users (Natarajan et al., 2023). Adopting this method demonstrates the broader potential of transliteration in speech technology as a linguistic transformation tool and a means of optimizing computational efficiency. However, extending this approach to other languages with rich phonetic variations may require more sophisticated mapping techniques to preserve phonemic integrity while ensuring compatibility with speech synthesis models.

Besides its role in TTS, transliteration is critical in automatic text summarization, particularly in multilingual contexts. The Hybrid Neural Text Summarization (HNTSumm) algorithm exemplifies this by leveraging word embeddings trained on transliterated text to generate summaries for news articles containing mixed-script data. By employing a Neural Embedding Model (NEM) and a hybrid Seq2seq model with an attention mechanism, HNTSumm effectively integrates transliteration to enhance the semantic understanding of transliterated words, thereby improving summary coherence and relevance (Muniraj et al., 2023). This integration highlights how transliteration supports cross-linguistic data processing, making it an indispensable tool for handling non-standardized script usage in digital communication. However, one limitation of this approach lies in its dependence on high-quality transliteration models; inconsistencies in script conversion could introduce semantic drift, potentially affecting summary accuracy.

In cross-language information retrieval, transliteration facilitates semantic alignment between languages by converting search queries into a script compatible with indexed text. A study on poetry retrieval demonstrated this by transliterating Romanized Marathi queries into Devanagari script, enabling more accurate semantic matching through an adapted cosine semiotic similarity algorithm (Jadhav, 2023). Transliteration in this context bridges orthographic gaps, ensuring that phonetic similarities between different scripts are accurately captured in retrieval models. This method proves particularly beneficial in literary and cultural datasets, where precise phonetic representation is crucial. However, challenges remain in refining phonetic transcription rules to minimize ambiguity, particularly for languages with complex syllabic structures.

Similarly, transliteration has shown promise in improving query expansion for information retrieval in mixed-script environments. Developing the Hindex and Weighted Hindex algorithms addresses the limitations of conventional phonetic matching systems like Soundex, which are primarily designed for English (Prabhakar et al., 2021). By generating phonetic codes tailored for Hindi transliteration, this method enhances the retrieval of transliterated search terms, significantly improving the accuracy of search results in multilingual databases. Additionally, incorporating back-transliteration (converting Romanized Hindi back into Devanagari) further refines query matching, demonstrating the versatility of transliteration in optimizing search functionality. However, a key challenge lies in balancing transliteration accuracy with computational efficiency, as extensive phonetic mapping may introduce latency in large-scale retrieval systems.

The diverse applications of transliteration in NLP, from improving speech intelligibility in TTS systems to enhancing retrieval accuracy in mixed-script environments, highlight its adaptability across multiple computational tasks. Expanding on its role in current NLP applications, transliteration remains crucial in enhancing script normalization, improving text representation, and refining retrieval accuracy. Its integration into computational models bridges orthographic differences and strengthens model interpretability by preserving phonetic integrity across languages. As multilingual data continues to expand, ensuring that transliteration techniques remain adaptable to diverse linguistic contexts will be essential for maintaining the efficiency and scalability of NLP systems.

Advancing transliteration in speech and text processing requires the development of adaptive phonetic models, robust neural-based transliteration frameworks, and refined phoneme alignment techniques. Such improvements will enable more precise handling of linguistic variations, enhance low-resource language processing, and ensure seamless integration into NLP applications. Refining these methods will also be essential in addressing script diversity while preserving linguistic precision. Unlike direct translation, transliteration preserves phonetic structures, ensuring linguistic nuances are maintained across orthographic systems. With ongoing developments, NLP applications will become more accurate, accessible, and robust across diverse languages and scripts.

E. Language Analysis and Identification

The role of transliteration in language analysis and identification has become increasingly significant as multilingual digital content continues to expand. Transliteration enables computational models to process linguistic variations across different scripts, facilitating improved text classification, information retrieval, and phonetic alignment. Unlike direct translation, which often alters semantic meaning, transliteration preserves phonetic properties, ensuring a more accurate representation of linguistic structures. This characteristic is particularly valuable in loanword identification, code-mixed text processing, and multilingual NLP tasks, where maintaining phonetic integrity is crucial. Given these advantages, recent studies have explored diverse

Table 14
Comparison of transliteration-based approaches in speech and text processing.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Speech and Text Processing
TTS System [65]	Phonemic concatenation with transliteration	Tamil text-to-speech conversion for vocally impaired individuals	- Converts Tamil text to Latin script before speech synthesis - Simplifies phoneme processing for intelligible speech	- Maintaining phonetic accuracy - Handling limited vocabulary scenarios	- Enhances speech intelligibility for Tamil TTS systems - Improves accessibility for vocally impaired users
HNTSumm Algorithm for Text Summarization Muniraj et al. (2023)	NEM and Seq2seq with attention	Automatic summarization of transliterated news articles	- Learns embeddings for transliterated words - Generates summaries using hybrid extractive–abstractive techniques	- Handling transliterated word variations - Processing large-scale multilingual datasets	- Improves summarization accuracy for multilingual text - Enhances coherence and relevance of generated summaries
Cross-Language Information Retrieval for Poetry Jadhav (2023)	Modified cosine similarity and customized transcription	Retrieval of Marathi poetry using Roman script queries	- Converts Roman script queries to Marathi - Uses cosine semiotic similarity for retrieval	- Ensuring phonetic and semantic accuracy - Handling diverse poetic styles	- Bridges linguistic gaps in poetry retrieval - Enhances cross-language retrieval accuracy
Hindex & Weighted Hindex for Query Expansion Prabhakar et al. (2021)	Customized phonetic matching for transliterated Hindi terms	Search query expansion in mixed-script domains	- Develop new phonetic codes suited to Hindi - Integrates back-transliteration into search queries	- Addressing spelling variations in transliterated terms - Improving recall and precision in search results	- Enhances search relevance for Hindi queries - Effectively handles mixed-script retrieval scenarios

transliteration-based methodologies to enhance the accuracy and efficiency of language identification models.

Recent advancements in computational linguistics have increasingly leveraged transliteration techniques to improve language analysis and identification. These methods are pivotal in refining classification accuracy, distinguishing native and foreign words, and processing code-mixed linguistic data.

