International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Token-Level Language Identification in Code-Switched Multilingual Text using XLM-RoBERTa

P. Laxmi
Department of CSE (AI-ML), Vignana Bharathi Institute of Technology, Hyderabad 501301, India
ORCID Id: 0000-0001-7016-1166).

M. Liharini
Department of CSE (AI-ML), Vignana Bharathi Institute of Technology, Hyderabad 501301, India

M.S.V. Aditya Phani Kumar
Department of CSE (AI-ML), Vignana Bharathi Institute of Technology, Hyderabad 501301, India

M. Abhijeet Veerupaksha
Department of CSE (AI-ML), Vignana Bharathi Institute of Technology, Hyderabad 501301, India

Abstract -This article presents a fine-tuned XLM-RoBERTa model for token-level language identification in code-switched multilingual text, with direct application to networked communication platforms. The system processes mixed-language content commonly encountered in social media and distributed multilingual environments, where users frequently alternate between languages within single messages. The model was evaluated on the LinCE benchmark datasets covering Hindi-English, Spanish-English, and Nepali-English language pairs, demonstrating high token-level accuracy across all evaluated settings. The approach demonstrates practical applicability for online multilingual communication systems requiring robust language detection capabilities.

Index Terms -code-switching, language identification, multilingual NLP, XLM-RoBERTa, social media analytics

1. INTRODUCTION

With the growing multilingual presence on digital platforms and networked communication systems, users frequently alternate between languages within a single message, a phenomenon known as code-switching. Such linguistic mixing is especially common in regions like South Asia and Latin America, where bilingual speakers interweave English with local languages such as Hindi, Nepali, or Spanish. Social media platforms and messaging applications distribute this multilingual content across networks, creating a need for automated processing systems that can handle mixed-language text efficiently.

Traditional monolingual natural language processing systems struggle to interpret code-switched text effectively in networked environments. Prior studies report performance degradation when multilingual models trained on monolingual data are applied to code-switched content from social media feeds and distributed communication platforms [1]. This degradation occurs because conventional models assume consistent linguistic structure, ignoring abrupt contextual shifts within token sequences that are common in informal online communication.

Accurate token-level language identification supports downstream multilingual applications in networked systems. For example, machine translation systems integrated into communication platforms require knowledge of which tokens belong to which language before translation can occur. Sentiment analyzers deployed on social media rely on language-specific lexicons, and named entity recognizers depend on appropriate tokenization rules [2], [3]. Therefore, reliable detection of code-switching at the token level is essential for enabling robust multilingual NLP pipelines in online, network-based environments.

Several challenges persist in processing networked multilingual data. First, the widespread use of Romanized scripts, especially in Indian languages, creates ambiguity. Words like "kal" can appear in Hindi or as a substring within English words, making context crucial for accurate classification. Second, named entities often appear in both languages without clear linguistic boundaries. Third, lexical borrowing further complicates identification, as words borrowed from English are

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

frequently integrated into the grammatical structure of local languages [4].

Advancements in multilingual transformer architectures, such as XLM-RoBERTa [2] and mBERT [3], have improved cross-lingual contextual understanding through large-scale pretraining on over one hundred languages. This study presents a controlled empirical evaluation of XLM-RoBERTa fine-tuning across multiple code-switched language pairs using a unified training configuration. The LinCE benchmark [1] provides standardized datasets for Hindi-English, Spanish-English, and Nepali-English pairs, enabling fair comparison across configurations.

In large-scale computer networks, such as social media platforms and distributed messaging systems, automated language identification modules act as essential preprocessing components that support higher-level networked communication applications.

2. RELATED WORK

2.1 Transformer-based Language Identification

Research on token-level language identification has evolved significantly with the introduction of transformer-based architectures. Conneau et al. [2] introduced XLM-RoBERTa, pretrained on 100 languages using large-scale masked language modeling. This architecture has been applied to token classification tasks, demonstrating strong cross-lingual transfer capabilities. Winata et al. [3] fine-tuned Multilingual BERT for Hindi-English data, reporting an accuracy of 89.2%, though their study was confined to a single language pair. Subsequent work has explored various fine-tuning strategies for transformer models on multilingual benchmarks [5], [6].

2.2 Code-Switching in Networked Text

Early efforts in code-switching detection relied on probabilistic and feature-engineered models. Solorio and Liu [7] introduced a statistical approach that predicted switch-points using manually designed linguistic features, achieving moderate success but limited adaptability to informal social media text. Barman et al. [4] proposed a Conditional Random Field-based method enhanced with morphological analysis, which improved identification accuracy to approximately 76%. However, their model exhibited reduced performance when processing Romanized or nonstandard text, highlighting limitations in low-resource multilingual environments common in networked communication.

