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

Asha Hegde[1], Fazlourrahman Balouchzahi[2], Sabur Butt[3], Sharal Coelho[1], Kavya G[1], Harshitha S Kumar[1], Sonith D[1], Shashirekha Hosahalli Lakshmaiah[1] and Ameeta Agrawal[4]

[1]*Department of Computer Science, Mangalore University, India,*
[2]*CIC, IPN, Mexico,*
[3]*IFE, Tecnologico de Monterrey, Mexico,*
[4]*Department of Computer Science, Portland State University, USA*

Abstract

Language Identification (LI) traditionally focuses on detecting languages in documents/sentences, primarily for high-resource languages like English, Spanish, German, and French. However, with growing technological advancements, LI challenges in multilingual countries like India, where users often create code-mixed content by blending local languages with English, have gained prominence. One such example is the combination of Dravidian languages: Tamil, Kannada, Malayalam, and Tulu, with English resulting in code-mixed texts. These code-mixed texts demand LI at word-level to analyze and process them under multilingual settings and acts as a preliminary step for many applications. Code-mixed Dravidian languages are rarely explored in the context of word-level LI. To address this lacuna, CoLI-Dravidian shared task focuses on word-level LI in code-mixed datasets of four Dravidian languages: Tamil, Kannada, Malayalam, and Tulu, written in Roman script. Participants of CoLI-Dravidian shared task are assigned the task of categorizing each word in the given sequence into one of the predefined categories. Out of ten teams who submitted the predictions of their models, the top-performing models achieved macro F1 scores of 0.7656, 0.9293, 0.8939, and 0.8678 for code-mixed Tamil, Kannada, Malayalam, and Tulu texts respectively, highlighting the difficulty and success of the task.

Keywords

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

1. Introduction

Dravidian languages, a family of approximately 80 languages spoken by more than 220 million people in South Asia, have a rich and ancient history. A recent study suggests that the Dravidian language family, which includes major languages such as Tamil, Telugu, Kannada, and Malayalam, is around 4,500 years old [1]. People speaking these local, native or regional languages, are at ease using even English for everyday communication. These multilingual individuals often prefer to use multiple scripts and languages when sharing their thoughts and opinions on social media platforms. As a result, code-mixing has become the standard linguistic practice on social media these days [2]. Code-mixing can occur at various levels, including the paragraph, sentence, or word level, and can even extend to the subword level [3]. One of the primary tasks in computational linguistics in multilingual scenario is to identify the language of each word in code-mixed sentences. LI is crucial as it enables the development of more accurate Natural Language Processing (NLP) tools, which can be applied in various applications such as machine translation, sentiment analysis, and social media monitoring [4].

To tackle the challenges of word-level LI in Dravidian languages, we organized a shared task titled ”CoLI-Dravidian: Word-level Code-Mixed Language Identification in Dravidian Languages”[1] as part of Forum for Information Retrieval Evaluation (FIRE) 2024[2]. The CoLI-Dravidian 2024 shared task provides code-mixed datasets in four languages - Kannada, Tamil, Malayalam, and Tulu - aiming to foster the

FIRE'24: Forum for Information Retrieval and Evaluation, December 12-15, 2024, DA-IICT, Gandhinagar, India

hegdekasha@gmail.com (A. Hegde); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); saburb@tec.mx (S. Butt); sharalmucs@gmail.com (S. Coelho); kavyamujk@gmail.com (K. G); harshiskumar94@gmail.com (H. S. Kumar); sonithksd@gmail.com (S. D); hlsrekha@mangaloreuniversity.ac.in (S. H. Lakshmaiah); ameeta@pdx.edu (A. Agrawal)

[1] https://sites.google.com/view/coli-dravidian-2024/home

[2] https://fire.irsi.org.in/fire/2024/home

development of advanced models for LI in these languages. The task was organized into two main phases: training and validation, followed by testing. In the first phase, participants were given labeled training and validation sets in the four languages to build and test their models respectively. During the testing phase, unlabeled test sets were provided in these languages and participants were required to run their models on the tests sets and submit their predictions via Codalab platform[3] for evaluation. The participating teams were given opportunity to make up to five submissions per language and the best result for each language was used for the final ranking. The predictions were evaluated based on macro averaged precision, recall, F1 score, and accuracy, and the final ranking was based on macro averaged F1 score. Out of 37 teams registered for this shared task, 10 teams submitted their predictions making it to the final rankings and 8 teams submitted the working notes.

