Overview of CoLI-Tunglish: Word-level Language Identification in Code-mixed Tulu Text at FIRE 2023

Asha Hegde[1], F. Balouchzahi[2], Sharal Coelho[1], H.L. Shashirekha[1], Hamada A. Nayel[3] and Sabur Butt[2]

[1] *Department of Computer Science, Mangalore University, India*
[2] *Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico*
[3] *Department of Computer Science Faculty of Computers and Artificial Intelligence, Benha University, Egypt*

Abstract

Word-level Language Identification (LI) aims to identify the language of individual words within a given sentence. It is a preliminary step in processing code-mixed text in which words or sub-words belonging to more than one language are used in a sentence/words/sub-words, for various applications. Though there are several tools/models for word-level LI for high-resource languages, under-resourced languages like Tulu, Kannada etc., are less explored in this direction due to lack of annotated data. To address these challenges, we have open-sourced a Tulu code-mixed dataset (a combination of Tulu, Kannada, and/or English words/sub-words/affixes) for word-level LI of Tulu, Kannada, English, and mixed-language words, written in Roman script in the CoLI-Tunglish shared task. The objective of the shared task is to assign one of the six predefined categories: Tulu, Kannada, English, Mixed (a combination of Tulu, Kannada, and/or English languages), Name, Location, and Other, to each word in a given sentence. A total of 14 teams had registered for the shared task and 10 different runs were submitted by 5 teams. Most of the teams have explored Machine Learning (ML) classifiers trained Term Frequency - Inverse Document Frequency (TF-IDF) of character n-grams. The top-performing model obtained weighted F1 score and macro F1 score of 0.89 and 0.81, respectively, among all the models submitted by the participants.

Keywords

Language Identification, Tulu, Sequence Labeling, Word-level

CEUR-WS.org/Vol-3681/T4-1.pdf

1. Introduction

Globally, South Asia stands out as the most linguistically diverse region boasting an astonishing array of over 650 distinct languages[1]. India, as a prominent South Asian country encapsulates the linguistic richness within its borders, with a rich tapestry of languages reflecting its cultural heritage and diversity. Tulu is one of the Dravidian languages having a rich cultural and literary

Forum for Information Retrieval Evaluation, December 15-18, 2023, India

hegdekasha@gmail.com (A. Hegde); fbalouchzahi2021@cic.ipn.mx (F. Balouchzahi); sharalmucs@gmail.com (S. Coelho); hlsrekha@mangaloreuniversity.ac.in (H.L. Shashirekha); hamada.ali@fci.bu.edu.eg (H. A. Nayel); sbutt2021@cic.ipn.mx (S. Butt)
https://sites.google.com/view/asha-hegde/home (A. Hegde); https://sites.google.com/view/fazlfrs/home (F. Balouchzahi); https://sites.google.com/view/sharalcoelho/home (S. Coelho); https://bu.edu.eg/staff/hamadaali14 (H. A. Nayel)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://www.deccanherald.com/content/ 652273/intl-meet-south-asian-languages.html

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

heritage and spoken by a community of over 4 million native speakers [1] in the coastal regions of the southern part of India, predominantly in Karnataka state [2]. Despite its significant speaker base, Tulu is facing the challenges of recognition and preservation for which a lot of efforts are ongoing to promote and sustain this unique linguistic tradition. As Tigalari - the Tulu script is not used much, Tulu text is often written in Kannada script. Further, Tulu was traditionally a spoken language, and since Kannada is taught from an early age, transcribing Tulu in Kannada script became widespread [3].