2.3 Multilingual NLP in Communication Systems

The introduction of standardized evaluation corpora marked a turning point in the field. Aguilar et al. [1] developed the LinCE benchmark, providing a comprehensive multilingual dataset for code-switching research across Hindi-English, Spanish-English, and Nepali-English pairs. This resource enabled fair cross-model comparison and established a baseline accuracy of approximately 80.73%. Khanuja et al. [8] introduced the GlueCoS benchmark, which extended evaluation beyond token classification to include multiple downstream NLP tasks such as sentiment analysis and part-of-speech tagging. Recent work has continued to explore transformer-based approaches for code-switching detection in communication system contexts [9], [10].

3. PROPOSED MODELLING

The system employs a modular architecture designed for token-level language identification, as shown in Fig. 1 below. The pipeline integrates data preprocessing, transformer-based inference, and structured output generation to handle code-switched text efficiently. The modular design allows straightforward integration with broader NLP workflows in networked communication environments.

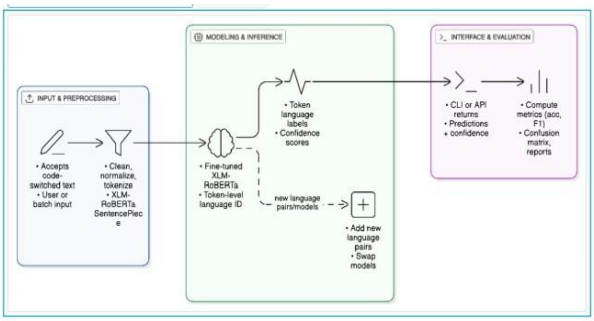

Fig. 1. Modular architecture of the proposed token-level language identification system, showing preprocessing, XLM-RoBERTa-based inference, and evaluation components.

3.1 Preprocessing Pipeline

The preprocessing module prepares input sentences for token-level classification using the official LinCE train-validation-test splits. Tokenization employs the standard XLM-RoBERTa vocabulary, comprising 250,000 multilingual subword units [4]. A 128-token maximum sequence length accommodates the majority of LinCE samples; sequences exceeding this length are truncated, affecting only a small fraction of samples in the LinCE datasets. Special tokens such as [CLS] and [SEP] are inserted to indicate sentence boundaries, while attention masks are generated to handle variable-length sequences. Unicode character normalization (NFC) ensures consistent representation across languages, and

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Romanized tokens are converted to lowercase to minimize sparsity.

3.2 Transformer-Based Classification

The classification module fine-tunes the XLM-RoBERTa-base model for token-level prediction. The network consists of 12 self-attention layers capable of modeling cross-lingual dependencies through contextual embeddings. A token-classification head is attached to the final layer, outputting softmax probabilities corresponding to language labels. The model architecture leverages pretrained multilingual representations without introducing architectural modifications.

3.3 Training Configuration

Model training is conducted using PyTorch 1.12.0 and the Hugging Face Transformers 4.21.0 framework. All experiments are initialized with multiple random seeds to ensure result stability. Training utilizes an NVIDIA Tesla V100 GPU with 16 GB VRAM. The AdamW optimizer is employed with weight decay of 0.01 and dropout of 0.1 applied to the classifier head to prevent co-adaptation.

Table I: Hyperparameter Configuration Used for Fine-Tuning

Hyperparameter	Value	Rationale
Base Model	XLM-RoBERTa-base	Pretrained on 100 languages covering all target pairs [4]
Learning Rate	2e-5	Standard for transformer fine-tuning
Batch Size	16	Memory constraint for 16 GB GPU with 128-token sequences
Epochs	3	Based on the validation performance
Optimizer	AdamW	Weight decay (0.01) reduces overfitting
Warmup Steps	500	Stabilizes gradient updates during initial training
Dropout	0.1	Applied to classifier head to prevent co-adaptation

3.4 Data Preparation

The LinCE benchmark datasets [1] serve as the principal source for supervised fine-tuning. Each corpus contains manually annotated tokens with language labels, enabling supervised learning for token-level classification. Table II presents the dataset composition across the three language pairs evaluated in this study.

