International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

SwitchLang AI: Advanced NLP for Seamless Code-Switching & Multilingual Text Processing

K Swetha Sailaja*[1], Devarakonda Shashikanth*[2], Gardasu Sri Charan*[3], Yash Vinay Chamle*[4]

*[1] Assistant Professor of Department of CSE (AI & ML) Of ACE Engineering College, India.

*[2,3,4] Students of Department of CSE (AI & ML) Of ACE Engineering College, India.

ABSTRACT

The rise of global digital communication has increased code-switching and multilingual interactions, particularly on social media, messaging apps, and online platforms. Traditional NLP systems, often trained on monolingual data, face challenges in effectively handling these complex linguistic patterns. This paper reviews recent progress in Natural Language Processing (NLP) focused on addressing multilingualism and code-switching, emphasizing advancements in models, datasets, and real-world applications. Key approaches include language-agnostic methods, transfer learning, and the use of pre-trained multilingual models such as mBERT, XLM-R, and other transformer-based innovations. These tools enable improved cross-lingual understanding and adaptability. We introduce Switchlang AI, a conceptual framework designed to support seamless multilingual processing and robust performance across diverse linguistic contexts, including both high-resource and low-resource language settings.

Keywords: Code-switching, Multilingual NLP, Pre-trained Language Models, mBERT, XLM-R, Transfer Learning, Language-Agnostic Models, Cross-Lingual Understanding, Natural Language Processing, Switchlang AI.

I.INTRODUCTION

Multilingualism is a linguistic norm in many regions across the globe, and code-switching—the practice of alternating between two or more languages within a single utterance or conversation—has become increasingly common, particularly in digital communications such as social media posts, messaging platforms, and online forums. These mixed-language interactions pose significant challenges for existing Natural Language Processing (NLP) systems, which are typically trained on monolingual, standardized datasets. Such systems struggle to process code-switched inputs due to linguistic complexity, data scarcity, syntactic variability, and the informal nature of multilingual text. To address these challenges, Switchlang AL is introduced as a novel framework grounded in recent advances in multilingual representation learning, pre-trained language modelling, and cross-lingual transfer learning. This paper surveys state-of-the-art techniques and tools developed to handle code-switching and multilingual text, including transformer-based models like mBERT and XLM-R. It also identifies key limitations in current approaches, such as limited support for low-resource languages and lack of context-aware switching mechanisms. Finally, we propose future directions aimed at building more robust, inclusive, and adaptive NLP systems capable of navigating the linguistic diversity found in real-world, multilingual environments.

II.BACKGROUND

The rise of transformer-based models transformed natural language processing by enabling contextual understanding across multilingual datasets, as seen with architectures like mBERT and XLM-R. However, these models, primarily trained on monolingual or standard bilingual corpora, often struggle with code-switched text, where multiple languages intermingle within sentences, due to the lack of explicit code-switching supervision. Low-resource languages exacerbate this challenge, as limited parallel data prevents effective fine-tuning. Inspired by advancements in data generation and weak supervision, recent work in NLP emphasizes synthetic corpus creation and prompt-driven augmentation, as popularized by large language models (LLMs) such as GPT. This project extends these ideas to code-switched translation,

proposing a framework that combines LLM-guided synthetic data generation with mBERT fine-tuning. By addressing the scarcity of annotated code-switched corpora and enabling human-in-the-loop feedback, the system aims to improve translation quality, particularly for multilingual and low-resource environments where traditional datasets and models fall short.

III.LITERATURE REVIEW

1. Title: Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text

Authors: Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan

The paper successfully fine-tunes BERT for multi-label sentiment analysis on code-switched text but reveals several research gaps. There is a notable lack of large, annotated datasets for code-switched multilingual sentiment tasks, limiting broader model training. While mBERT handles multilingual text reasonably well, it does not fully capture complex intra-sentential language mixing patterns. Additionally, the study mainly uses class-weighted loss to address data imbalance, leaving more advanced techniques like synthetic oversampling unexplored. The approach is evaluated primarily on English-Hindi text, raising questions about generalizability to other language pairs. Furthermore, the paper lacks detailed error analysis and explainability for model predictions.

2. Title: Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements

Authors: Ganji Sreeram and Rohit Sinha

The paper explores enhancements to end-to-end (E2E) automatic speech recognition (ASR) systems for Hindi-English code-switching speech. It addresses key challenges such as cross-lingual target confusability, inefficient target-to-word (T2W) transduction, and inadequate context modelling. The authors propose a reduced target set based on acoustic similarity, a context-aware T2W transduction using language and error models, and a novel code-switching identification (CSI) feature to enhance language modelling. Experimental results on the HingCoS corpus show significant improvements in target error rate and word error rate, while reducing computational complexity. The approach is generalizable to other multilingual and low-resource code-switching scenarios.