Tulu is the regional language of Dakshina Kannada and Kannada is the official language of Karnataka. Tuluvas (people whose mother tongue is Tulu) usually know both Tulu and Kannada languages fluently to read, write, and speak. In addition, many Kannada words are used in Tulu language. Moreover, English is widely spoken among Tulu-speaking individuals, particularly among those who are active on social media platforms. Tulu content such as songs, videos, movies, comedy programs, and skits, are immensely popular on social media and comments posted by Tulu users often comprise a mix of Tulu, Kannada, and/or English. Due to the limitations of technology in computer keyboards and smartphone keypads and the intricacies of composing words with consonant conjuncts in Kannada script, many Tulu users opt to employ Roman script or a combination of Kannada and Roman script when interacting on social media, resulting in code-mixed text [4]. This code-mixing can occur at various linguistic levels, including the paragraph, sentence, word, or sub-word, where users blend their native and/or local language like Tulu and/or Kannada with English [5, 6]. Due to the prevalence of Roman alphabets on computer keyboard layouts and smartphone keypads, people often prefer to write code-mixed content in Roman script rather than their native script.

Social media platforms have granted users the liberty to compose text informally, often disregarding the grammar conventions of the specific languages used. This has led to a substantial influx of user-generated content characterized by incomplete words or sentences, catchy phrases, user-defined abbreviations ("gm" for "good morning"), slang terms ("meme", "Gmeet", "WhatsApp"), common abbreviations ("OMG" for "Oh my God"), and the repetition of characters ("soooooo sad" for "so sad"), among others [7, 8]. These informal language elements can make the content challenging to comprehend. Additionally, the prevalence of code-mixing, where words of one language interwoven with words of another language as prefixes or suffixes, complicates text analysis particularly due to conflicting phonetics. The expanding user base on social media platforms results in a continuous surge of user-generated content, making manual management and understanding of this text increasingly impractical. This underscores the need for automated tools and techniques capable of processing user-generated code-mixed text.

The preliminary step in handling code-mixed text for many of the Natural Language Processing (NLP) tasks like Machine Translation [9], Parts-Of-Speech tagging [10], Sentiment Analysis [11, 12], Emotion Analysis [13, 14], Detecting Sign of Depression [15], Hate Speech and Offensive Language Identification [6, 16], Hope Speech Detection [17, 18], etc., is identifying the language of each word/phrase/sentence [7] and this task is known as a Language Identification (LI). Traditionally, LI has been predominantly studied at the document level, with a focus on high-resource languages, often overlooking low-resource languages. However, in recent times, due to technological advancements and the multilingual nature of countries like India, there has been a growing trend of users posting comments in code-mixed texts [7, 19]. Some of the prominent code-mixed Indian languages are: Hindi-English [19], Bengali-English [20], Kannada-English [7], Telugu-English [21], and Malayalam-English [22]. These code-mixed texts demand LI at word-level as each word in the text belongs to anyone language or combination of languages. Identifying the language of the words in code-mixed social media text gives insight into the linguistic intervention and can also be helpful in multilingual text processing.

Word-level LI can be modeled as sequence labeling problem, where each word in the sequence is tagged with one of the predefined languages including mixed language. Inspired by Shashirekha et al. [7], to address the challenges of word-level LI in code-mixed text, CoLI-Tunglish shared task introduces a gold standard corpus for word-level LI in Tulu code-mixed text. The objective of this task is to determine the language of each word in the given Tulu code-mixed data sourced from social media text [4]. CoLI-Tunglish dataset serves as a valuable resource for researchers and practitioners working on word-level LI in multilingual contexts, allowing them to develop and evaluate models that can effectively handle code-mixed data.

The rest of the paper is organized as follows: Section 2 describes the related work and Section 3 describes the task description. Section 4 gives details about the evaluation metrics followed by the brief description about the baselines in Section 5. Overview of the submitted systems are described in Section 6 and Results are discussed in Section 7. The paper concludes in Section 8 along with some future avenues.