Table 15 presents a structured comparative overview of transliteration-based methods in language analysis and identification. It outlines each method's methodology, applications, key features, challenges, and impact, providing insights into their role in multilingual text processing. Some techniques emphasize phonetic fidelity to improve classification accuracy, while others address complex orthographic variations in mixed-script datasets. Understanding these approaches is crucial for exploring transliteration's role in loanword detection, language segmentation, and code-mixed text classification.

One innovative approach utilizes an unsupervised character-based n-gram classifier to detect loanwords and transliterated foreign terms within Korean text. The method employs a Bayesian classifier with two distinct character n-gram models: one representing native vocabulary and the other foreign-origin terms. A grapheme-to-phoneme conversion process enhances detection accuracy, enabling the classifier to capture phonetic characteristics of embedded foreign words (Koo, 2015). A key advantage of this technique is its minimal reliance on extensive linguistic resources, requiring only a monolingual corpus and a lexicon for support. Moreover, the model demonstrates adaptability for other languages with similar phonetic transformation processes, such as Japanese, underscoring its broader potential for multilingual language analysis.

Another sophisticated method integrates multichannel neural networks, combining CNN and BiLSTM models to identify languages at the word level in code-mixed text. Focused on English and Romanized Hindi, this approach utilizes a transliteration module to convert Roman-script Hindi words into phonetic equivalents, allowing the neural network to more effectively classify linguistic components (Shekhar et al., 2020). The inclusion of transliteration is particularly significant in handling mixed-script data, where overlapping linguistic structures complicate conventional language identification. By aligning Roman Hindi words with their original phonetic forms, this method enhances segmentation, boosts classification accuracy, and strengthens NLP models in processing social media discourse. Given the prevalence of multilingual digital communication, transliteration emerges as a fundamental component in enhancing model adaptability to evolving linguistic patterns.

Beyond individual methodologies, transliteration continues to be a crucial mechanism in advancing language analysis and identification. By transforming non-Latin scripts into standardized formats, transliteration facilitates more efficient multilingual text processing, enabling classifiers and neural networks to operate more effectively within complex linguistic environments.

As transliteration-based approaches gain prominence, future research must focus on refining existing methodologies to improve their adaptability across low-resource languages and highly dynamic linguistic contexts. Integrating self-learning phonetic models, transformer-based transliteration frameworks, and multilingual transfer learning could further enhance model performance in distinguishing complex language patterns. Optimizing cross-lingual transliteration pipelines would allow more accurate language identification in social media

Table 15
Comparative analysis of transliteration-based methods for language analysis and identification.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Language Analysis & Identification
Unsupervised Character-Based n-Gram Classifier Koo (2015)	Bayesian classifier with character n-gram models	Identification of loanwords and transliterated foreign words in Korean text	- Utilizes a grapheme-to-phoneme converter for phonetic representation - Requires only a large monolingual corpus and a lexicon	- Handling variations in loanword phonetics - Applicability to other languages with vowel insertion	- Enhances distinction between native and foreign words - Improves classification efficiency without extensive resources
Multichannel Neural Network (CNN + BiLSTM) Shekhar et al. (2020)	Combination of CNN and BiLSTM	Word-level language identification in code-mixed social media text (English and Roman Hindi)	- Converts Hindi words in Roman script to phonetic equivalents - Processes mixed-language data effectively	- Managing variations in Romanized Hindi spelling - Handling highly intermingled multilingual text	- Increases classification accuracy of code-mixed texts - Strengthens the robustness of language identification systems

analytics, digital archiving, and linguistic forensics domains. Expanding these capabilities will ensure that transliteration remains a key enabler of advanced NLP systems, bridging linguistic gaps and reinforcing the accuracy of language processing technologies.

F. Transliteration Analyzing and Word Normalization

Transliteration analysis and word normalization play fundamental roles in improving the accuracy and efficiency of NLP systems. The need for robust transliteration models becomes increasingly crucial as languages evolve, particularly in digital and social media contexts. The complexity of transliteration arises from variations in phonetics, orthography, and writing conventions across different languages. While rule-based approaches offer interpretability, statistical and deep learning-based techniques provide adaptability and precision. Integrating linguistic knowledge with computational methods has become key to advancing transliteration accuracy. Understanding the advantages and limitations of different transliteration techniques is essential for developing both linguistically sound and computationally efficient systems.

A comparative analysis of various transliteration-based methods highlights the diversity of language analysis and identification approaches. Table 16 outlines that these methods range from rule-based systems, such as character-mapping techniques and phonetic conversion, to more advanced deep learning models, including convolutional and recurrent neural networks. While rule-based techniques, such as Bigram models and HMM, exhibit strong performance in controlled environments, they often struggle with linguistic variability. Deep learning models have demonstrated superior adaptability in handling complex transliteration patterns. Among these models, bidirectional LSTM and transformer-based approaches are particularly effective, especially for informal and code-mixed texts. The comparative analysis underscores the necessity of selecting appropriate transliteration methods based on each application's specific linguistic challenges and computational constraints. These findings inform model selection for NLP tasks and highlight the ongoing need for refinement and hybridization of transliteration approaches. Understanding these methodological distinctions provides a foundation for exploring their practical applications, particularly in text normalization, phonetic consistency, and adaptation to informal writing styles.

A study on Punjabi text normalization from Roman to Gurumukhi script proposes a rule-based method integrating a mapping table, bigram model, and HMM for selecting the most probable transliteration (Kaur and Singh, 2015b). This approach significantly enhances the processing of spelling variations, phonetic similarities, and text shortening, which are common challenges in social media texts. Integrating a bigram model enhances contextual accuracy, ensuring word normalization aligns with natural language usage. However, while the method outperforms traditional rule-based approaches, it still faces challenges adapting to less structured linguistic variations. These limitations indicate that future refinements should explore hybrid models incorporating statistical and deep learning techniques for greater robustness.