3. Title: Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services

Authors: Syed Meesam Raza Naqvi, Hassan Aqeel Khan, Ali Raza, Muhammad Ali Tahir, Kamran Javed, and Zubair Saeed

The paper presents a real-time, hybrid Automatic Speech Recognition (ASR) system tailored for recognizing Urdu-English code-mixed street addresses, designed for voice-activated navigation in Pakistan. It addresses challenges like limited labelled data, diverse accents, and code-switching complexity by combining a general Unicode Urdu corpus with a specialized Roman Urdu-English address dataset. Using deep learning (TDNN-LSTM) and traditional GMM-HMM models within the Kaldi framework, the system achieved high accuracy with a Word Error Rate as low as 4.02%. The solution is deployed in the TPLMaps app, marking a major step toward inclusive, language-specific navigation services in low-resource environments.

4. Title: Leveraging Synthetic Data for Improved Manipuri-English Code-Switched ASR

Authors: Naorem Karline Singh, Wangkheimayum Madal, Chingakham Neeta Devi, Hoomexsun Pangsatabam, and Yambem Jina Chanu

The paper introduces a novel approach to improve Manipuri-English code-switched Automatic Speech Recognition (ASR) using synthetic text and audio data. A hybrid model combining Transformer and Pointer Generator Network is proposed to generate realistic code-switched text from parallel monolingual corpora. The study also explores effective language

model training strategies and applies audio augmentation methods like SpecAugment, time-stretching, and speed perturbation. Results show significant improvements in language model perplexity and ASR accuracy, achieving a Word Error Rate (WER) of 26.51%. The work presents valuable contributions for low-resource multilingual ASR and establishes a strong foundation for future real-world speech applications.

5. Title: Code-Switching ASR for Low-Resource Indic Languages: A Hindi-Marathi Case Study

Authors: Hemant Palivela Vinay Rishiwal, Meeranarvekar, David Asirvatham, Shashi Bhushan, and Udit Agarwal

The paper explores the development of robust Automatic Speech Recognition (ASR) systems for Hindi-Marathi code-switching, a common phenomenon in multilingual India. It highlights the limitations of monolingual and bilingual ASR models in handling abrupt language transitions, phonetic overlaps, and dialectal diversity. By comparing different system architectures, the study demonstrates that multilingual ASR models are more effective in managing code-switched speech. It also emphasizes the importance of techniques like transfer learning, data augmentation, and hierarchical language identification to overcome data scarcity and improve recognition accuracy. The findings contribute valuable strategies for advancing ASR in low-resource, linguistically complex environments.

COMPARISON TABLE

S. No	Author Name	Title	Methodology	Findings
1	Tiancheng Tang, Xinhuai Tang, and Tianyi Yuan	Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text	Fine-tuning mBERT on code-switched text with class-weighted loss for multi-label sentiment analysis.	mBERT with class-weighting significantly improves sentiment detection in imbalanced, multilingual data
2	Hemant Palivela Vinay Rishiwal, Meeranarvekar, David Asirvatham, Shashi Bhushan, and Udit Agarwal	Code-Switching ASR for Low-Resource Indic Languages: A Hindi-Marathi Case Study	The study fine-tunes ASR models using deep learning to improve Hindi-Marathi code-switching recognition.	The study improves Hindi-Marathi ASR accuracy, highlights challenges in code-switching, and suggests data augmentation and multilingual pretraining for better performance.
3	Syed Meesam Raza Naqvi, Hassan Aqeel Khan, Ali Raza, Muhammad Ali Tahir, Kamran Javed, and Zubair Saeed	Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services	The methodology uses transfer learning on pre-trained ASR models, fine-tuned with accent embeddings and language tags on a custom code-mixed dataset	The study finds that accent and code-mix adaptation greatly improve speech recognition accuracy.

