Overview of CoLI-Dravidian 2025: Word-level Code-Mixed Language Identification in Dravidian Languages

Asha Hegde[1], Fazlourrahman Balouchzahi[2], Sabur Butt[3], Sharal Coelho[1], Sudha V[1], Shashirekha Hosahalli Lakshmaiah[1] and Ameeta Agrawal[4]

[1]*Department of Computer Science, Mangalore University, India,*

[2]*Universidad Nacional Autónoma de México (UNAM), Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas (IIMAS), Mexico,*

[3]*IFE, Tecnologico de Monterrey, Mexico,*

[4]*Department of Computer Science, Portland State University, USA*

Abstract

Language Identification (LI) has traditionally been performed at the document or sentence level and specifically for high-resource languages. The rise of social media communication in multilingual regions such as India has seen users generate code-mixed texts, which typically combine English with local languages. Here, word-level LI is of primary importance: a system assigns a language label to each word in the text. It provides a fine granularity essential for capturing frequent language switches in informal, transliterated content. Indeed, it becomes an important step in most downstream NLP tasks such as machine translation, sentiment analysis, and conversation systems. This becomes all the more important because many languages of the Dravidian family are spoken by millions but are characteristically low-resourced. The CoLI-Dravidian shared task attempts to fill this gap by taking up word-level LI in Roman-script code-mixed Dravidian datasets. A total of eight teams participated in this shared task, and the top models achieved macro F1 scores 0.743, 0.921, 0.827, 0.952, and 0.823 for Tamil, Kannada, Malayalam, Telugu, and Tulu, respectively, indicating both the complexity and the progress in this domain.

Keywords

Word-level Language Identification, Code-mixed, Dravidian Languages, Data Collection

1. Introduction

The Dravidian languages are a family of approximately 80 languages spoken by more than 220 million people in South Asia, with a rich and ancient history. As per a recent study, the Dravidian language family, consisting of major languages such as Tamil, Telugu, Kannada, and Malayalam, is approximately 4,500 years old [1]. Usually, speakers of these regional languages are comfortable using more than one language, including English, for daily communication. This multilingual environment causes users to frequently switch between languages and scripts, especially on informal platforms like social media. Hence, code-mixing has emerged as one of the widespread and natural linguistic phenomena in online communication [2]. The code-mixing can occur at several levels-paragraph, sentence, word, or even subword-depending on the speaker's fluency and communicative intent [3]. This dynamic multilingual behavior introduced new opportunities and challenges for the research of computational linguistics.

One of the major tasks in multilingual computational linguistics is the identification of the language for every word in a code-mixed sentence. It forms the base for more accurate NLP tools that enable applications such as machine translation, sentiment analysis, and social media analytics [4]. Since Dravidian languages are highly morphologically rich, the challenge becomes even more complex. These languages have extensive inflection, agglutination, and derivational morphology, which contributes to their large vocabularies and numerous word variations. Morphological complexity introduces ambiguity, complicates tokenization, and makes feature extraction problematic for NLP models. Moreover, most code-mixed data lacks a standard for spellings, transliterations, and grammar. Hence, it calls for

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

hegdekasha@gmail.com (A. Hegde); fbalouc@iimas.unam.mx (F. Balouchzahi); saburb@tec.mx (S. Butt); sharalmucs@gmail.com (S. Coelho); srisudhavenkatesh22@gmail.com (S. V); hlsrekha@mangaloreuniversity.ac.in (S. H. Lakshmaiah); ameeta@pdx.edu (A. Agrawal)

specialized approaches toward the development of robust NLP systems which can take code-mixing and heavy morphological structures into consideration effectively.

We have thus organized a shared task[1] titled "CoLI-Dravidian: Word-level Code-Mixed Language Identification in Dravidian Languages" to address the challenges associated with word-level LI within the Dravidian languages. This is part of FIRE 2025: Code-mixed datasets in five languages, namely Kannada, Tamil, Malayalam, Telugu and Tulu, were provided to develop advanced models for accurate LI systems in morphologically rich languages. The shared task consists of two major phases: a training and validation phase, and a testing phase. In the first phase, the participants will be provided with labeled training and validation datasets in all five languages to develop and tune their systems. In the testing phase, the unlabeled test sets will be released and participants will submit their predictions through the Codabench platform. for automatic evaluation. A maximum of five submissions per language were allowed per team, with only the best performing submission taken into account for the final ranking. Of 35 registered teams, a total of 10 teams have submitted valid predictions for the final evaluation and 8 of them provided detailed working notes documenting system descriptions.