Rest of the paper is organized as follows: an overview of previous shared tasks on word-level LI in Dravidian languages and the various approaches used by participants are briefed in Related Works - section 2. The datasets used in current version of the task, together with their description and statistics, are detailed in Datasets - section 3. A discussion of different models submitted by the participants is presented in System Description - section 4, followed by the final rankings and results in Ranking - section 5. Finally, the findings are discussed in Findings - section 6, and Conclusion and Future Works - section 7 outline the overall conclusions and potential directions for future research.

2. Related Works

Code-mixing has emerged as the default language of communication on social media allowing blending of words/sub-words from multiple languages and has gathered significant research attention, especially in the area of word-level LI with several notable studies contributing to the understanding of this complex linguistic behavior. Recently several studies have focused on LI tasks in code-mixed Dravidian languages. The description of CoLi-Kanglish [2] and CoLi-Tunglish [5] - our earlier shared tasks on word-level LI, and the summary of the models submitted to this shared task are given below:

2.1. CoLI-Kanglish 2022

In CoLI-Kanglish - a shared task [2] on word-level LI in Kannada-English code-mixed texts, participants were tasked with identifying each word belonging to one of six categories: Kannada, English, Kannada-English, Name, Location, and Other. The dataset was built by processing around 100,000 comments from Kannada YouTube videos and words in the dataset were annotated with six categories. Thirty submissions received from eight teams used several Machine Learning (ML) and Deep Learning (DL) models, including transformers like Distil Bidirectional Encoder Representations from Transformers (BERT) and multilingual (mBERT) and the best-performing model achieved an averaged macro F1 score of 0.62. Models utilizing neural networks and transformers generally outperformed traditional ML classifiers. Table 1 presents statistics of the dataset used in this shared task and descriptions of the best performing models are presented below:

Vajrobol [6] employed fine-tuning a DistilBERT-cased model - a pre-trained transformer model for CoLI-Kanglish task. Their model performed exceptionally well, achieving the highest averaged macro F1 score of 0.62 in the competition. The team's approach of leveraging a pre-trained transformer model proved effective in tackling the complex nature of code-mixed texts. Tonja et al. [7] explored a variety of transformer models (BERT, mBERT, Robustly Optimized BERT Pretraining Approach (RoBERTa), and Cross-lingual Language Modeling-RoBERTa (XLM-R)) in combination with Long Short-Term Memory (LSTM) architecture to capture word-level dependencies in code-mixed Kannada-English dataset. Among these models, their proposed BERT model demonstrated the best performance, achieving averaged macro F1 score of 0.61. Their extensive experimentation with multiple transformer models positioned them second in the overall ranking, highlighting the effectiveness of multilingual transformers for this task. Yigezu et al. [8] focused on character-level models by implementing LSTM and Bidirectional LSTM

[3] https://codalab.lisn.upsaclay.fr/competitions/19357

Table 1
Statistics of CoLI-Kanglish 2022 shared task dataset

Category	Train Set	Test Set
Kannada (kn)	6,526	2,194
English (en)	4,469	1,812
Kannada-English (kn-en)	1,379	93
Name	708	354
Location	102	31
Other	1,663	100
Total	**14,847**	**7,241**

(BiLSTM) architectures with attention mechanisms, designed to read text as a sequence of characters. BiLSTM model outperformed the LSTM, likely due to its ability to capture more complex patterns in code-mixed text and the attention mechanism further enhanced the model's ability to focus on important parts of the text. Their model achieved an averaged macro F1 score of 0.61, placing them in a tie for second place with Tonja et al. [7]. Deka et al. [9] experimented with multiple transformer models for LI and among the models they experimented, BERT-based model demonstrated solid performance, securing an averaged macro F1 score of 0.57. This placed them fourth in the overall rankings. Their approach showcased the strength of transformer models in handling code-mixed text, particularly in identifying Kannada and English at the word level.