Similarly, research on Arabic-English transliteration highlights the complexities arising from linguistic ambiguity and homophones, which pose significant challenges for machine translation tools (Alharbi, 2024). A mixed-methods study, combining qualitative interviews with human translators and quantitative analysis of transliteration accuracy, emphasizes the superiority of human expertise in managing subtle linguistic nuances. While machine-based transliteration models offer scalability, their struggle with ambiguous phonetic variations demonstrates the irreplaceable role of human intervention in ensuring transliteration reliability. This study reinforces the importance of hybrid approaches, where human expertise complements automated tools to enhance transliteration effectiveness, particularly in high-stakes translation contexts.

The accuracy of transliteration methods is further examined in a study on Latin-to-Balinese script transliteration using the mobile application Transliterasi Aksara Bali (TAB) (Jampel et al., 2018). The method involves analyzing certain documents of Balinese script writing rules and examples to evaluate the accuracy of the transliteration process. The results show that while the TAB mobile application effectively transliterates Latin text to Balinese script, there are areas for improvement, particularly in handling special words that require unique treatment. The study suggests that the transliteration method used in TAB is effective, but further enhancements are required to address specific linguistic intricacies. These findings emphasize the role of transliteration in preserving linguistic accuracy and the need for continual refinement of transliteration tools.

In Javanese manuscript transliteration, a study addresses the scriptio continua problem, where texts are written without spaces between words (Widiarti, 2020). The proposed methods use a backtracking algorithm and depth-first search to segment contiguous syllables into words, utilizing both forward and backward approaches. The transliteration method employed here is significant for its ability to accurately concatenate syllables into words, demonstrating the effectiveness of using advanced algorithms to solve complex transliteration

Table 16
Comparative analysis of transliteration-based methods for transliteration analyzing and word normalization.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Transliteration Analyzing and Word Normalization
Punjabi Transliteration Normalization Kaur and Singh (2015b)	Rule-based approach with Bigram model and HMM	Normalization of Punjabi text from Roman to Gurumukhi in social media	- Uses a mapping table for Roman-to-Gurmukhi conversion - Generates multiple Gurumukhi combinations - Selects the most probable word and sentence	- Handling different spellings of the same word - Managing phonetic similarities and text shortening	- Improves transliteration accuracy for informal social media text - Enhances the normalization process for multilingual content
Arabic-to-English Transliteration Challenges Alharbi (2024)	Mixed methods (qualitative interviews and quantitative analysis)	Analysis of transliteration issues in Arabic-English translation	- Identifies ambiguity markers and homophones - Compares manual and machine-based transliteration	- Machine translation struggles with linguistic nuances - Homophones increase translation complexity	- Highlights the importance of human transliterators - Emphasizes linguistic adaptability in formal translation
Latin-to-Balinese Transliteration Using TAB Mobile App Jampel et al. (2018)	Evaluation of transliteration accuracy	Balinese script transliteration through mobile applications	- Analyzes Balinese script writing rules - Evaluate mobile app transliteration effectiveness	- Struggles with special words requiring unique treatment	- Highlights the need for continual improvement in transliteration tools - Supports indigenous script preservation
Javanese Manuscript Transliteration (Scriptio Continua) Widiarti (2020)	Backtracking algorithm and depth-first search	Javanese script transliteration with word segmentation	- Segments syllables into words using forward and backward approaches - Accurately represents historical manuscripts	- Handling word segmentation without spaces	- Demonstrates the effectiveness of computational algorithms for transliteration - Improves readability of historical texts
N-Gram Models for Transliteration Detection Nabende (2015)	N-gram correspondence models	Detection of transliterated words across languages	- Higher-order n-gram models (4-gram and above) improve accuracy - Captures more contextual information	- Model complexity varies depending on the language and writing system	- Enhances transliteration detection accuracy - Increases efficiency in multilingual text processing
Statistical Models for Transliteration Mining Sajjad et al. (2017)	Supervised, semi-supervised, and unsupervised statistical models	Mining transliteration pairs from parallel corpora	- Semi-supervised approach effectively learns from minimal labeled data	- Balancing labeled and unlabeled data for optimal results	- Improves transliteration accuracy with limited linguistic resources - Supports cross-language transliteration
Comparative Analysis of Latin-to-Balinese Transliteration (BAB vs. TAB) Indrawan et al. (2018a)	Mobile application comparison	Balinese script transliteration	- TAB outperforms BAB inaccuracy - Continuous evaluation of transliteration effectiveness	- Need for ongoing refinement of transliteration models	- Highlights transliteration tools' role in language preservation - Demonstrates comparative efficiency of different models
Transliteration in Multilingual Neural Machine Translation (MNMT) Marton and Zitouni (2014)	Neural machine translation with standardized script (ITRANS)	Machine translation for lexically rich languages (Malayalam, Tamil)	- Standardized script improves translation accuracy	- Ensuring adaptability across diverse languages	- Enhances multilingual translation quality - Improves linguistic consistency in NLP applications
Romanized Assamese Social Media Text Baruah (2024)	Character-level transliteration models (PBSMT, BiLSTM seq2seq with attention, Neural Transformer)	Transliteration variation handling in social media	- Neural Transformer model outperforms other approaches - Handles user-generated informal transliterations	- High variation in transliteration schemes among users	- Demonstrates the adaptability of neural networks for informal text - Strengthens robustness in processing transliterated social media content

issues. This method accurately represents historical manuscripts by carefully segmenting the text into coherent words. The study highlights the importance of combining computational techniques with linguistic knowledge to enhance transliteration accuracy.

The evaluation of n-gram correspondence models for transliteration detection is another area of focus (Nabende, 2015). This study examines different classes of n-gram models, revealing that higher-order models (such as 4-gram and above) significantly improve transliteration detection performance compared to lower-order models. This finding is crucial because higher-order n-gram models can capture more contextual information, leading to better detection and accuracy in transliteration tasks. The study emphasizes the need for appropriate model selection based on the specific language and writing system involved. This finding highlights the impact of model complexity on transliteration effectiveness.

Statistical models for transliteration mining, including unsupervised, semi-supervised, and supervised approaches, are explored in another study (Sajjad et al., 2017). The research demonstrates that semi-supervised systems, which combine unlabeled data with a small amount of labeled data, achieve high precision by learning contextual information. The transliteration method utilized here is instrumental in mining transliteration pairs from parallel corpora, showing that even with minimal labeled data, semi-supervised approaches can significantly enhance transliteration accuracy. This study showcases the potential of statistical models to improve transliteration processes across different languages and scripts.