2. Related Works

In recent years, there has been a growing interest among researchers in the field of code-mixed text, particularly in low-resource and under-resource languages for various applications [5] [6] [7] [2] [3] [2] [8]. To address the challenges of LI in code-mixed text, several studies have been conducted employing various ML and Deep Learning (DL) algorithms. Gundapu and Mamidi [9] performed LI on TeluguEnglish code-mixed text using Conditional Random Fields (CRF) classifiers and obtained an accuracy of 91.28% by considering previous, current, and next words, their POS tags, word length, and character n-grams in the range (1, 3) as features. Veena et al. [10] explored SVM models trained with word and character 5-gram embeddings, for LI in code-mixed Hindi-English text and achieved better accuracy.

Chaitanya et al. [11] proposed word-level LI model in Hindi-English code-mixed data using word embeddings (Continuous Bag of Words (CBOW) and Skip-gram) to effectively capture word semantics and relationships. They achieved 67.33% and 67.34% for CBOW and Skip-gram model respectively using Support Vector Machine (SVM) classifiers. Veena et al.[12] implemented a word-level LI system for code-mixed Malayalam-English and Tamil-English Facebook data and generated character embedding features using skip-gram architecture. They employed 10-fold cross-validation to train and evaluate the SVM model, ensuring the robust performance of the model, and obtained 93% and 95% accuracies for Malayalam-English and Tamil-English text, respectively.

In the context of word-level LI in code-mixed Kannada-English texts, particularly during the ICON 2022 competition, a notable trend emerged. As documented by Balouchzahi et al. [2], teams that used neural network (NN) architectures and transformer-based models consistently outperformed traditional machine learning classifiers and baseline models. Hegde et al. [7] provided an in-depth overview of the methodologies and outcomes of the "CoLI-Tunglish: Word-level Language Identification in Code-mixed Tulu Texts" shared task. This task featured participation from five distinct teams, each employing LI in code-mixed Tulu texts. Among the approaches, a machine learning model that employed a stacking ensemble of multiple classifiers trained on character n-grams emerged as the top performer. This model achieved a significant macro F1 score of 0.813, highlighting its effectiveness in addressing the complexities of code-mixed Tulu text processing

3. Datasets

The shared task is built on a multilingual corpus of user-generated social media text spanning five Dravidian and Indo-Aryan languages: Kannada (KAN), Malayalam (MAL), Telugu (TL), Tamil (TM), and

[1] https://www.codabench.org/competitions/7902/

Dataset Split	EN	LANG	SYM	OTHER	NAME	MIXED/TMEN	LOCATION / PLACE	NUMBER
KANNADA (KAN)								
Train	17905	4058	3833	2588	1353	1031	142	–
Dev	922	546	273	47	50	176	2	–
Test	1204	289	264	117	55	141	5	–
MALAYALAM (ML)								
Train	5762	11749	2930	2084	1985	761	123	601
Dev	407	860	217	263	149	61	8	43
Test	380	938	211	164	158	88	8	50
TELUGU (TL)								
Train	3326	1248	789	377	189	64	27	260
Dev	301	57	57	52	11	32	–	5
Test	238	148	63	19	15	4	2	5
TAMIL (TM)								
Train	2756	7062	1189	82	1134	1253	11	–
Dev	496	1000	183	1	160	144	–	–
Test	534	986	230	16	139	152	9	–
TULU								
Train	7451	11644	4254	658	1511	540	526	–
Dev	742	1251	422	85	135	57	41	–
Test	813	1330	455	59	133	65	56	–

Table 1
Distributions of label tags in datasets for all languages. "LANG" indicates the language-specific tag (KAN, MAL, TL, TM, TULU) for each dataset.