2.2. CoLI-Tunglish 2023

Hegde et al. [5] presented the CoLI-Tunglish shared task, which focuses on word-level LI in code-mixed Tulu texts [10]. This task aims to assign one of six predefined categories to each word in code-mixed Tulu-Kannada-English texts written in Roman script. The dataset used in this shared task consists of user-generated comments from YouTube, which were tokenized and annotated by native speakers. The final dataset includes words categorized into Tulu, Kannada, English, mixed-language words, names, locations, and other categories and the mixed category posed challenges due to its complexity. The shared task attracted 14 teams, with 10 different submissions from 5 teams. Most teams used traditional ML methods exploring Support Vector Machine (SVM), k-Nearest Neighbors (kNN), and Random Forest (RF), trained on character n-grams and one team used Transfer Learning (TL) approach with mBERT. The highest-performing team, achieved a macro F1 score of 0.813 with a context-sensitive Logistic Regression (LR) model trained on character n-grams. Table 2 presents statistics of the dataset used in this shared task and the descriptions of the best performing models are presented below:

Table 2
Statistics of CoLI-Tunglish 2023 shared task dataset

Category	Train Set	Development Set	Test Set
Tulu	8,647	1,461	4,118
English	5,499	889	2,617
Kannada	2,068	344	1,173
Name	1,104	162	513
Other	506	102	200
Mixed	403	69	194
Location	369	54	190
Total	**18,596**	**3,081**	**9,005**

Bestgen [11] developed two systems for the CoLI-Tunglish task: a basic system and a context-sensitive one. The basic system used a LIBLinear L2-regularized LR model trained on character n-grams ranging from 1 to 5. The context-sensitive system built upon the basic system trained the LR model with additional context-based information. Their approach was highly effective, achieving the highest macro F1 score of 0.813 and securing first rank in the shared task. The team's use of both basic and context-sensitive models demonstrated the importance of incorporating contextual information for word-level LI in code-mixed text. Fetouh and Nayel [12] explored a variety of ML models, including SVM, Stochastic Gradient Descent (SGD), kNN, and Multilayer Perceptron (MLP). These models were trained on Term Frequency-Inverse Document Frequency (TF-IDF) of character n-grams in the range of 1 to 4, along with word length as an additional feature. Among their experiments, SVM model performed the best, achieving a macro F1 score of 0.812, placing them second in the competition. Shetty [13] used TF-IDF of character n-grams in the range of 1 to 4 to train a range of models (Multinomial Naive Bayes (MNB), RF, LR, LinearSVC, Decision Tree (DT), kNN, AdaBoost, One Vs Rest, and Gradient Boost). Among the models proposed, LinearSVC model achieved a macro F1 score of 0.799 placing them in third position. The author's experimentation with multiple classifiers and n-gram ranges showcased the value of using robust ML models to handle the challenges of word-level LI in code-mixed data. Chanda et al. [14] adopted a TL approach by fine-tuning mBERT model to generate word embeddings for Tulu code-mixed text and applied a softmax activation function to obtain language predictions for each word. By tuning the hyperparameters of BiLSTM layer added to the mBERT model, the team achieved a macro F1 score of 0.602, placing fifth in the competition. While their approach using TL with mBERT was novel, it did not outperform the traditional ML models used by other teams, indicating the complexity of code-mixed text handling in low-resource languages.

