# Advancing Language Identification in Code-Mixed Tulu Texts: Harnessing Deep Learning Techniques

Supriya Chanda[1], Anshika Mishra[2] and Sukomal Pal[1]

[1]Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, INDIA, 221005
[2]Department of Computer Science and Engineering, Vellore Institute of Technology Bhopal, Madhya Pradesh, INDIA

Abstract

This study focuses on the task of word-level language identification in code-mixed Tulu-English texts, which is crucial for addressing the linguistic diversity observed on social media platforms. The CoLI-Tunglish shared task served as a platform for multiple teams to tackle this challenge, aiming to enhance our understanding of and capabilities in handling code-mixed language data. To tackle this task, we employed a methodology that leveraged Multilingual BERT (mBERT) for word embedding and a Bi-LSTM model for sequence representation. Our system achieved a Precision score of 0.74, indicating accurate language label predictions. However, our Recall score of 0.571 suggests the need for improvement, particularly in capturing all language labels, especially in multilingual contexts. The resulting F1 score, a balanced measure of our system's performance, stood at 0.602, indicating a reasonable overall performance. Ultimately, our work contributes to advancing language understanding in multilingual digital communication.

**Keywords**

Social Media, Code-Mixed, Multilingual BERT, Language Identification, Tulu,

## 1. Introduction

In the era of widespread social media usage, Natural Language Processing (NLP) techniques have become indispensable tools, facilitating global communication and information sharing. However, this digital age of user-generated content (UGC) has introduced a unique challenge to NLP systems - code-mixing. Code-mixing, the concurrent use of multiple languages within a single text, often arises when users are not proficient with their native language keyboards, leading to the transliteration of text or the blending of languages.

Code-mixing encompasses two related phenomena: code-switching and code-mixing. Code-switching involves the deliberate alternation between languages, typically at sentence boundaries, to aid comprehension. On the other hand, code-mixing is the unconscious and frequent use of multiple languages within a single phrase or sentence, involving various linguistic components like phonology, morphology, grammar, and lexicon. While these terms are often used interchangeably, we refer to both as code-mixing for simplicity in this context.

*FIRE'23: Forum for Information Retrieval Evaluation, December 15-18, 2023, India*
supriyachanda.rs.cse18@itbhu.ac.in (S. Chanda); anshika.mishra2019@vitbhopal.ac.in (A. Mishra); spal.cse@itbhu.ac.in (S. Pal)
0000-0002-6344-8772 (S. Chanda); 0000-0002-9421-8566 (S. Pal)

While code-mixing in spoken language has been extensively studied for decades, the analysis of code-mixed text, particularly in social media, is a relatively new frontier. Modern NLP models excel in tasks such as sentiment analysis [1, 2, 3], language identification, hate speech identification [4, 5, 6], information retrieval [7] and named-entity recognition for monolingual text but struggle when confronted with code-mixed content.

This shared task, CoLI-Tunglish (Code-mixed Tulu-English Language Identification), addresses the intricate problem of word-level language identification in code-mixed Tulu-English texts. Tulu, a regional language, coexists with Kannada and English, especially in social media discourse among Tulu-speaking individuals. The fusion of these languages in roman script has generated a unique and largely unexplored dataset.

Participants in this task are challenged to develop methods for the precise identification and categorization of words within code-mixed Tulu-English sentences. This task is a vital step in advancing NLP capabilities to handle the intricacies of code-mixed text, bridging the gap between the linguistic diversity of social media and automated language processing.

In this paper, we provide an overview of the CoLI-Tunglish shared task, emphasizing the importance of accurate word-level language identification in code-mixed Tulu-English texts and highlighting the unique linguistic characteristics of this dataset. We encourage researchers to delve into this challenging domain, advancing the capabilities of NLP systems in understanding the nuances of code-mixing within digital communication platforms.

The remainder of the paper is structured as follows. Section 2 provides a synopsis of some earlier work. The datasets that we have used are discussed in Section 3. Section 4 presents our computational methodologies, model descriptions, and evaluation methodology, followed by Section 5 results and analysis. We conclude in Section 6.

## 2. Related Work

In recent years, the field of computational linguistics has witnessed a surge of interest in addressing code-mixing challenges. This growing focus on code-switching and code-mixing tasks has led to various initiatives and investigations.

The first and second workshops on a computational approach to code-switching, held in conjunction with Empirical Methods in Natural Language Processing (EMNLP) in 2014 [1] and 2016, featured shared tasks on language identification across diverse language pairs. Additionally, the Forum for Information Retrieval Evaluation (FIRE) organized multiple shared tasks centered on language identification for Indian language pairs during the years 2014, 2015, and 2016.