Tulu. The raw data were sourced from YouTube comments collected from web scraper and underwent preprocessing steps such as script normalization, tokenization, and removal of sensitive information. Each dataset is annotated at the token level using a unified labeling scheme to support consistent cross-language modeling.

The tagset includes:

- LANG, which represents the primary language label of the respective dataset (KAN, MAL, TL, TM, or TULU),
- EN for English tokens, SYM for symbols, emoji, punctuation, and informal markers,
- NUMBER for numeric expressions,
- NAME for named entities (primarily person names),
- LOCATION for geographic references,
- MIXED for intra-token code-mixed forms,
- and OTHER for residual or ambiguous cases.

Annotation was carried out by native speaker of each language following unified guidelines. Table 1 presents the distribution of tags across Train, Development, and Test splits for all languages. In the shared-task setup, the Train and Development sets were released with gold-standard labels, whereas the Test sets were provided in blinded form to enable unbiased system evaluation.

4. System Description

This section gives concise system descriptions for each of the eight teams that submitted. These descriptions highlight the core methodologies and performance highlights of their approaches to the shared tasks.

Team 1: The authors employ a SVM-based framework enriched with morpho-phonological features designed specifically for Dravidian languages. They combine TF-IDF vectorization over character

(1–4) and word (1–2) n-grams with affix patterns and phonotactic cues to handle agglutination and Romanization variability. A class-weighted, linear-kernel SVM (C=5) optimized on the Kannada word-level LI dataset delivers strong performance, achieving 0.911 Macro F1 securing 4th rank in the shared task.

Team 2: This working note, the authors describe a word-level LI system for the Dravidian languages Tamil, Telugu, Malayalam, Kannada, and Tulu by combining TF-IDF character n-grams in range 2 to 5, handcrafted linguistic features like word length, capitalization and alphanumeric check, and FastText embeddings. Their proposed system obtained macro F1 scores of 0.908, 0.921, 0.734,0.818, and 0.791 for Tulu, Kannada, Tamil, Malayalam, and Telugu respectively securing first place in Tulu and Kannada tracks and second place in rest of the tracks.

Team 3: The authors of this system, approached the shared task with a traditional yet robust pipeline built around language-specific feature engineering and Conditional Random Fields (CRF). They designed lexical, orthographic, and contextual features tailored to the morphological properties and word formation patterns of each Dravidian language, enabling the models to adapt effectively to code-mixed and noisy inputs. Separate CRF models were trained for each language, leveraging these handcrafted features to capture fine-grained linguistic cues often missed by end-to-end neural systems. Their proposed system obtained macro F1 scores of 0.729, 0.796, 0.762, 0.904, and 0.596 for Telugu, Tulu, Malayalam, Kannada, and Tamil languages securing 4th, 5th, 6th, 6th, and 8th ranks, respectively, in the shared task.

Team 4: The authors fine-tuned LaBSE on the shared-task datasets for the five Dravidian languages. The model was trained for 10 epochs using Adam optimizer with a batch size of 32, where the classification head was adapted to the number of language labels in the dataset (four or five). Input words were tokenized, embeddings were generated, and the [CLS] representation from LaBSE was passed through the classifier to predict the language of each word. The system achieved strong results, with macro-F1 scores of 0.8995 for Kannada, 0.7434 for Tamil, 0.8271 for Malayalam, 0.9515 for Telugu, and 0.8224 for Tulu, securing 1st place for Tamil, Malayalam, and Telugu, 2nd place for Tulu, and 7th place for Kannada on the leaderboard.

Team 5: The authors propose a lightweight approach to word-level LI across five Dravidian languages (Kannada, Malayalam, Telugu, Tamil, and Tulu) using character-level TF-IDF features combined with classical machine learning classifiers. Words are vectorized as character n-grams (1–4), transformed into sparse TF-IDF matrices, and then passed to classifiers, with ExtraTrees showing the strongest performance. Label encoding is used to convert language tags into numeric form for supervised learning. The method was evaluated on the shared-task datasets, achieving macro-F1 scores of 0.8987 for Kannada, 0.7938 for Malayalam, 0.7084 for Tamil, 0.7572 for Telugu, and 0.7925 for Tulu. On the leaderboard, this corresponded to 8th place for Kannada, 4th for Malayalam, 3rd for Tamil and Telugu, and 6th for Tulu.