2. Related Work

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 [4, 5, 8, 23]. To address the challenges of LI in code-mixed text, several studies have been conducted employing various ML and Deep Learning (DL) algorithms and the description of some relevant works are given below:

Chaitanya et al. [19] have explored LI of Hindi-English code-mixed data, employing feature vectors generated by the Continuous Bag of Words (CBOW) and Skipgram models, to train ML models (Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), k-Nearest Neighbor (kNN), and Adaptive Boosting (AdaBoost)). Among these models, SVM classifiers achieved highest accuracies of 67.33% and 67.34% using CBOW and Skipgram models, respectively. Gundapu and Mamidi [24] performed LI on Telugu-English 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. Mandal and Singh [25] proposed a multichannel Neural Network (NN) model of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) models combined with Bidirectional LSTM (BiLSTM) and CRF, for LI in code-mixed Hindi-English and Bengali-English text. This multichannel NN model achieved accuracies of 93.32% and 93.28% for Hindi-English and Bengali-English data, respectively. Thara and Poornachandran [22] introduced a dataset for LI in code-mixed English-Malayalam text and utilized transformer-based model with fine-tuned Enhanced Light Efficiency Cophasing Telescope Resolution Actuator (ELECTRA) model and obtained a best performance with a macro F1 score of 0.9933. Veena et al. [26] explored SVM models trained with word and character 5-gram embeddings, for LI in code-mixed Hindi-English text and achieved better accuracy.

Table 1
Statistics of CoLI-Kenglish dataset

Tag	Train set	Test set
Kannada	6,526	2,194
English	4,469	1,812
Kannada-English	1,379	93
Name	708	354
Location	102	31
Other	1,663	100
Total	**14,847**	**7,241**

To address the specific challenge of word-level LI in Kannada-English code-mixed texts, our previous work - the CoLI-Kanglish shared task [23], aimed to provide a solution by open-sourcing a dataset comprising Kannada-English code-mixed text written in the Roman script [7]. The task's objective was to classify each word within the text into one of six predefined categories: Kannada, English, Kannada-English, Name, Location, or Other. The CoLI-Kenglish dataset used in CoLI-Kanglish shared task [23] is described in [7] and the statistics of the dataset are given in Table 1. The study reported the performance of various models submitted by participants in the CoLI-Kanglish shared task. Table 2 borrowed from [23] shows the final leaderboard in the CoLI-Kanglish shared task and the summary of some top-performing models is given below:

Team Tiya1012 [27] achieved the top position in the competition by fine-tuning DistilBERT - a transformer-based model on the CoLI-Kenglish dataset and obtained a macro F1 score of 0.62 indicating promising progress in the field of word-level LI for code-mixed texts. Team Abyssinia [28] conducted experiments using various Language Models (LM) (Bidirectional Encoder Representations from Transformers (BERT), Multilingual BERT (mBERT), XLM-R, and RoBERTa from HuggingFace) in combination with a LSTM architecture. Notably, mBERT and XLM-R outperformed the other models, achieving a macro F1 score of 0.61 and securing the second rank in the competition. Team PDNJK [29] explored multiple transformer-based models for the LI task in code-mixed Kannada-English words. Their top-performing model based on BERT achieved a macro F1 score of 0.57, earning them the fourth position in the shared task. Team Habesha [30] took a different approach by training character-level LSTM and BiLSTM models with attention mechanisms. Their BiLSTM model outperformed the LSTM model, achieving a macro F1 score of 0.61 and securing the second place in the competition. Team Lidoma [31] investigated the use of character n-grams to generate character TF-IDF representation for training traditional ML classifiers. Among their experiments, a simple kNN classifier performed the best, achieving a macro F1 score of 0.58. Team NLP_BFCAI [32] converted Bag-of-Characters into character vectors and introduced a character representation model known as Bag-of-n-Characters. They experimented with several traditional ML algorithms and found that the RF model, utilizing the proposed features, achieved a macro F1 score of 0.43 in the competition.

To summarize, a considerable amount of research works are reported on word-level LI in code-mixed Indo-Aryan texts like Hindi-English and Bengali-English. However, word-level LI in code-mixed Dravidian language (Tamil, Malayalam, Kannada, and Telugu) texts are seen with very limited attention in this direction. Further, it is also clear that word-level LI in code-mixed

Tulu text has not yet been explored by the researchers and this is the first-ever research attempt that focuses on word-level LI in code-mixed Tulu text.