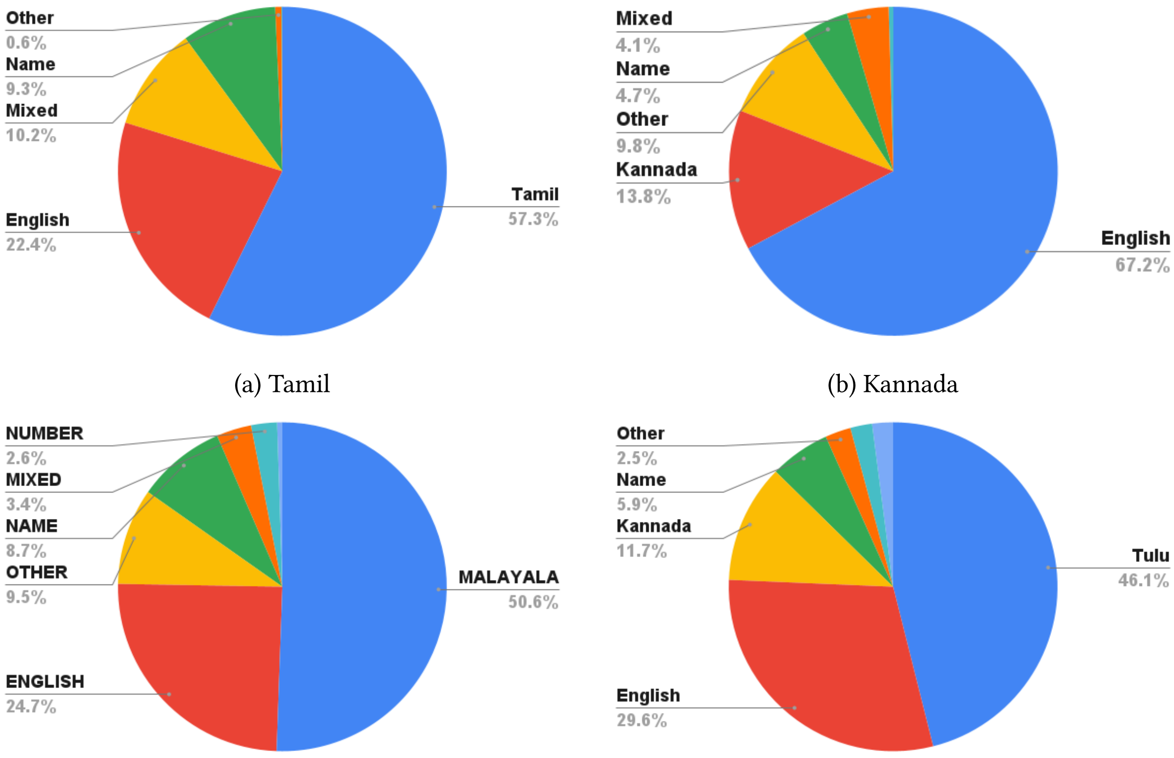
The literature review summarizes that the word-level LI shared tasks in Kannada and Tulu languages have given ample opportunities for researchers to process code-mixed texts and explore various learning models for word-level LI in these languages.

3. CoLI-Dravidian 2024 Dataset

In continuation with our earlier shared tasks - CoLi-Kanglish [2] and CoLi-Tunglish [5], CoLI-Dravidian 2024 shared task aims to advance research in word-level LI in four code-mixed Dravidian languages - Tamil, Kannada, Malayalam, and Tulu. The goal of this shared task is to invite researchers to develop models that categorize each word in the given text into one of the predefined labels: Tamil/Kannada/-Malayalam/Tulu/English, mixed language content (Mixed), named entities such as names (Name) and locations (Location), numbers (Number), and words that do not fit to any category (Other). While Tamil, Kannada, and Malayalam datasets have two distinct language classes: Tamil/Kannada/Malayalam and English, Tulu dataset has three distinct languages: Tulu, Kannada, and English. Digits are denoted as 'Number', 'Name' class is assigned to person names, 'Location' class is used for geographical locations, 'Mixed' class is designated for words that blend words/suffixes from Dravidian languages and/or English language in any order and the remaining words fall into 'Other' class for unclassified terms. The 'Mixed' category presents a significant challenge for LI task because these words are formed by the combination of Dravidian languages and/or English words, often mixed with corresponding affixes (prefixes and suffixes) from these languages. The beauty and complexity of these mixed-language words emerge from the unique word patterns created by social media users highlighting the diversity and adaptability of language in digital communication.