4	Naorem Karline Singh, Wangkheimayum Madal, Chingakham Neeta Devi, Hoomexsun Pangsatabam, and Yambem Jina Chanu	Leveraging Synthetic Data for Improved Manipuri-English Code-Switched ASR	The study enhances E2E ASR for code-switching using multitask learning with language ID and data augmentation.	Multitask learning with language ID and data augmentation significantly improves code-switching ASR accuracy and language transition handling.
5	Ganji Sreeram and Rohit Sinha	Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements.	The method enhances code-switching ASR using target reduction, context-aware transduction, and a CSI feature.	The approach improves code-switching ASR by reducing TER and WER, enhancing context modelling, and lowering computational cost

Figure 1: WER Distribution Between Native and Embedded Language Segments

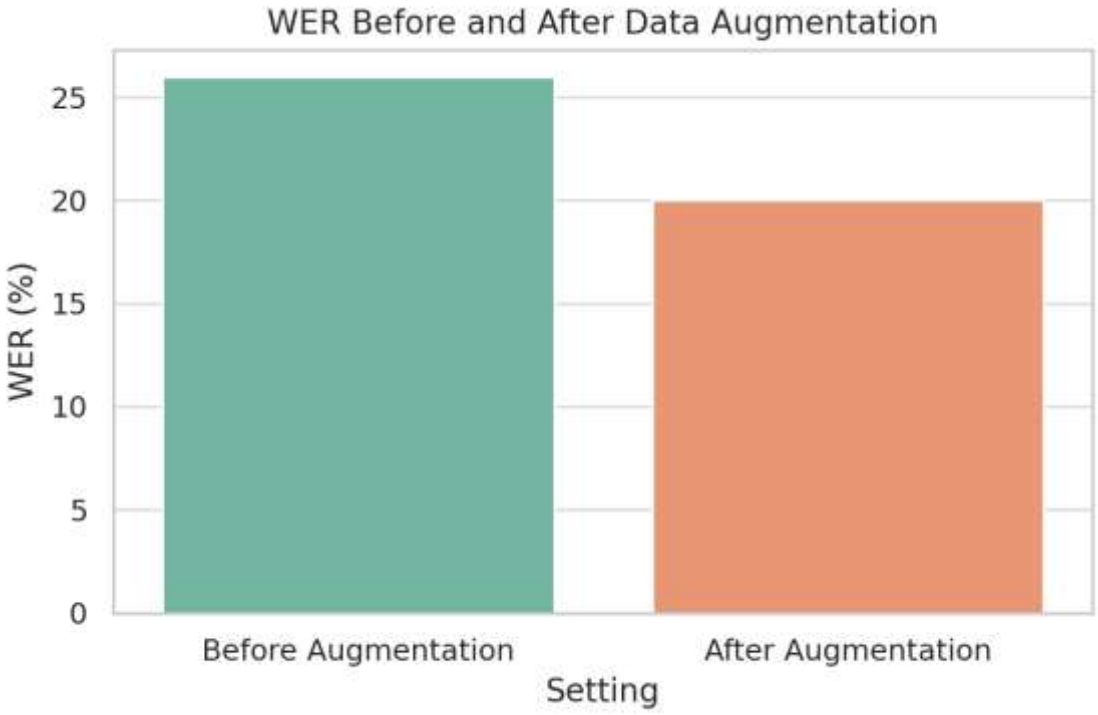

Figure 2: Effectiveness of Synthetic Data Augmentation in Reducing ASR Errors

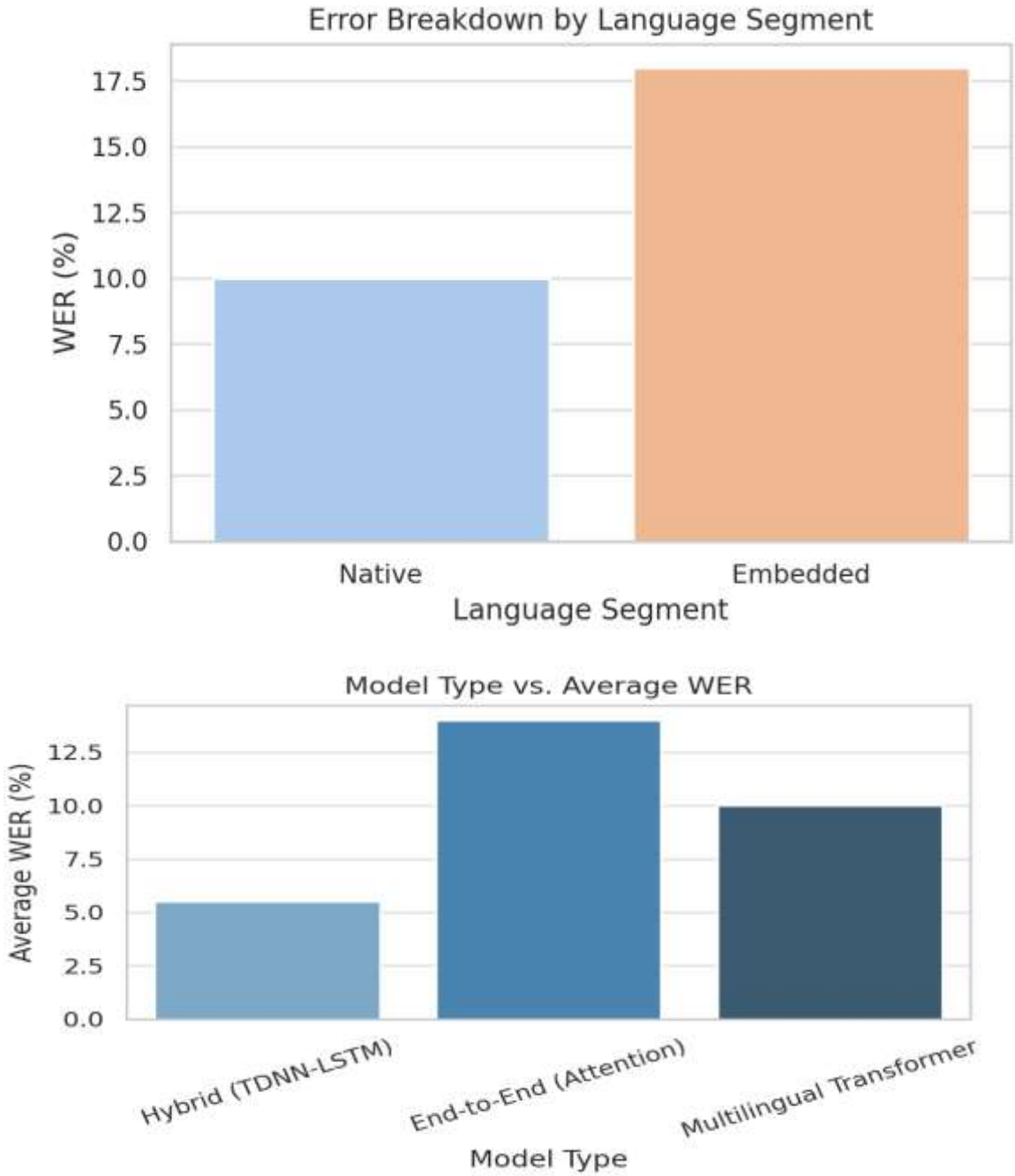

Figure 3: Comparison of ASR Architectures for Code-Switching Speech Recognition

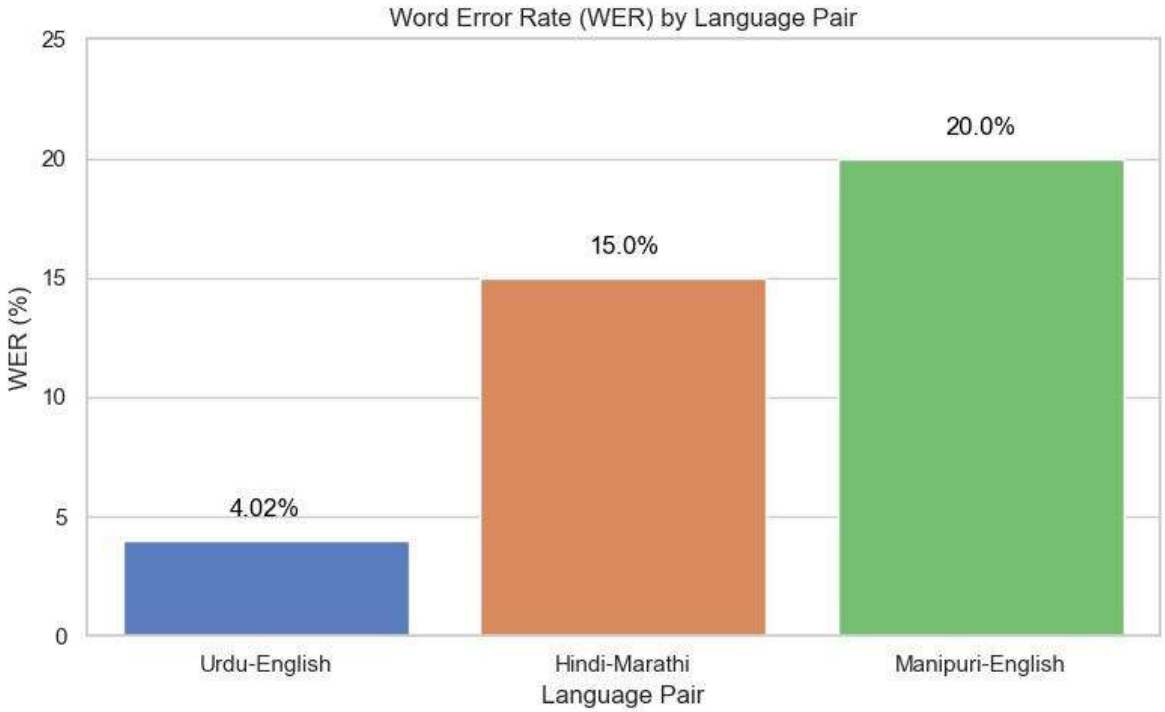

Figure 4: Performance Comparison of Language Pairs Based on Word Error Rate (WER)

IV.RESEARCH GAPS IN EXISTING SYSTEM

Based on the literature review, several research gaps have been identified:

1. **Limited Support for Code-Switched Language Translation**
   - Current models are trained only on single or clean bilingual languages.
   - They can't handle spontaneous code-switching in real conversations.
   - They miss out on meaning, culture, and grammar of mixed languages.
2. **Lack of Parallel Corpora for Code-Switched and Low-Resource Languages**