A comparative analysis of two Latin-to-Balinese script transliteration methods on mobile applications, Belajar Aksara Bali (BAB) and Transliteracy Aksara Bali (TAB), highlights the effectiveness of different transliteration techniques (Indrawan et al., 2018a). The study reveals that the TAB method outperforms the BAB method in accuracy. This comparison underscores the importance of transliteration tools in preserving and promoting indigenous scripts. The results suggest that continuous development and evaluation of transliteration methods are essential for enhancing their accuracy and usability in practical applications.

Transliteration normalization is pivotal in machine translation, particularly for lexically rich languages (Marton and Zitouni, 2014). A study on MNMT shows that transliteration significantly enhances translation accuracy, especially for complex languages like Malayalam and Tamil. The transliteration method used in this research involves converting text to a standardized script (ITRANS), which improves the model's ability to handle linguistic variations. This approach demonstrates the value of transliteration in improving the quality of machine translation for multiple languages, emphasizing the need for standardized transliteration schemes.

The study on transliteration characteristics in Romanized Assamese social media text explores how users do not adhere to fixed Romanization schemes, leading to significant variations in transliterations (Baruah, 2024). The research employs three character-level transliteration models: a traditional Phrase-based Statistical Machine Transliteration model (PBSMT), a BiLSTM neural seq2seq model with attention, and a Neural Transformer model. The findings indicate that the Neural Transformer model outperforms the other models, highlighting the adaptability and effectiveness of advanced neural network approaches in handling informal and variable transliterations on social media platforms. This study emphasizes the need for robust transliteration models to manage user-generated content diversity.

Transliteration analysis and word normalization are essential for improving NLP accuracy, particularly in multilingual and low-resource language contexts. This study highlights the need for transliteration methods that balance linguistic integrity with computational adaptability. While rule-based methods offer interpretability, deep learning techniques such as transformer models enhance adaptability in handling complex transliteration patterns. Beyond improving NLP, transliteration supports text standardization in digital communication, machine translation, and social media analytics. Effective normalization reduces ambiguity, benefiting tasks like sentiment analysis and named entity recognition. However, challenges remain in adapting transliteration frameworks to dynamic linguistic environments, especially for code-mixed and informal texts.

Additionally, transliteration is key to preserving linguistic heritage, as demonstrated in research on Balinese and Javanese scripts. Future studies should integrate textual, phonetic, and image-based data to enhance transliteration accuracy, particularly for endangered scripts with intricate orthographic rules. As methodologies evolve, assessing their real-world effectiveness is crucial. Advancing transliteration techniques requires collaboration between computational linguists, data scientists, and language experts to ensure phonetic and semantic fidelity, promoting linguistic inclusivity in NLP applications.

G. Security

Traditionally seen as a linguistic process, transliteration is increasingly recognized for its role in cybersecurity. As digital communication evolves, safeguarding sensitive data becomes a priority, driving the need for innovative approaches that integrate linguistic techniques with security mechanisms. With its ability to represent one script in another, transliteration presents unique possibilities for secure data encoding, particularly in steganography.

Table 17 provides an overview of how transliteration techniques can be leveraged for data concealment. The table outlines the key components of a transliteration-based steganography method, highlighting its encoding strategy, security advantages, and potential challenges. This approach ensures that the hidden message remains imperceptible while maintaining linguistic coherence by embedding secret information within transliterated text. The structured analysis demonstrates how transliteration can serve as a means of text representation and an effective tool for secure communication.

One notable example is a steganography method utilizing Bengali transliteration for covert data hiding (Khairullah, 2019). This method exploits the phonetic characteristics of Bengali keyboard layouts, encoding secret messages by mapping different bits to specific transliteration choices. Unlike conventional cryptographic techniques, this approach preserves the readability of the text while embedding hidden data, making it difficult for unauthorized users to detect. The ability to seamlessly integrate concealed information within natural language structures illustrates the broader potential of transliteration-based security methods.

A significant advantage of this approach is its low risk of detection by automated systems, particularly due to its reliance on non-roman script structures. Many standard text-processing tools are optimized for Latin-based alphabets, making identifying hidden patterns in transliterated text challenging. Additionally, the method is adaptable to various non-roman scripts, broadening its applicability in different linguistic contexts. These findings underscore the importance of understanding script-specific transliteration patterns when designing secure communication frameworks.

Beyond its immediate applications, this transliteration-based steganography approach raises important considerations for future cybersecurity developments. The balance between linguistic accuracy and data concealment is crucial in ensuring that secure communication methods remain functional and undetectable. As security threats continue to evolve, exploring hybrid approaches that integrate machine learning with transliteration-based encoding could enhance robustness and adaptability across languages. This perspective further highlights how transliteration is a linguistic tool and a key enabler of advanced security solutions, which will be explored in the following discussion.

Transliteration plays a crucial role in NLP by enhancing multilingual text processing. It facilitates script normalization, improves cross-lingual information retrieval, and supports tasks such as machine translation, sentiment analysis, and named entity recognition. As transliteration becomes more integrated into these applications, its role extends beyond simple character mapping to a crucial component in ensuring consistency and interoperability across different languages and writing systems. However, as NLP models evolve, so do the complexities associated with transliteration, necessitating more sophisticated approaches to handle phonetic and orthographic variations effectively.

Table 17
Analysis of transliteration-based steganography for security applications.

Study/Reference	Approach	Application	Key features	Challenges	Impact on Security
Transliteration-Based Steganography for Bengali Text Khairullah (2019)	Steganography using transliteration of Bengali text	Data hiding and secure communication	- Leverages Bengali phonetic keyboard layouts for encoding secret information - Assign bits to specific transliteration options - Embeds hidden data without altering the text's apparent meaning	- Requires careful selection of transliteration options to maintain readability - Adaptability to other non-Roman scripts may require further refinement	- Reduces risk of machine detection by exploiting non-Roman script structures - Ensures hidden data remains inconspicuous to text processing tools - Enhances security in communication by protecting sensitive data from interception

With transliteration playing an increasingly central role in NLP, its influence extends beyond linguistic processing into broader computational applications. It impacts data representation, supports multilingual model adaptation, and is even applied in security-related tasks, such as encoded text processing and access control in multilingual environments. These diverse applications highlight the growing demand for transliteration models that enhance NLP performance and contribute to more robust and adaptable data management mechanisms.