To address word-level LI in code-mixed Kannada, Tamil, and Malayalam texts, YouTube comments were collected using a custom-built scraper. The comments underwent pre-processing to remove punctuation and control characters, followed by tokenization into individual words. Each word was then manually annotated by a native speaker fluent in the regional language (Kannada, Tamil, or Malayalam) and English. Further, the dataset used in CoLi-Tunglish[4] 2023 shared task is used for word-level LI in code-mixed Tulu text in this shared task also [4, 5]. This task challenges the researchers to create models that effectively handle the linguistic complexity and diversity of code-mixed Dravidian texts. The statistics of the class-wise distribution of the Coli-Dravidian datasets are shown in Figure 1.

[4] https://sites.google.com/view/coli-tunglish/home



(a) Tamil

(b) Kannada

(c) Malayalam

(d) Tulu

Figure 1: Statistics of CoLi-Dravidian datasets

4. System Description

To benchmark datasets used in Coli-Dravidian shared task, experiments were conducted with different ML classifiers (SVM, MLP, DT, LR, RF, and AdaBoost) trained with TF-IDF of character n-grams in the range (1, 5). Among these classifiers, SVM, LR, and DT performed better and are therefore used as baselines for the shared task. More than 100 distinct predictions per language were submitted by 10 different teams. The description of models submitted by the participants and their performances are as follows:

Team PonsubashRaj explored MNB, LR, DT, SVM, and voting classifiers, trained with count vectorizers and TF-IDF, of character sequences. Their proposed voting classifiers trained with count vectorizers of character sequences secured 1st, 5th, 2nd, and 4th ranks for Tamil, Kannada, Malayalam, and Tulu texts respectively.

Team Kaivalya fine-tuned Multilingual Representations for Indian Languages (MuRiL) and mBERT pre-trained models for word-level LI task for all the four languages and found that MuRiL models outperformed mBERT models achieving 3rd, 1st, 2nd, and 2nd ranks for Tamil, Kannada, Malayalam, and Tulu texts respectively.

Team NLPnorth used MACHAMP[5] model to fine-tune a wide range of transformer models and picked the best five language models based on their performances on the development sets. Further, they added Conditional Random Field (CRF) layer to the MACHAMP model to capture the likelihood between the consecutive words and obtained 4th, 2nd, 1st, and 1st ranks for Tamil, Kannada, Malayalam, and Tulu texts respectively.

Team Awsathama conducted a wide range of experiments using ML classifiers (MNB, LR, Support Vector Classifier (SVC), kNN, DT, RF, Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Categorical Boosting (CatBoost)) trained with count vectors and TF-IDF

[5] https://github.com/machamp-nlp/machamp

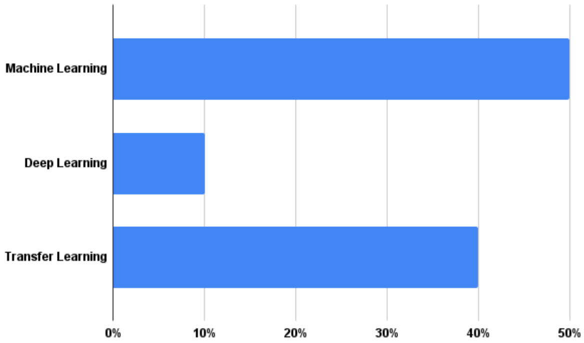


Figure 2: Learning approaches used by participants

vectors, of character sequences. Their proposed XGBoost and SVC models trained with count vectors obtained 2nd and rank 3rd ranks for Tamil and Malayalam texts respectively. Further, SVC models trained with TF-IDF vectors obtained 3rd and 4th ranks for Tulu and Kannada texts respectively.

Team MUCS employed deep neural network models to implement two sequence labeling models: i) CoLi_CNN - a Convolutional Neural Network (CNN) model trained with MuRiL embeddings and ii) CoLi_TNN - a transformer neural networks trained from scratch, and a sequence-to-sequence learning model with BiLSTM encoder and LSTM decoder. Their proposed CoLi_CNN model obtained 6th rank for all the four languages.

Team MUCSNLPLAB trained CRF models with text features - word length, previous word, next word, and by tuning the hyperparameters, their proposed models obtained 3rd, 7th, 9th, and 7th ranks for Tamil, Kannada, Malayalam, and Tulu texts respectively.