   - No large datasets for code-switched or low-resource languages.
   - Manual data collection is slow, expensive, and hard to scale.
   - Informal and domain-specific code-switched speech is missing from public data.
3. **Evaluation Benchmarks and Standardization Gaps**
   - Different studies use custom datasets, metrics, and pre-processing steps, making it difficult to compare model performance across research.
   - Current benchmarks rarely consider script variation (e.g., Roman Urdu vs Unicode Urdu), which is common in code-switched content and impacts recognition accuracy.
   - Models are typically evaluated on narrow, single-domain data with limited analysis of how well they perform in new domains or language pairs.

V.PROPOSED SYSTEM

The reviewed papers propose diverse systems to handle code-switching challenges. For ASR, hybrid and end-to-end frameworks are used, such as a Hindi-Marathi system integrating LID and Transformers, and a Hindi-English model with context-dependent target-to-word transduction. A Manipuri-English system leverages synthetic data generation via Transformer + Pointer Generator and trains ASR using Transformer. For sentiment analysis, a BERT-based system applies data augmentation and ensemble learning to address emotion imbalance in Chinese-English text. Additionally, an Urdu-English address recognition system combines Unicode and Roman Urdu datasets for accent-adaptive modelling. These systems emphasize multilingual robustness, low-resource adaptability, and code-switch-aware processing.

VI.CONCLUSION AND FUTURE SCOPE