However, advancing transliteration in these expanding domains presents several challenges, including reliance on predefined mappings, inconsistencies in accuracy, and the need for more adaptable learning mechanisms. High performance across heterogeneous language pairs and scripts depends on improving transliteration methods to accurately reflect contextual and phonetic variation. By integrating transliteration with explainable AI and adaptive learning techniques, NLP models can achieve greater efficiency, scalability, and linguistic precision across diverse digital environments.

5. Conclusion and future direction

This systematic review has explored machine transliteration's diverse methodologies, challenges, and applications in various linguistic contexts. The findings highlight the evolution of transliteration techniques from rule-based systems to statistical and machine-learning approaches, emphasizing their adaptability to different scripts and languages. While significant progress has been made in addressing phonetic and orthographic challenges, the review underscores persistent issues such as data scarcity, linguistic complexity, and the need for standardization across transliteration systems. These challenges particularly affect underrepresented languages, limiting the universality of current methodologies.

The review also demonstrates the critical role of transliteration in supporting broader NLP tasks, including machine translation, sentiment analysis, and language identification. Transliteration facilitates data creation for languages with low-resource settings, enabling advancements in machine-learning applications. Moreover, transliteration proves invaluable in preserving cultural heritage, as evidenced by its application in processing historical scripts and manuscripts. Despite these achievements, transliteration systems often face difficulties handling diverse linguistic nuances, such as *scriptio continua*, and require enhanced contextual adaptability.

This review contributes to a deeper understanding of transliteration's multifaceted applications and its significance in multilingual NLP. The identified limitations and gaps present an opportunity to refine transliteration techniques further, paving the way for more robust and inclusive models that cater to the linguistic diversity of global communities. Despite these opportunities, existing transliteration approaches still face various technical and linguistic challenges that hinder their effectiveness across different scripts and languages.

Existing transliteration methods, including rule-based, statistical, and machine-learning approaches, often face challenges in handling phonetic variations, complex script structures, and low-resource languages. Many of these approaches require extensive language-specific rule engineering or large amounts of parallel training data, which are not always available for underrepresented languages such as Javanese and Balinese. Despite the advancements in hybrid methods combining statistical and neural approaches, limitations remain in addressing orthographic complexity and maintaining high transliteration accuracy across diverse scripts. Javanese presents particularly complex transliteration challenges among these underrepresented languages due to its intricate script structure and linguistic rules.

Future research should address the unique transliteration challenges for the Javanese script, one of the most complex and underrepresented scripts in computational linguistics. Its extensive character set, complex vowel-consonant combinations, and integration of diacritical marks demand innovative approaches that combine linguistic insights with advanced computational models. Addressing these challenges requires sophisticated models capable of handling scriptio continua, where words are written without spaces, adding to the complexity of transliteration. Further exploration should include creating large, annotated datasets of Javanese script to support model training and evaluation and leveraging transfer learning techniques such as multi-script training that utilizes similarities with Balinese and Sundanese. Additionally, incorporating hybrid approaches that integrate rule-based and neural methods may solve the script's phonological and orthographic challenges, ensuring phonetic and semantic fidelity in transliteration. Given these persistent challenges, exploring more advanced computational approaches that can enhance transliteration accuracy and adaptability across diverse scripts is essential.

To overcome these challenges, recent advancements in deep learning offer promising directions for improving transliteration performance. Transformer-based architectures and large language models (LLMs) have significantly enhanced various NLP tasks, including transliteration. Unlike traditional rule-based or statistical approaches, these models utilize self-attention mechanisms to capture long-range dependencies and contextual relationships between characters and words. This capability allows them to handle complex script conversions while preserving phonetic accuracy and linguistic structure. Transformer models, such as BERT, GPT, and mT5, have demonstrated superior adaptability in multilingual tasks, while LLMs like GPT-4 and LLaMA can generate transliterations with high contextual accuracy. These advancements highlight the potential of deep learning-based transliteration in overcoming challenges that traditional methods struggle with, such as handling phonetic variations and script complexity. Evaluating the effectiveness of Transformer-based models and LLMs through comparative studies with traditional methods will be crucial, particularly by benchmarking these models with diverse linguistic datasets,

including low-resource scripts like Javanese and Balinese. Moreover, integrating multimodal approaches, combining textual and image-based data, could enhance transliteration for historical manuscripts, where handwritten script variations pose significant challenges. Developing high-quality annotated transliteration datasets is essential to improving model training and evaluation. These efforts will advance transliteration research and facilitate the digital preservation of endangered scripts, contributing to their accessibility in linguistic and cultural studies.

CRediT authorship contribution statement

A'la Syauqi: Writing – original draft, Data curation, Conceptualization. **Aji Prasetya Wibawa:** Supervision, Investigation, Formal analysis.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Microsoft Copilot to improve the work's readability and language. After using this tool/service, the proofreader reviewed and the author(s) edited the content as needed and take(s) full responsibility for the publication's content.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Covidence. [Online]. Available: https://www.covidence.org/.

Abbas, M.R., 2020. Punjabi to ISO 15919 and roman transliteration with phonetic rectification. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 19 (2), http://dx.doi.org/10.1145/3359991.

Abebaw, Z., Rauber, A., Atnafu, S., 2023. Transliterating Latin to Amharic scripts using user-defined rules and character mappings. Int. J. Digit. Libr. 24 (1), 63–75. http://dx.doi.org/10.1007/s00799-023-00346-5.

Ahmadi, S., 2019. A rule-based Kurdish text transliteration system. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 18 (2), http://dx.doi.org/10.1145/3278623.

Alam, M., 2021. Deep learning-based Roman-Urdu to Urdu transliteration. Int. J. Pattern Recognit. Artif. Intell. 35 (4), http://dx.doi.org/10.1142/S0218001421520017.