Team TextTitans proposed a prompt based method using GPT 3.5 `turbo`, a large language model, to perform word-level LI in Tamil and Kannada texts and obtained 10th rank for both the languages.

Team abadian trained ML classifiers (SVM, SGD, kNN, and MLP) with TF-IDF of character sequences for word-level LI in Kannada and Malayalam texts and their proposed SVM model obtained 8th and 10th ranks for Kannada and Tulu texts respectively.

The findings reveal that a significant number of participants experimented with different transformer models, while few others opted for traditional ML techniques, and a smaller group focused on DL models. This diversity in approaches highlights the evolving landscape of the techniques used in the shared task.

5. Ranking

Conventionally, word-level LI datasets are imbalanced and this can skew the model evaluation. Hence, using both macro and weighted F1 scores provide a more comprehensive assessment, as macro treats all classes equally and weighted-average accounts for class imbalance based on their frequency. Together, these metrics offer a better evaluation of model performance across all the classes. The predictions submitted by the participants of the shared task was evaluated based on macro F1 scores to rank the teams and ranking ties are resolved considering the weighted F1 score. Tables 3 and 4 presents the performances of the participating teams in the shared task along with the baselines.

The top four teams surpassed the baseline models, achieving higher macro F1 scores of 0.7656, 0.9293, 0.8939, and 0.8678 for code-mixed Tamil, Kannada, Malayalam, and Tulu texts, respectively, reflecting the difficulty and competitiveness of the shared task. This result underscores the advancement made by the top teams in addressing the task's challenges.

Table 3
Rank list of Tamil and Kannada languages

Team	M_F1	M_Pr	M_Re	W_F1	W_Pr	W_Re	Acc	Rank
Tamil								
PonsubashRaj	0.7657	0.7603	0.7884	0.9173	0.9197	0.9165	0.9165	1
Awsathama	0.7301	0.7784	0.7041	0.9205	0.9192	0.9234	0.9234	2
Team_Kaivalya	0.7233	0.7190	0.7290	0.9338	0.9322	0.9358	0.9358	3
NLPnorth	0.7179	0.7297	0.7088	0.9355	0.9332	0.9387	0.9387	4
denis_gordeev	0.6999	0.6912	0.7133	0.9402	0.9439	0.9373	0.9373	5
MUCS	0.6994	0.7343	0.6997	0.9257	0.9279	0.9279	0.9279	6
Srihari V K	0.6833	0.7738	0.6587	0.9076	0.9066	0.9125	0.9125	7
SVM-Baseline	0.6817	0.8773	0.6369	0.8960	0.9004	0.9037	0.9037	-
DT-Baseline	0.6534	0.6945	0.6373	0.8761	0.8742	0.8794	0.8794	-
CUFE	0.6279	0.6260	0.6319	0.8759	0.8761	0.8760	0.8760	8
MUCSNLPLab	0.6080	0.6269	0.5958	0.8594	0.8562	0.8641	0.8641	9
LR-Baseline	0.5891	0.7262	0.5582	0.8564	0.8693	0.8735	0.8735	-
TextTitans	0.3312	0.3260	0.3658	0.7022	0.7559	0.6690	0.6690	10
Kannada								
Team_Kaivalya	0.9294	0.9242	0.9359	0.9669	0.9671	0.9668	0.9668	1
NLPnorth	0.9225	0.9240	0.9211	0.9606	0.9605	0.9608	0.9608	2
MUCSNLPLab	0.8688	0.9010	0.8469	0.9287	0.9311	0.9293	0.9293	3
Awsathama	0.8570	0.8482	0.8716	0.9413	0.9435	0.9404	0.9404	4
PonsubashRaj	0.8516	0.8718	0.8580	0.9370	0.9430	0.9380	0.9380	5
MUCS	0.8400	0.8304	0.8582	0.9346	0.9382	0.9333	0.9333	6
Srihari V K	0.8205	0.8115	0.8401	0.9224	0.9278	0.9209	0.9209	7
abadian	0.8192	0.8909	0.7792	0.9130	0.9172	0.9173	0.9173	8
SVM-Baseline	0.8185	0.8899	0.7787	0.9121	0.9163	0.9165	0.9165	-
LR-Baseline	0.7633	0.8816	0.7092	0.8643	0.8750	0.8749	0.8749	-
DT-Baseline	0.7619	0.7496	0.7862	0.8657	0.8739	0.8661	0.8661	-
denis_gordeev	0.7186	0.7063	0.7593	0.9264	0.9405	0.9209	0.9209	9
TextTitans	0.4494	0.5475	0.4242	0.6725	0.7191	0.6994	0.6994	10
Macro F1-Score (M_F1), Macro precision (M_Pr), Macro Recall (M_Re), Weighted F1-Score (W_F1), Weighted precision (W_Pr), Weighted Recall (W_Re) and Accuracies (Acc)								