Early approaches to language identification included Cavnar and Trenkle's [8] character n-gram method, which achieved remarkable accuracy of 99.8% on newsgroup article data but proved inadequate when applied to code-mixed social media content. This discrepancy can be attributed to the formal and standardized nature of newsgroup articles, in contrast to the informal and dynamic nature of social media text.

Nguyen et al. [9] explored various techniques, including dictionary lookup, language models, logistic regression classifiers, and conditional random fields (CRF) classifiers, for Turkish-Dutch code-mixed data. For Indian language pairs, Barman et al. [10] tackled word-level

[1] https://emnlp2014.org/workshops/CodeSwitch/call.html

language identification using methods such as supervised classification (Support Vector Machine (SVM)), sequence classification (CRF), and dictionary lookup on Bengali-Hindi-English Facebook comments. Similarly, Das and Gamback [11] applied these methods with diverse features to code-mixed chat message corpora (English-Bengali and English-Hindi) and introduced the Code Mixed Index (CMI) to evaluate code-mixing levels in a corpus. Vyas et al. [12] created a multi-level annotated corpus of Hindi-English code-mixed text from Facebook forums, exploring language identification, back-transliteration, normalization, and part-of-speech tagging.

Recent research endeavors have explored the utility of non-textual features, like neighborhood-based features, for multi-language Language Identification tasks. Two notable benchmarks, Linguistics Code-switching Evaluation (LinCE) [13] and GLUECoS [14], have emerged for specific language pairs and tasks. LinCE covers four language pairs: Spanish-English [15], Nepali-English [16], Hindi-English, and Modern Standard Arabic-Egyptian Arabic. GLUECoS provides a framework for evaluating language understanding in Code-Switched NLP, focusing on language pairings such as English-Hindi and English-Spanish. CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts was organised at ICON 2022 [17].

In the realm of deep learning, Joshi et al. [18] evaluated the effectiveness of character, sub-word, and word-based representations for language identification in Hindi-English code-mixed data. They formulated this task as a token classification problem, employing convolutional neural networks (CNN) and LSTM networks on top of word representations. Jamatia et al. [19] leveraged pre-trained word embeddings (GloVe) along with LSTM layers and Character-level Recurrent Neural Networks (RNNs) with CRF classifiers. Their research demonstrated that deep learning models achieved competitive accuracy compared to supervised approaches like CRF. Recent advancements, such as BERT models, have further elevated language understanding by constructing contextual word representations based on surrounding words, enhancing the field of dynamic language representation.

## 3. Dataset

The CoLI-Tunglish dataset comprises words from Tulu, Kannada, and English, all transcribed in Roman script. These words are categorized into seven main groups: 'Tulu,' 'Kannada,' 'English,' 'Mixed-language,' 'Name,' 'Location,' and 'Other.' Table 1 provides detailed descriptions of these labels within the CoLI-Tunglish dataset [20]. Additionally, Table 2 illustrates the distribution of these labels across the training set, and development set.

## 4. Methodology

In our experimental framework, we have adopted a layered architecture comprising three fundamental components. The initial layer involves word embedding, which represent individual words within a text, considering the contextual information provided by neighboring words. This embedding layer plays a crucial role as it transforms each word in the input sequence into a vector format, which is essential for subsequent processing. To enhance the contextual language representation, we have harnessed the power of BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art model. Specifically, we utilized the multilingual BERT

**Table 1**
Description of labels in CoLI-Tunglish dataset

| Category | Descriptions | Samples |
|---|---|---|
| Name | Words that indicate name of a person (including Indian names) | Koragajja, daiva, thaniye |
| Location | Words that indicate the location | Padil, Kudla, Kapikad |
| English | Pure English words | super, style, comedy |
| Tulu | Tulu words written in Roman script | maste (very), tikkund (if we get), moke (Love) |
| Kannada | Kannada words written in Roman script | ashirvada (Blessings), namma (our), Kushi (happy) |
| Mixed | Combination of Kannada, Tulu and/or English words in Roman script | teamda (team + da, in a team), Lovetha (Love + tha, for love), Actorsnakl (actors + nakl, all actors) |
| Other | Words not belonging to any of the above categories and words of other languages | Znjdjfjbj - not a word, Kannada words in Kannada script, Hindi words in Devanagari script, Hindi words in Roman script, Malayalam words in Roman script |

**Table 2**
Class wise distribution of Train and Development data set

| Category | Count in training data | Count in development data |
|---|---|---|
| Tulu | 8647 | 1461 |
| English | 5499 | 889 |
| Kannada | 2068 | 344 |
| Name | 1104 | 162 |
| Other | 506 | 102 |
| Mixed | 403 | 69 |
| Location | 369 | 54 |

(mBERT) pre-trained model to generate word embedding tailored for the task of code-mixed language identification.