Alginahi, Y., Al Binali, A.M., Dekkak, M., Kushk, A., 2018. A computerized reversible Arabic transliteration system. Arab. J. Sci. Eng. 43 (2), 759–776. http://dx.doi.org/10.1007/s13369-017-2737-2.

Alharbi, M.A., 2024. Transliteration of Arabic words/phrase into English: An exploration of ambiguity markers. World J. Engl. Lang. 14 (4), 404–410. http://dx.doi.org/10.5430/wjel.v14n4p404.

Anbukkarasi, S., 2023. Phonetic-based forward online transliteration tool from English to Tamil language. Int. J. Reliab. Qual. Saf. Eng. 30 (3), http://dx.doi.org/10.1142/S021853932350002X.

Ansari, M.Z., 2022. Hindi to English transliteration using multilayer gated recurrent units. Indones. J. Electr. Eng. Comput. Sci. 27 (2), 1083–1090. http://dx.doi.org/10.11591/ijeecs.v27.i2.pp1083-1090.

Bala Das, S., Panda, D., Kumar Mishra, T., Kr. Patra, B., Ekbal, A., 2024. Multilingual neural machine translation for Indic to Indic languages. ACM Trans. Asian Low Res. Lang. Inf. Process. 23 (5), http://dx.doi.org/10.1145/3652026.

Baruah, H., 2024. Transliteration characteristics in romanized Assamese language social media text and machine transliteration. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 23 (2), http://dx.doi.org/10.1145/3639565.

Biadgligne, Y., 2023. Baseline transliteration corpus for improved English-Amharic machine translation. Inform. (Slovenia) 47 (6), 191–202. http://dx.doi.org/10.31449/inf.v47i6.4395.

Chatterjee, S., 2021. Machine transliteration using SVM and HMM. Int. J. Adv. Intell. Parad. 19 (1), 3–27. http://dx.doi.org/10.1504/IJAIP.2021.114584.

Chen, M., Wang, Y., Yang, H., 2023. Review of machine translation. In: 2023 2nd International Conference on Artificial Intelligence and Computer Information Technology. AICIT, IEEE, Yichang, China, pp. 1–4. http://dx.doi.org/10.1109/AICIT59054.2023.10277892.

Ciubotaru, C., 2021. Backtracking algorithm for lexicon generation. Comput. Sci. J. Mold. 29 (1), 135–152, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108561320&partnerID=40&md5=e2bdedbd590c0549c7078b241cb8b54c.

Danilov, A.V., Salekhova, L.L., Khakimov, B.E., 2016. The design of Cyrillic-Latin converter for Tatar language. J. Lang. Lit. 7 (2), 275–279. http://dx.doi.org/10.7813/jll.2016/7-2/52.

Dencker, T., 2020. Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. PLoS ONE 15 (12), http://dx.doi.org/10.1371/journal.pone.0243039.

Dinh, D., Nguyen, P., Nguyen, L.H.B., 2021. Transliterating Nôm scripts into Vietnamese national scripts using statistical machine translation. Intl. J. Adv. Comput. Sci. Appl. 12 (2), 37–45. http://dx.doi.org/10.14569/IJACSA.2021.0120205.

Garg, K.D., 2019. Hidden Markov model based Punjabi to English machine transliteration system. Int. J. Control. Autom. 12 (4), 199–206, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85082398921&origin=inward.

Goyal, K.D., 2022. Forward-backward transliteration of Punjabi Gurmukhi script using N-gram language model. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 22 (2), http://dx.doi.org/10.1145/3542924.

Guellil, I., et al., 2021. A semi-supervised approach for sentiment analysis of Arab(ic+izi) messages: Application to the Algerian dialect. SN Comput. Sci. 2 (2), http://dx.doi.org/10.1007/s42979-021-00510-1.

Hajbi, S., 2024. Moroccan Arabizi-to-Arabic conversion using rule-based transliteration and weighted Levenshtein algorithm. Sci. Afr. 23 (Query date: 2024-10-03 09:44:22), http://dx.doi.org/10.1016/j.sciaf.2024.e02073.

Hoseinmardy, A., Momtazi, S., 2021. Recognizing transliterated English words in Persian texts. J. Inf. Syst. Telecommun. 8 (30), 84–92. http://dx.doi.org/10.29252/jist.8.30.84.

Indrawan, G., 2020. Handling of line breaking on Latin-to-Balinese script transliteration web application as part of Balinese language ubiquitous learning. In: 2020 6th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Disaster Management, ICSITech 2020, No. Query Date: 2024-10-03 09:28:50. pp. 40–44. http://dx.doi.org/10.1109/ICSITech49800.2020.9392035.

Indrawan, G., 2021. A method to accommodate backward compatibility on the learning application-based transliteration to the Balinese script. Int. J. Adv. Comput. Sci. Appl. 12 (6), 280–286. http://dx.doi.org/10.14569/IJACSA.2021.0120631.

Indrawan, G., Paramarta, I.K., Agustini, K., Sariyasa, S., 2018a. Latin-to-Balinese script transliteration method on mobile application: A comparison. IJEECS 10 (3), 1331. http://dx.doi.org/10.11591/ijeecs.v10.i3.pp1331-1342.

Indrawan, G., Puspita, N.N.H., Paramarta, I.K., Sariyasa, S., 2018b. LBtrans-bot: A Latin-to-Balinese script transliteration robotic system based on Noto Sans Balinese font. IJEECS 12 (3), 1247. http://dx.doi.org/10.11591/ijeecs.v12.i3.pp1247-1256.

Islam, S., 2019. Bodo to English machine translation through transliteration. Int. J. Innov. Technol. Explor. Eng. 8 (12), 4825–4830. http://dx.doi.org/10.35940/ijitee.L3695.1081219.

Jadhav, R.S., 2023. Cross-language information retrieval for poetry form of literature-based on machine transliteration using CNN. J. Intell. Fuzzy Syst. 45 (2), 3025–3037. http://dx.doi.org/10.3233/JIFS-223591.

Jaf, A.A., Kayhan, S.K., 2021. Machine-based transliterate of Ottoman to Latin-based script. Sci. Program. 2021, http://dx.doi.org/10.1155/2021/7152935.