6. Findings

37 teams registered for this shared task and 10 teams submitted their results for all the four languages. Figure 2 gives a glimpse of the number of teams and the learning approaches used by them to address word-level LI. Most of the teams have incorporated ML models using language-independent feature extraction techniques, like TF-IDF and CountVectorizer, while few teams have leveraged TL to improve their model's performance for low-resource languages like Tulu. This approach demonstrates the flexibility of the models in handling languages that are not part of the original training data. Only one team has employed DL models by incorporating MuRiL embbedings - a language dependent representation and Keras embeddings - a language independent representation. Their proposed methodology found DL classifier trained with MuRiL embeddings to be more beneficial for performing the word-level LI task. This suggests that language-specific embeddings like MuRiL can provide a significant advantage in handling tasks for specific languages.

Participants have also encountered challenges while working with code-mixed text in Roman script. To overcome this, they have either fine-tuned suitable pre-trained models for the datasets or employed language-independent feature extraction methods. However, language dependent resources for Tulu remain limited compared to other languages. Further, the issue of extreme class imbalance in the given datasets is not addressed by any of the participants.

Table 4
Rank list of Malayalam and Tulu languages

Team	M_F1	M_Pr	M_Re	W_F1	W_Pr	W_Re	Acc	Rank
Malayalam								
NLPnorth	0.8939	0.9092	0.8840	0.9389	0.9397	0.9400	0.9400	1
PonsubashRaj	0.8755	0.9192	0.8499	0.9230	0.9259	0.9264	0.9264	2
Awsathama	0.8688	0.9159	0.8381	0.9221	0.9235	0.9268	0.9268	3
SVM-Baseline	0.8396	0.9295	0.7913	0.9104	0.9184	0.9184	0.9184	-
Srihari V K	0.8396	0.8859	0.8111	0.8967	0.8980	0.9040	0.9040	4
DT-Baseline	0.8259	0.8641	0.7990	0.8927	0.8931	0.8968	0.8968	-
CUFE	0.8047	0.8745	0.7614	0.8848	0.8849	0.8892	0.8892	5
MUCS	0.8028	0.8861	0.7751	0.9086	0.9105	0.9132	0.9132	6
MUCSNLPLab	0.7671	0.8970	0.7061	0.8688	0.8790	0.8792	0.8792	7
Ram	0.7635	0.7964	0.7483	0.8152	0.8310	0.8093	0.8093	8
LR-Baseline	0.7496	0.9274	0.7035	0.8717	0.8932	0.8880	0.8880	-
denis_gordeev	0.6050	0.6295	0.5905	0.9166	0.9215	0.9192	0.9192	9
abadian	0.1067	0.1063	0.1319	0.2683	0.2721	0.3031	0.3031	10
Tulu								
NLPnorth	0.8679	0.8798	0.8589	0.9168	0.9183	0.9162	0.9162	1
Team_Kaivalya	0.8585	0.8988	0.8282	0.9179	0.9191	0.9187	0.9187	2
Awsathama	0.8390	0.8872	0.8036	0.9085	0.9095	0.9111	0.9111	3
PonsubashRaj	0.8157	0.8426	0.7988	0.8952	0.8961	0.8961	0.8961	4
Srihari V K	0.7888	0.8316	0.7585	0.8769	0.8767	0.8803	0.8803	5
MUCS	0.7854	0.8224	0.7659	0.8769	0.8799	0.8779	0.8779	6
SVM-Baseline	0.7853	0.8931	0.7330	0.8851	0.8909	0.8906	0.8906	-
MUCSNLPLab	0.7717	0.8674	0.7187	0.8736	0.8797	0.8791	0.8791	7
DT	0.7706	0.7839	0.7599	0.8666	0.8661	0.8678	0.8678	-
denis_gordeev	0.7470	0.7616	0.7362	0.9052	0.9097	0.9022	0.9022	8
LR-Baaseline	0.6931	0.8889	0.6462	0.8472	0.8638	0.8605	0.8605	-
PNB	0.2664	0.2697	0.2656	0.4418	0.4386	0.4465	0.4465	9