Following the word embedding layer, we have employed a Sequence Layer responsible for generating a comprehensive word sequence representation. This layer takes as input the sequence of embedding vectors from the text sequence and leverages bidirectional Long Short-Term Memory (Bi-LSTM) networks. In the context of Bi-LSTM, the term "bidirectional" indicates that information flows both forward and backward in time. Consequently, the Bi-LSTM processes the input sequence in two directions, producing a hidden forward sequence and a

hidden backward sequence. The final encoded vector results from the concatenation of these forward and backward hidden unit outputs. It's noteworthy that, in our study, we opted for a simplified LSTM model, where a single Bi-LSTM with 256 hidden units was employed.

The concluding layer in our architecture is a softmax feedforward network. This network generates a probability distribution for each word in the sequence across a predefined list of tags. During the prediction phase, the tag with the highest associated probability is selected as the predicted tag for each word. In the training phase, we set specific hyperparameters, including a learning rate of 0.01, a batch size of 16, and a maximum of 10 training epochs. These parameters are crucial in guiding the learning process and fine-tuning the model for optimal performance.

## 5. Results

The organizers employed a comprehensive set of performance metrics, including accuracy, precision, recall, and F1 scores, to meticulously assess the submissions from various participating teams. Subsequently, they made the top scores achieved by each team publicly available. Table 3 presents a detailed overview of the performance metrics for all participating teams, along with their respective rankings [21].

**Table 3**
Evaluation results on test data and rank list

| Team Name | Precision | Recall | $F_1$ score | Rank |
|---|---|---|---|---|
| SATLAB | 0.851 | 0.783 | 0.813 | 1 |
| BFCAI | 0.859 | 0.777 | 0.812 | 2 |
| Poorvi | 0.821 | 0.781 | 0.799 | 3 |
| MUCS | 0.807 | 0.743 | 0.770 | 4 |
| IRLab@IITBHU | 0.740 | 0.571 | 0.602 | 5 |

Our team achieved a commendable Precision score of 0.74, indicating the accuracy of our language label predictions within the code-mixed text. However, our system obtained a Recall score of 0.571, suggesting potential improvements in identifying all language labels, particularly in multilingual contexts.

The F1 score, which balances Precision and Recall, reached 0.602, reflecting reasonable overall performance. Nevertheless, there is room for enhancement to achieve a higher F1 score.

In the team rankings, we secured the 5th position with a single submission. While our system exhibited accuracy, other teams outperformed us in both Precision and Recall, impacting our F1 score and final ranking

## 6. Conclusion

In conclusion, the CoLI-Tunglish shared task addressed the intricate challenge of word-level language identification in code-mixed Tulu-English texts. This task highlighted the growing

importance of understanding and processing code-mixed language data, which is prevalent on social media platforms. The participating teams showcased diverse approaches, with varying degrees of success, as reflected in metrics such as Precision, Recall, and F1 scores. While significant strides have been made in addressing code-mixing challenges, there remains ample room for improvement, particularly in enhancing Recall and achieving a more balanced F1 score.

The shared task provided valuable insights into the state of the art in code-mixed language identification and encouraged further research in this evolving field. It underlines the need for advanced NLP techniques to bridge the gap between linguistic diversity in digital communication and automated language processing. Future endeavors in this domain will likely yield more robust solutions for handling code-mixed text, enabling more accurate language understanding and information retrieval in multilingual contexts.

## Acknowledgements

We are thankful to the organizers for providing the opportunity to work on this interesting and important task.

## References

[1] S. Chanda, S. Pal, Irlab@ iitbhu@ dravidian-codemix-fire2020: Sentiment analysis for dravidian languages in code-mixed text., in: FIRE (Working Notes), 2020, pp. 535–540.

[2] S. Chanda, R. Singh, S. Pal, Is meta embedding better than pre-trained word embedding to perform sentiment analysis for dravidian languages in code-mixed text?, Working Notes of FIRE (2021).