Jampel, I.N., Indrawan, G., Widiana, I.W., 2018. Accuracy analysis of Latin-to-Balinese script transliteration method. Int. J. Electr. Comput. Eng. 8 (3), 1788–1797. http://dx.doi.org/10.11591/ijece.v8i3.pp1788-1797.

Jha, V., 2019. Braille transliteration of hindi handwritten texts using machine learning for character recognition. Int. J. Sci. Technol. Res. 8 (10), 1188–1193, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85074032132&origin=inward.

Jha, A., 2023. A review of machine transliteration, translation, evaluation metrics and datasets in Indian languages. Multimedia Tools Appl. 82 (15), 23509–23540. http://dx.doi.org/10.1007/s11042-022-14273-1.

Joseph, S., 2024. Machine transliteration of handwritten MODI script to Devanagari using deep neural networks. Int. J. Comput. 23 (2), 219–225. http://dx.doi.org/10.47839/ijc.23.2.3540.

Karakanta, A., Dehdari, J., van Genabith, J., 2018. Neural machine translation for low-resource languages without parallel corpora. Mach. Transl. 32 (1–2), 167–189. http://dx.doi.org/10.1007/s10590-017-9203-5.

Karimi, S., Scholer, F., Turpin, A., 2011. Machine transliteration survey. ACM Comput. Surv. 43 (3), 1–46. http://dx.doi.org/10.1145/1922649.1922654.

Karmani, N., 2019. Tunisian Arabic chat alphabet transliteration using probabilistic finite state transducers. Int. Arab. J. Inf. Technol. 16 (2), 295–303, [Online]. Available: https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=85065705403&origin=inward.

Kasparaitis, P., 2023. Automatic transliteration of Polish and English proper nouns into Lithuanian. Inf. Technol. Control. 52 (1), 128–139. http://dx.doi.org/10.5755/j01.itc.52.1.32353.

Kaur, K., Singh, P., 2015a. Multi script text transliteration to Punjabi. Indian J. Sci. Technol. 8 (34), http://dx.doi.org/10.17485/ijst/2015/v8i34/85944.

Kaur, J., Singh, J., 2015b. Toward normalizing romanized Gurumukhi text from social media. Indian J. Sci. Technol. 8 (27), http://dx.doi.org/10.17485/ijst/2015/v8i27/81666.

Kesiman, M.W.A., Burie, J.-C., Ogier, J.-M., Grangé, P., 2017. Knowledge representation and phonological rules for the automatic transliteration of balinese script on palm leaf manuscript. Comput. Sist. 21 (4), 739–747. http://dx.doi.org/10.13053/CyS-21-4-2851.

Kesiman, M.W.A., et al., 2018. Benchmarking of document image analysis tasks for palm leaf manuscripts from southeast Asia. J. Imaging 4 (2), http://dx.doi.org/10.3390/jimaging4020043.

Khairullah, M., 2019. A novel steganography method using transliteration of Bengali text. J. King Saud Univ. - Comput. Inf. Sci. 31 (3), 348–366. http://dx.doi.org/10.1016/j.jksuci.2018.01.008.

Knight, K., Graehl, J., 1998. Machine transliteration. Comput. Linguist. 24 (4), 599–612.

Koo, H., 2015. An unsupervised method for identifying loanwords in Korean. Lang. Resour. Eval. 49 (2), 355–373. http://dx.doi.org/10.1007/s10579-015-9296-5.

Kumbhar, M., 2024. Language identification and transliteration approaches for code-mixed text. J. Eng. Sci. Technol. Rev. 17 (1), 63–70. http://dx.doi.org/10.25103/jestr.171.09.

Laksminy, L.P., 2024. Nyastra: Nyurat aksara transliterasi (a computer design of Balinese transliteration apps). Multidiscip. Sci. J. 7 (1), http://dx.doi.org/10.31893/multiscience.2025035.

Le, N.T., 2019. Low-resource machine transliteration using recurrent neural networks. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 18 (2), http://dx.doi.org/10.1145/3265752.

Li, P., Wang, M., Wang, J., 2021. Named entity translation method based on machine translation lexicon. Neural Comput. Appl. 33 (9), 3977–3985. http://dx.doi.org/10.1007/s00521-020-05509-y.

Liberati, A., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. Ann. Intern. Med. 151 (4), p. W. http://dx.doi.org/10.7326/0003-4819-151-4-200908180-00136.

Liu, L., Finch, A., Utiyama, M., Sumita, E., 2020. Agreement on target-bidirectional recurrent neural networks for sequence-to-sequence learning. J. Artificial Intelligence Res. 67, 581–606. http://dx.doi.org/10.1613/JAIR.1.12008.

Londhe, D., Kumari, A., 2022. Multilingual sentiment analysis using the social eagle-based bidirectional long short-term memory. Int. J. Intell. Eng. Syst. 15 (2), 479–493. http://dx.doi.org/10.22266/ijies2022.0430.43.

Maimaiti, M., Liu, Y., Luan, H., Sun, M., 2019. Multi-round transfer learning for low-resource NMT using multiple high-resource languages. ACM Trans. Asian Low Res. Lang. Inf. Process. 18 (4), http://dx.doi.org/10.1145/3314945.

Mammadzada, S., 2023. A review of existing transliteration approaches and methods. Int. J. Multiling. 20 (3), 1052–1066. http://dx.doi.org/10.1080/14790718.2021.1958821.

Marton, Y., Zitouni, I., 2014. Transliteration normalization for information extraction and machine translation. J. King Saud Univ. - Comput. Inf. Sci. 26 (4), 379–387. http://dx.doi.org/10.1016/j.jksuci.2014.06.011.

Masmoudi, A., 2019. Transliteration of Arabizi into Arabic script for Tunisian dialect. ACM Trans. Asian Low- Resour. Lang. Inf. Process. 19 (2), http://dx.doi.org/10.1145/3364319.

Muniraj, P., Sabarmathi, K.R., Leelavathi, R., S. Balaji, B., 2023. HNTSumm: Hybrid text summarization of transliterated news articles. Int. J. Intell. Netw. 4, 53–61. http://dx.doi.org/10.1016/j.ijin.2023.03.001.