7. Conclusion and Future Works

This paper describes Coli-Dravidian 2024 - a word-level LI shared task and presents findings of the task. The task is focused on four low-resource Dravidian languages - Tamil, Kannada, Malayalam, and Tulu, intertwined with English, reflecting the real-world linguistic dynamics of multilingual communities in the digital age. Further, it underscores the importance of recognizing the unique characteristics of these low-resource languages and highlights the efforts to preserve linguistic diversity in an increasingly interconnected world.

The fine-tuned MuRiL model excelled for Kannada, achieving the highest macro F1 score of 0.9294, and also performed well for Tulu with a macro F1 score of 0.8585, underscoring its versatility in handling less commonly studied languages in the Dravidian family. For Malayalam, the MACHAMP model, with an added CRF layer, achieved the best result with a macro F1 score of 0.8939, showcasing its effectiveness in capturing language sequences. In case of Tamil, a voting classifier trained on character sequences produced the highest score of 0.7656, which highlights the need for further refinement of models for this language, potentially through more sophisticated contextual understanding. The effectiveness of these methods depends heavily on the linguistic and code-mixing properties of each Dravidian language.

By using the datasets of this shared task, researchers can focus on adding more context and improving transformer models to better understand the unique details of Dravidian languages in real-world tasks like sentiment analysis, translation, and monitoring social media. The shared task's outcomes emphasize the importance of continued research into code-mixed LI, which is crucial for preserving linguistic diversity in the digital age.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT in order to: Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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, Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts (2022) 38.

[3] H. Shashirekha, LAs for HASOC-Learning Approaches for Hate Speech and Offensive Content Identification., in: In FIRE (working notes), 2020, pp. 145–151.

[4] 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.

[5] 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.

[6] V. Vajrobol, CoLI-Kanglish: Word-Level Language Identification in Code-Mixed Kannada-English Texts Shared Task using the Distilka Model, 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. 7–11.

[7] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbukh, Transformer-based Model for Word Level Language Identification in Code-mixed Kannada-English Texts, 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. 18–24.

[8] M. G. Yigezu, A. L. Tonja, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbukh, Word Level Language Identification in Code-mixed Kannada-English Texts using Deep Learning Approach, 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. 29–33.

[9] P. Deka, N. J. Kalita, S. K. Sarma, BERT-based Language Identification in Code-Mix Kannada-English Text at the CoLI-Kanglish Shared Task@ 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. 12–17.

[10] A. Hegde, F. Balouchzahi, S. Coelho, S. H L, 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, FIRE '23, Association for Computing Machinery, 2024, p. 25–26.

[11] Y. Bestgen, Using Character Ngrams for Word-Level Language Identification in Trilingual Code-Mixed Data (and Even More)., in: FIRE (Working Notes), 2023, pp. 191–197.

[12] A. M. Fetouh, H. Nayel, BFCAI at CoLI-Tunglish@ FIRE 2023: Machine Learning Based Model for Word-level Language Identification in Code-mixed Tulu Texts., in: FIRE (Working Notes), 2023, pp. 205–212.

[13] P. Shetty, Word-Level Language Identification of Code-Mixed Tulu-English Data., in: FIRE (Working Notes), 2023, pp. 198–204.

[14] S. Chanda, A. Mishra, S. Pal, Advancing Language Identification in Code-Mixed Tulu Texts: Harnessing Deep Learning Techniques., in: FIRE (Working Notes), 2023, pp. 223–230.