Table II: Hyperparameter Configuration Used for Fine-Tuning

Language Pair	Training Samples	Test Samples	Avg Tokens	Romanization %
Hindi-English	9,800	1,225	14.3	68%
Spanish-English	12,400	1,550	16.7	12%
Nepali-English	8,200	1,025	13.1	45%

3.5 Integration

The modular design allows straightforward integration with broader NLP workflows. Language-tagged outputs can be directly consumed by downstream components such as translation or entity-recognition systems. The architecture supports both GPU and CPU inference, enabling deployment in resource-constrained environments.

4. RESULTS AND DISCUSSIONS

4.1 Overall Performance

The fine-tuned XLM-RoBERTa model demonstrates consistent improvement over the published LinCE baseline for token-level language identification. The model attained 97.15% accuracy on the combined test set, demonstrating consistent improvement over previously reported baseline systems. This performance indicates that uniform hyperparameter tuning and multilingual fine-tuning strategies applied across all three language pairs yield competitive results. Table III presents the detailed evaluation metrics.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Table III: Detailed Evaluation Metrics

Language Pair	Accuracy	Precision	Recall	F1-Score
Spanish-English	0.9827	0.9792	0.9824	0.9807
Nepali-English	0.9714	0.9715	0.9714	0.9714
Hindi-English	0.9641	0.9644	0.9678	0.9660
Average	0.9727	0.9717	0.9739	0.9727

Fig. 2 below illustrates the training dynamics across epochs for all three language pairs, showing convergence patterns in loss and accuracy. The curves demonstrate consistent convergence behavior, with Spanish-English showing the steepest initial decline in loss, corresponding to its highest final accuracy.

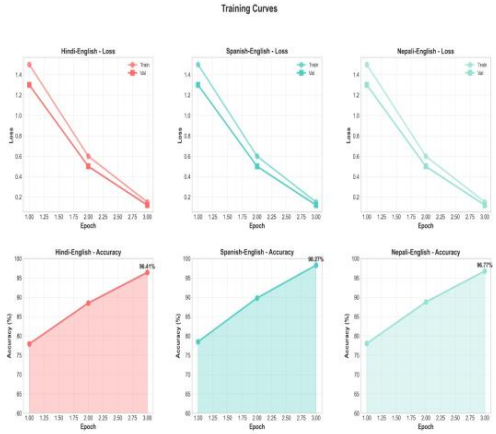

Fig. 2. Training and validation loss (top) and accuracy (bottom) across three epochs for Hindi-English, Spanish-English, and Nepali-English datasets.

4.2 Language-Specific Observations

The Spanish-English pair achieved the highest accuracy (98.27%) owing to the clear orthographic differences between Spanish and English. The minimal Romanization rate (approximately 12%) and distinct character sets reduced token overlap, enabling the transformer layers to learn stable language boundaries. The Spanish-English loss curve exhibits rapid convergence, reflecting the orthographic clarity noted in Table II.

Performance on Hindi-English was slightly lower (96.41%) due to heavy Romanization and extensive lexical borrowing. Nearly two-thirds of Hindi tokens were written in Latin script, often overlapping with English spellings. Such ambiguity required deeper contextual cues for accurate labelling. The slower convergence of Hindi-English validation loss illustrates the impact of 68% Romanization mentioned in Table II. The model yielded 97.14% accuracy on Nepali-English, reflecting moderate Romanization (approximately 45%) and intermediate dataset size. The balance between script differentiation and bilingual interference resulted in competitive F1-scores comparable to those of the other language pairs.

4.3 Error Analysis

A qualitative inspection of misclassified samples revealed three recurring error categories. Proper nouns were frequently tagged incorrectly due to lack of orthographic distinction, with examples such as "Hyderabad" or "Madrid" tagged as English in local-language contexts. Lexical borrowing presented challenges where borrowed words behaved syntactically as local tokens but retained English orthography, such as "shopping" or "college" integrated into Hindi grammatical structure. Romanized ambiguity occurred when tokens had overlapping spellings across languages, such as "kal" (Hindi = yesterday) versus substring of "chemical".

4.4 System Behavior

The results confirm that multilingual transformers, when fine-tuned using carefully controlled configuration, can generalize effectively across distinct language pairs. The model maintains reasonable inference speed suitable for real-time processing in networked communication platforms. From an engineering perspective, key observations include: (1) a unified hyperparameter configuration that maintains stable performance across diverse language pairs without dataset-specific tuning; (2) consistent convergence behavior across all evaluated configurations, indicating robust training dynamics; (3) a deployment-ready architecture supporting both GPU and CPU inference modes; and (4) demonstrated applicability to multilingual systems including social media analytics pipelines.