[3] S. Chanda, A. Mishra, S. Pal, Sentiment analysis and homophobia detection of code-mixed dravidian languages leveraging pre-trained model and word-level language tag, in: Working Notes of FIRE 2022-Forum for Information Retrieval Evaluation (Hybrid). CEUR, 2022.

[4] A. Saroj, S. Chanda, S. Pal, Irlab@ iitv at semeval-2020 task 12: multilingual offensive language identification in social media using svm, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 2012–2016.

[5] S. Chanda, S. Ujjwal, S. Das, S. Pal, Fine-tuning pre-trained transformer based model for hate speech and offensive content identification in english, indo-aryan and code-mixed (english-hindi) languages, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.

[6] S. Chanda, S. Sheth, S. Pal, Coarse and fine-grained conversational hate speech and offensive content identification in code-mixed languages using fine-tuned multilingual embedding, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE). CEUR-WS. org, 2022.

[7] S. Chanda, S. Pal, The effect of stopword removal on information retrieval for code-mixed data obtained via social media, SN Computer Science 4 (2023) 494.

[8] W. B. Cavnar, J. M. Trenkle, N-gram-based text categorization, in: In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pp. 161–175.

[9] D. Nguyen, A. S. Doğruöz, Word level language identification in online multilingual communication, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Seattle, Washington, USA, 2013, pp. 857–862. URL: https://aclanthology.org/D13-1084.

[10] U. Barman, A. Das, J. Wagner, J. Foster, Code mixing: A challenge for language identification in the language of social media, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 13–23. URL: https://aclanthology.org/W14-3902. doi:10.3115/v1/W14-3902.

[11] A. Das, B. Gambäck, Identifying languages at the word level in code-mixed Indian social media text, in: Proceedings of the 11th International Conference on Natural Language Processing, NLP Association of India, Goa, India, 2014, pp. 378–387. URL: https://aclanthology.org/W14-5152.

[12] Y. Vyas, S. Gella, J. Sharma, K. Bali, M. Choudhury, POS tagging of English-Hindi code-mixed social media content, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 974–979. URL: https://aclanthology.org/D14-1105. doi:10.3115/v1/D14-1105.

[13] G. Aguilar, S. Kar, T. Solorio, LinCE: A centralized benchmark for linguistic code-switching evaluation, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 1803–1813. URL: https://aclanthology.org/2020.lrec-1.223.

[14] S. Khanuja, S. Dandapat, A. Srinivasan, S. Sitaram, M. Choudhury, GLUECoS: An evaluation benchmark for code-switched NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 3575–3585. URL: https://aclanthology.org/2020.acl-main.329. doi:10.18653/v1/2020.acl-main.329.

[15] G. Molina, F. AlGhamdi, M. Ghoneim, A. Hawwari, N. Rey-Villamizar, M. Diab, T. Solorio, Overview for the second shared task on language identification in code-switched data, in: Proceedings of the Second Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Austin, Texas, 2016, pp. 40–49. URL: https://aclanthology.org/W16-5805. doi:10.18653/v1/W16-5805.

[16] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Ghoneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang, P. Fung, Overview for the first shared task on language identification in code-switched data, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 62–72. URL: https://aclanthology.org/W14-3907. doi:10.3115/v1/W14-3907.

[17] F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, H. Shashirekha, G. Sidorov, A. Gelbukh, Overview of coli-kanglish: Word level language identification in code-mixed kannada-english texts at icon 2022, in: Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in

Code-mixed Kannada-English Texts, 2022, pp. 38–45.
[18] R. Joshi, R. Joshi, Evaluating input representation for language identification in hindi-english code mixed text, in: ICDSMLA 2020, Springer, 2022, pp. 795–802.
[19] A. Jamatia, A. Das, B. Gambäck, Deep learning-based language identification in english-hindi-bengali code-mixed social media corpora, Journal of Intelligent Systems 28 (2019) 399–408. URL: https://doi.org/10.1515/jisys-2017-0440. doi:doi:10.1515/jisys-2017-0440.
[20] A. Hegde, M. D. Anusha, S. Coelho, H. L. Shashirekha, B. R. Chakravarthi, Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text, in: Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, 2022, pp. 33–40.
[21] A. Hagde, F. Balouchzahi, S. Coelho, S. Hosahalli Lakshmaiah, H. A Nayel, S. Butt, Overview of coli-tunglish: Word-level language identification in code-mixed tulu texts at fire 2023, in: Forum for Information Retrieval Evaluation FIRE - 2023, 2023.