INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Hemanta Baruah
Roll Number	:	186155001
Programme of Study	:	Ph.D.
Thesis Title:		Back Transliteration of Romanized Assamese Social Media Texts (Corpus, Analysis and Models)
Name of Thesis Supervisor(s)	:	Prof. Sanasam Ranbir Singh, Prof. Priyankoo Sarmah
Thesis Submitted to the Academic Division	:	Centre for Linguistic Science and Technology, IIT Guwahati
Date of completion of Thesis Viva-Voce Exam	:	10/10/2025
Key words for description of Thesis Work	:	Back-transliteration, Social Media, Transformer, LLM

SHORT ABSTRACT

Natural Language Processing (NLP) research has largely focused on resource-rich languages, leaving low-resource ones like Assamese underrepresented. Assamese language, spoken by millions in northeast India, faces challenges due to its linguistic diversity and lack of standardized resources. This thesis tackles back-transliteration of Romanized Assamese-common on social media platforms like Facebook, YouTube, and Twitter (X)-where informal, noisy, and code-mixed content complicates processing. Transliteration converts text between scripts while preserving phonetics; back-transliteration reverses this process. These tasks are increasingly relevant in multilingual contexts like India. Assamese poses unique difficulties due to inconsistent Romanization, phonetic variation, and orthographic diversity. This work presents a detailed analysis of grapheme-level and phoneme-level variations and introduces a new dataset of 60,312 sentence pairs and 65,614 word pairs from social media. Various transliteration models-including statistical, neural, transformer and LLM-based-are benchmarked, with a focus on word-level vs. sentence-level performance. Results show the importance of phonetic and contextual factors in accuracy. The thesis also demonstrates how back-transliteration improves downstream tasks like sentiment analysis, offering valuable tools and insights for advancing NLP in low-resource languages.