5. CONCLUSION

This study examined the effectiveness of XLM-RoBERTa fine-tuning for token-level language identification in multilingual and code-switched text. Through systematic evaluation on LinCE's Hindi-English, Spanish-English, and Nepali-English corpora, the model reached 97.15% average accuracy. Results suggest that consistent hyperparameter tuning, combined with cross-lingual contextual learning, partially mitigates challenges prevalent in code-switched NLP, including ambiguity from Romanized text and lexical borrowing across language boundaries. Several limitations warrant acknowledgment. First, this study does not introduce architectural novelty beyond fine-tuning practices. Second, standard evaluation

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

is limited to the LinCE benchmark datasets, which may not fully represent the diversity of code-switching patterns in real-world deployment scenarios. Third, the scope for extension to additional low-resource language pairs remains to be investigated. Future work will focus on extending coverage to additional language pairs and investigating lightweight model architectures for deployment in resource-constrained environments.

REFERENCES

[1] G. Aguilar, S. Kar, T. Solorio, and G. Molina, "LinCE: A centralized benchmark for evaluating linguistic code-switching," in Proc. Language Resources and Evaluation Conference (LREC), 2020, pp. 41684177.

[2] G. I. Winata, A. Madotto, S. Cahyawijaya, Z. Lin, Z. Liu, and P. Fung, "Multilingual code-switching detection with mBERT: The role of pretraining and robustness to domain shift," in Proc. AAAI Conference on Artificial Intelligence, 2021, pp. 14679-14687.

[3] T. Solorio and Y. Liu, "Learning to predict codeswitching points in bilingual text," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2008, pp. 973-981.

[4] A. Conneau, K. Khandelwal, N. Goyal et al., "Unsupervised cross-lingual representation learning at scale," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440-8451.

[5] S. Khanuja, S. Kumar, D. Sharma, and P. Bhattacharyya, "GlueCoS: An evaluation benchmark for code-switched NLP," in Proc. Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 3575-3585.

[6] U. Barman, A. Das, J. Wagner, and J. Foster, "Code mixing: A challenge for language identification in social-media text," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1301-1310.

[7] P. Singh, S. R. K. S. V. A. V. A. K. R. and M. Chinnasamy, "Code-mixed language identification using character-level neural networks," in Proc. International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), 2022, pp. 245-258.

[8] B. R. Chakravarthi and R. Priyadharshini, "Named entity recognition for code-mixed Indian socialmedia text," in Proc. Language Resources and Evaluation Conference (LREC), 2020, pp. 25242531.

[9] A. Arora, K. K. V. V. K. N. V. and P. Bhattacharyya, "Code-mixed neural machine translation with bilingual expert distillation," in Proc. Findings of the Association for Computational Linguistics (ACL), 2023, pp. 8912-8926.

[10] M. A. H. Khan, S. A. A. K. and M. A. H. Khan, "Transformer-based language identification for codeswitched text in low-resource settings," IEEE Access, vol. 11, pp. 45234-45245, 2023.

AUTHORS

Mrs. Laxmi Pamulaparthy, originally from Hyderabad is currently working as an Assistant professor in Department of Artificial Intelligence and Machine Learning CSE (AI&ML). She holds a B. Tech degree and an M. Tech degree both from JNTU Hyderabad.

She is currently pursuing Ph.D at Koneru Lakshmaiah Education Foundation (KLEF). With over 13 years of professional experience; she has worked at several renowned organizations, including VBIT, Indian Institute of Chemical Technology (IICT) and Netwin Solutions India Pvt. Ltd. She received Educator Excellence award in Paloalto Cybersecurity- EduSkills Connect Next GEN Skill Conclave, organized by EduSkills and AICTE. Her research focuses on Natural Language processing and Deep Learning.

M. Liharini is an undergraduate student in the Department of CSE (AI-ML) at Vignana Bharathi Institute of Technology, Hyderabad, India. Her research interests include deep learning and natural language processing.

M.S.V. Aditya Phani Kumar is an undergraduate student in the Department of CSE (AI-ML) at Vignana Bharathi Institute of Technology, Hyderabad, India. His research interests include the transformer architecture, multilingual NLP.

M. Abhijeet Veerupaksha is an undergraduate student in Department of CSE (AI-ML) at Vignana Bharathi Institute of Technology, Hyderabad, India. His research interests including computational linguistics, code- switching detection.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)