Team 6: In this work, authors present a system that combines mBERT and GRU models (mBERT+GRU) for word-level LI in five Dravidian languages—Kannada, Malayalam, Tamil, Telugu, and Tulu. This model merges the contextual multilingual transformer embeddings from mBERT with GRU-based sequential modeling to learn both global and local linguistic patterns in code-mixed text. Training is performed using a learning rate of 2e-5, weight decay of 0.01, batch size of 16, and up to 150 epochs with early stopping to avoid overfitting. The Focal Loss and oversampling techniques are used to combat class imbalance, with post-processing prediction cleaning aimed at removing invalid or irrelevant tags. Evaluation on the official FIRE 2025 shared task dataset elicits competitive performance across all languages, with macro F1 scores of 0.642, 0.620, 0.493, 0.732, and 0.791 for Malayalam, Tamil, Telugu, Tulu and Kannada languages respectively obtaining 7th, 7th 8th, 8th, and 9th ranks in the shared task.

Team 7: The author employs a CountVectorizer-based representation using character n-grams in the range of 1–4 to capture subword patterns common in code-mixed Dravidian text. These features are fed into a linear SVM classifier. The model is trained on the shared task datasets and optimized for word-level LI across multiple Dravidian languages. Their proposed model obtained macro F1 scores of 0.917, 0.817, 0.790, 0.684, and 0.706 for Kannada, Tulu, Malayalam, Tamil, and Telugu languages securing 3rd, 3rd, 5th, 5th, and 6th ranks respectively, in the shared task.

Table 2
Rank lists of Malayalam and Tamil languages

Malyalam					Tamil			
Username	M.F1	Accuracy	Rank		Username	M.F1	Accuracy	Rank
wordsj	0.827	0.884	1		wordsj	0.743	0.925	1
sruthi_s	0.818	0.883	2		sruthi_s	0.734	0.900	2
nguyentriet	0.808	0.873	3		sharal	0.708	0.898	3
sharal	0.794	0.874	4		nguyentriet	0.690	0.891	4
poorvi	0.790	0.869	5		poorvi	0.684	0.915	5
rachanabn	0.762	0.857	6		ewenburban	0.642	0.864	6
suhani_verma	0.642	0.833	7		suhani_verma	0.620	0.897	7
ewenburban	0.621	0.820	8		rachanabn	0.596	0.875	8

Table 3
Rank lists of Telugu and Tulu languages

Telugu					Tulu			
Username	M.F1	Accuracy	Rank		Username	M.F1	Accuracy	Rank
wordsj	0.952	0.968	1		sruthi_s	0.823	0.908	1
sruthi_s	0.791	0.957	2		wordsj	0.822	0.903	2
sharal_coelho	0.757	0.929	3		poorvi	0.817	0.906	3
rachanabn	0.729	0.921	4		nguyentriet	0.809	0.897	4
nguyentriet	0.715	0.935	5		rachanabn	0.796	0.883	5
poorvi	0.706	0.933	6		sharal_coelho	0.793	0.894	6
ewenburban	0.613	0.879	7		ewenburban	0.758	0.850	7
suhani_verma	0.493	0.891	8		suhani_verma	0.732	0.856	8

5. Ranking and Findings

Out of thirty-three registrations for the shared task, Ten teams took part, and seven teams submitted their working notes. Tables 2, 3 and 4 show the results of the models submitted by the participants, and the ranks decided by their macro F1 scores (M.F1). The top-performing models achieved macro F1 scores of 0.921, 0908, 0.952, 0.827, and 0.743 for code-mixed Kannada, Tulu, Telugu, Malayalam, and Tamil code-mixed texts respectively, emphasizing the challenges and accomplishments of the word-level LI task.

Most of the teams employed a variety of ML models (SVM, ExtraTrees, etc.) for LI in code-mixed text. In addition, participants also explored Conditional Random Fields (CRF) and fine-tuned LaBSE, mBERT+GRU models. Further, ML models proposed by the participants are commonly trained with TF-IDF of character n-grams, CountVectorizer-based representation using character n-grams. The models proposed and the features employed by the participating teams highlight the scarcity of computational resources for handling Dravidian texts.