Murat, A., Osman, T., Yang, Y., Zhou, X., Wang, L., Li, X., 2017. Using semantic knowledge in the Uyghur-Chinese person name transliteration. J. Inf. Process. Syst. 13 (4), 716–730. http://dx.doi.org/10.3745/JIPS.02.0065.

Nabende, P., 2015. An evaluation of n-gram correspondence models for transliteration detection. Lect. Notes Electr. Eng. 312, 615–622. http://dx.doi.org/10.1007/978-3-319-06764-3_79.

Natarajan, K., Chelliah, J., Mariyarose, J.V., Andi, S., Venkatachalam, B., Alagarsamy, M., 2023. TTS system for deafened and vocally impaired persons in native language. J. Intell. Fuzzy Syst. 45 (5), 8159–8169. http://dx.doi.org/10.3233/JIFS-231680.

Ngo, G.H., 2019. Phonology-augmented statistical framework for machine transliteration using limited linguistic resources. IEEE/ ACM Trans. Audio Speech Lang. Process. 27 (1), 199–211. http://dx.doi.org/10.1109/TASLP.2018.2875269.

Prabhakar, D.K., Pal, S., 2018. Machine transliteration and transliterated text retrieval: a survey. Sadhana 43 (6), http://dx.doi.org/10.1007/s12046-018-0828-8.

Prabhakar, D.K., Pal, S., Kumar, C., 2021. Query expansion for transliterated text retrieval. ACM Trans. Asian Low Res. Lang. Inf. Process. 20 (4), http://dx.doi.org/10.1145/3447649.

Radjenović, D., Heričko, M., Torkar, R., Živkovič, A., 2013. Software fault prediction metrics: A systematic literature review. Inf. Softw. Technol. 55 (8), 1397–1418. http://dx.doi.org/10.1016/j.infsof.2013.02.009.

Rani, S., Kumar, A., Kumar, N., 2021. Hybrid deep neural model for duplicate question detection in trans-literated bi-lingual data. Recent. Adv. Comput. Sci. Commun. 14 (3), 926–933. http://dx.doi.org/10.2174/2213275912666190710152709.

Raulji, J.K., Saini, J.R., Pal, K., Kotecha, K., 2022. A novel framework for sanskrit-gujarati symbolic machine translation system. Intl. J. Adv. Comput. Sci. Appl. 13 (4), 374–380. http://dx.doi.org/10.14569/IJACSA.2022.0130444.

Sajjad, H., Schmid, H., Fraser, A., Schütze, H., 2017. Statistical models for unsupervised, semi-supervised, and supervised transliteration mining. Comput. Linguist. 43 (2), 350–375. http://dx.doi.org/10.1162/COLI_a_00286.

Sen, D., Garg, K.D., 2015. A survey on machine transliteration systems. Int. J. Appl. Eng. Res. 10 (55), 1646–1649, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84942419235&partnerID=40&md5=4d2d85038cd456bf0fcf6b8895d1bab7.

Shahariar, G.M., Shawon, M.T.R., Shah, F.M., Alam, M.S., Mahbub, M.S., 2024. Bengali fake reviews: A benchmark dataset and detection system. Neurocomputing 592, http://dx.doi.org/10.1016/j.neucom.2024.127732.

Shahroz, M., Mushtaq, M.F., Mehmood, A., Ullah, S., Choi, G.S., 2020. RuTUT: Roman Urdu to Urdu translator based on character substitution rules and unicode mapping. IEEE Access 8, 189823–189841. http://dx.doi.org/10.1109/ACCESS.2020.3031393.

Shekhar, S., 2023. Hatred and trolling detection transliteration framework using hierarchical LSTM in code-mixed social media text. Complex Intell. Syst. 9 (Query date: 2024-10-03 09:44:22), 2813–2826. http://dx.doi.org/10.1007/s40747-021-00487-7.

Shekhar, S., Sharma, D.K., Sufyan Beg, M.M., 2020. An effective bi-LSTM word embedding system for analysis and identification of language in code-mixed social media text in English and Roman Hindi. Comput. Sist. 24 (4), 1415–1427. http://dx.doi.org/10.13053/CYS-24-4-3151.

Taguchi, K., Finch, A., Yamamoto, S., Sumita, E., 2015. Automatic induction of romanization systems from bilingual corpora. IEICE Trans. Inf. Syst. E98D (2), 381–393. http://dx.doi.org/10.1587/transinf.2014EDP7236.

Unterkalmsteiner, M., Gorschek, T., Islam, A.K.M.M., Cheng, Chow Kian, Permadi, R.B., Feldt, R., 2012. Evaluation and measurement of software process improvement—A systematic literature review. IIEEE Trans. Softw. Eng. 38 (2), 398–424. http://dx.doi.org/10.1109/TSE.2011.26.

Vathsala, M.K., Holi, G., 2020. RNN based machine translation and transliteration for Twitter data. Int. J. Speech Technol. 23 (3), 499–504. http://dx.doi.org/10.1007/s10772-020-09724-9.

Vukčević, M.M., 2020. Turns of the centuries. The Transkribus automated tool for recognition, transcription and translation of handwritten historical documents: A German-Serbian case study. Babel 66 (2), 294–310. http://dx.doi.org/10.1075/babel.00159.vuk.

Widiarti, A.R., 2020. A method for solving scriptio continua in Javanese manuscript transliteration. Heliyon 6 (4), http://dx.doi.org/10.1016/j.heliyon.2020.e03827.

Wu, C.K., 2022. Improving low-resource machine transliteration by using 3-way transfer learning. Comput. Speech Lang. 72 (Query date: 2024-10-03 09:41:33), http://dx.doi.org/10.1016/j.csl.2021.101283.

Yadav, M., 2023. Image processing-based transliteration from Hindi to English. Int. J. Perform. Eng. 19 (5), 334–341. http://dx.doi.org/10.23940/ijpe.23.05.p5.334341.

Younes, J., 2022. Romanized Tunisian dialect transliteration using sequence labelling techniques. J. King Saud Univ. - Comput. Inf. Sci. 34 (3), 982–992. http://dx.doi.org/10.1016/j.jksuci.2020.03.008.