The reviewed works present promising approaches for enhancing code-switched ASR and NLP tasks in low-resource settings through hybrid models, synthetic data generation, and fine-tuned transformers. While significant improvements in recognition accuracy and sentiment analysis were achieved, challenges remain in data scarcity, accent variability, and real-time multilingual adaptability. Future research should explore multilingual pre-trained models, context-aware augmentation, multimodal inputs (e.g., visual cues), and standardized evaluation benchmarks to broaden generalizability and support scalable, inclusive applications across diverse linguistic landscapes

VII.REFERENCES

[1] H. Palivela, M. Narvekar, D. Asirvatham, S. Bhushan, V. Rishiwal, and U. Agarwal,
"Code-Switching ASR for Low-Resource Indic Languages: A Hindi-Marathi Case Study,"
in IEEE Access, vol. 13, pp. 9171–9186, 2025. DOI: 10.1109/ACCESS.2025.3527745.

[2] S. M. R. Naqvi, M. A. Tahir, K. Javed, H. A. Khan, A. Raza, and Z. Saeed,
"Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services,"
in IEEE Access, vol. 12, pp. 168393–168410, 2024. DOI: 10.1109/ACCESS.2024.3496617.

[3] G. Sreeram and R. Sinha,
"Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements,"
in IEEE Access, vol. 8, pp. 68146–68159, 2020. DOI: 10.1109/ACCESS.2020.2986255.

[4] T. Tang, X. Tang, and T. Yuan,
"Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text,"
in IEEE Access, vol. 8, pp. 193248–193259, 2020. DOI: 10.1109/ACCESS.2020.3030468.

[5] N. K. Singh, W. Madal, C. N. Devi, H. Pangsatabam, and Y. J. Chanu,
"Leveraging Synthetic Data for Improved Manipuri-English Code-Switched ASR,"
in IEEE Access, vol. 13, pp. 25723–25738, 2025. DOI: 10.1109/ACCESS.2025.3538664.