Table 2
Results of the CoLI-Kenglish shared task [23]

Rank	Team name	Weighted			Macro		
		Precision	Recall	F1-score	Precision	Recall	F1-score
1	Tiya1012 [27]	0.87	0.85	0.86	0.67	0.61	0.62
2	Abyssinia [28]	0.85	0.84	0.84	0.62	0.62	0.61
2	Habesha [30]	0.85	0.83	0.84	0.66	0.6	0.61
-	LSVM-Baseline	0.84	0.84	0.83	0.67	0.57	0.59
3	Lidoma [31]	0.83	0.83	0.83	0.64	0.56	0.58
4	PDNJK [29]	0.86	0.85	0.86	0.58	0.58	0.57
-	MLP-Baseline	0.84	0.81	0.82	0.60	0.60	0.57
-	LR-Baseline	0.84	0.84	0.83	0.69	0.53	0.56
5	NLP_BFCAI [32]	0.73	0.73	0.72	0.52	0.41	0.43
6	iREL	0.68	0.62	0.64	0.38	0.45	0.39
7	JUNLP	0.69	0.67	0.67	0.33	0.34	0.3
8	PresiUniv	0.57	0.59	0.53	0.22	0.22	0.2

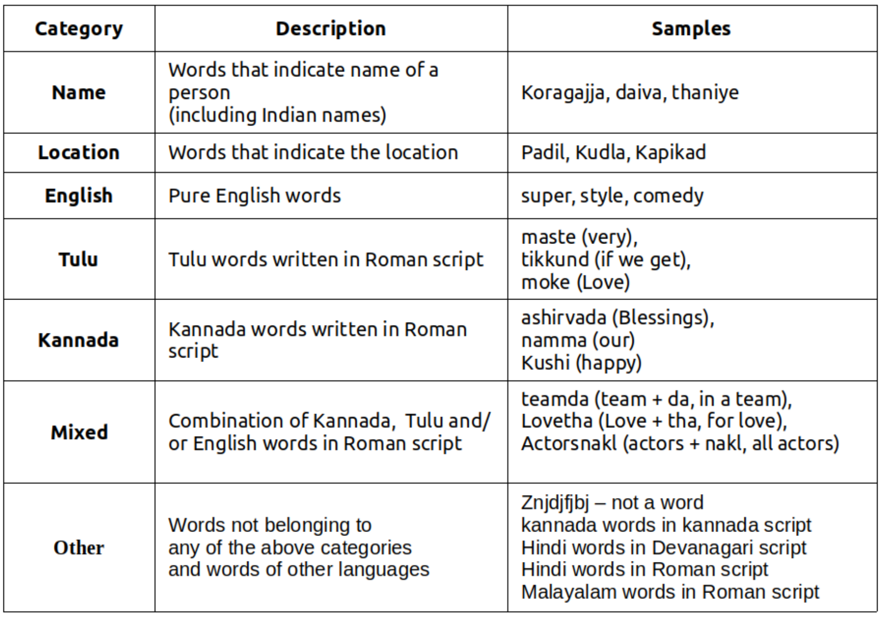
3. Task Description

To address word-level LI in code-mixed Tulu texts, CoLI-Tunglish dataset is constructed by using the YouTube comments collected by Hegde et al. [4]. The comments are preprocessed by removing digits, punctuation, and control characters, and the remaining content is tokenized into individual words. These words are then manually annotated by native Tulu speakers who have fluency in both Kannada and English.