The team (Team 2) that utilized TF-IDF character n-grams (2,5), handcrafted linguistic features (word length, capitalization and alphanumeric check) and FastText embeddings outperformed the other models, including fine-tuned LaBSE, CRF model. Most participating teams opted for language-independent features (TF-IDF of character n-grams) rather than investigating the possibility of very few available pre-trained models.

6. Conclusion and Future Works

Although LI is frequently overlooked in low-resource languages, it is an essential first step for many NLP projects. The amount of text data in low-resource languages has significantly increased as a result of recent technological developments, especially on social media platforms where code-mixed content—a combination of local and regional languages and English—is frequently found. Word-level LI is required in code-mixed texts when many languages are combined at the word level. For various

Table 4
Rank list of Kannada languages

Kannada			
Username	**M.F1**	**Accuracy**	**Rank**
sruthi_s	0.921	0.960	1
nguyentriet	0.918	0.962	2
poorvi	0.917	0.962	3
abdollah	0.911	0.955	4
ewenburban	0.910	0.956	5
rachanabn	0.904	0.959	6
wordsj	0.900	0.968	7
sharal_coelho	0.899	0.944	8
suhani_verma	0.791	0.936	9

ranges of ”n,” the majority of teams have investigated machine learning models trained with TF-IDF of character n-grams. The results obtained by the models of the participating teams suggest a promising avenue for addressing LI challenges in low-resource and code-mixed language scenarios. As future work, we plan to extend the task by incorporating additional languages to widen its applicability.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat GPT-4 in order to:Grammar and spelling check. Paraphrasing was handled via QuillBot. With this tool, the author(s) reviewed and revised the content as required, while assuming full responsibility for the publication’s integrity.

References

[1] V. Kolipakam, F. M. Jordan, M. Dunn, S. J. Greenhill, R. Bouckaert, R. D. Gray, A. Verkerk, A Bayesian Phylogenetic Study of the Dravidian Language Family, Royal Society open science 5 (2018) 171504.

[2] 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.

[3] 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.

[4] N. Sushma, A. Hegde, H. L. Shashirekha, Word-level language identification in code-mixed tulu texts., in: FIRE (Working Notes), 2023, pp. 213–222.

[5] A. Hegde, F. Balouchzahi, S. Butt, S. Coelho, K. G, H. S. Kumar, S. HL, A. Agrawal, Coli@ fire2024: Findings of word-level code-mixed language identification in dravidian languages, in: Proceedings of the 16th Annual Meeting of the Forum for Information Retrieval Evaluation, 2024, pp. 7–10.

[6] A. Hegde, F. Balouchzahi, S. Coelho, H. Shashirekha, H. A. Nayel, S. Butt, Overview of coli-tunglish: Word-level language identification in code-mixed tulu text at fire 2023., in: FIRE (Working Notes), 2023, pp. 179–190.

[7] A. Hegde, F. Balouchzahi, S. Coelho, S. HL, H. A. Nayel, S. Butt, Coli@ fire2023: Findings of word-level language identification in code-mixed tulu text, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, 2023, pp. 25–26.

[8] H. L. Shashirekha, F. Balouchzahi, M. D. Anusha, G. Sidorov, Coli-machine learning approaches

for code-mixed language identification at the word level in kannada-english texts, arXiv preprint arXiv:2211.09847 (2022).

[9] S. Gundapu, R. Mamidi, Word level language identification in english telugu code mixed data, in: Proceedings of the 32nd Pacific Asia conference on language, information and computation, 2018.

[10] P. Veena, M. Anand Kumar, K. Soman, Character embedding for language identification in hindi-english code-mixed social media text, Computación y Sistemas 22 (2018) 65–74.

[11] I. Chaitanya, I. Madapakula, S. K. Gupta, S. Thara, Word level language identification in code-mixed data using word embedding methods for indian languages, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2018, pp. 1137–1141.

[12] P. Veena, M. A. Kumar, K. Soman, An effective way of word-level language identification for code-mixed facebook comments using word-embedding via character-embedding, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2017, pp. 1552–1556.