Inspired by Balouchzahi et al. [23], the aim of the CoLI-Tunglish task is to promote research in word-level LI in Tulu - a low-resource Indian language. Participants are invited to use the dataset comprising of Tulu, Kannada, and English language content and develop models to categorize each word in the dataset into one of English, Tulu, Kannada, a mixture of two or three of the above languages (Mixed), a Named Entity denoting a name (Name) or location (Location), or designated as "other" (Other), categories. The Coli-Tunglish dataset consists of words categorized into three distinct language classes: “Tulu”, “Kannada”, and “English”, to denote words from these respective languages. The code-mixed nature is depicted by the “Mixed” class which is designated for words that blend word/prefixes/suffixes from Tulu, Kannada, and/or English languages in any order. Further, while “Name” class is assigned to the name of a person, “Location” class is used for geographical or place names, and any other words fall into the category of “Other” class. The “Mixed” category in the dataset presents a significant challenge for the LI task because these words are formed by combining Tulu, Kannada, 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. The categories, description of the categories and the sample tokens of the

CoLI-Tunglish dataset are shown in Table 3 and statistics of the class-wise distribution of the CoLI-Tunglish dataset is shown in Table 4.

Table 3
Description and samples tokens of the classes in CoLI-Tunglish dataset



Category	Description	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 4
Class-wise distribution of Train, Development, and Test set

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

4. Evaluation Metrics

In an imbalanced dataset, categories with a larger number of samples may affect the weighted F1 scores. The model may achieve high accuracy by simply predicting the majority class, and hence, the evaluation measure of accuracy may be misleading. Further, the weighted F1 score that gives the average weight of the number of samples available in that class fails to address data imbalance. On the other hand, the macro F1 score is often used in evaluating models trained on imbalanced data as they provide a balanced assessment of model performance across all classes, regardless of class distribution. Further, the macro F1 score gives equal importance to each class, making it a suitable metric to evaluate model performance in scenarios where class imbalances exist. As CoLI-Tunglish dataset is imbalanced, macro F1 score is used to evaluate the performance of the submitted models. The classification report[2] tool which provides comprehensive metrics and insights for evaluating the performance of the systems available at Scikit-learn library is used to compute macro F1 score.

5. Baselines

To benchmark the CoLI-Tunglish dataset, several ML classifiers (Multinomial Naive Bayes (MNB), SVM, Multilayer Perceptron (MLP), Decision Tree (DT), LR, RF, and Adaboost) are trained with TF-IDF of character n-grams in range (1, 3) considering top 5,000 features. Among these ML classifiers, as RF, DT, and SVM models, gave better performance, they are used as baselines for CoLI-Tunglish shared task.

6. Overview of the Submitted Systems

A total of ten different runs were submitted by five different teams for the CoLI-Tunglish 2023 shared task and all five teams submitted their working notes. While 90% of the participants experimented different ML models, 10% of the participants implemented Transfer Learning (TL) approach. A summary of the models submitted by all five teams is given below:

Team SATLAB developed two different working systems: i) Basic System: LIBLinear L2-regularized LR model trained with character n-grams in the range (1, 5) and ii) Context-Sensitive System: LIBLinear L2-regularized LR model trained with the output obtained by the Basic System. Their Context-sensitive system achieved a macro F1 score of 0.813 and secured the 1st rank in the competition.

Team BFCAI explored ML models (SVM, Stochastic Gradient Descent, kNN and MLP) trained with TF-IDF of character n-grams with range (1, 4) and word length. Among their experiments, SVM model performed the best and secured the 2nd rank, achieving a macro F1 score of 0.812.

Team Poorvi used TF-IDF of character n-grams with various ranges to train MNB, RF, LR, LinearSVC, DT, kNN, AdaBoost, One Vs Rest, and Gradient Boost. Their LinearSVC model trained with TF-IDF of character n-grams in the range (1, 4) was the most effective configuration observed for the word-level LI task and their proposed model obtained macro F1 score of 0.799 securing 3rd rank in the shared task.

Team MUCS proposed three different models: i) CRF model trained with text-based features (word, Length of word, Beginning of sentence, End of sentence, etc), ii) Ensemble of ML classifiers (SVM, LR, and RF) with hard voting trained with fastText embeddings for words, and characters embeddings for Roman letters, and iii) Ensemble of ML classifiers (SVM, LR, and RF) with hard voting trained with TF-IDF of character n-grams, for the given datasets. Among all the models, the highest macro F1 score of 0.77 was reported for CRF model securing 4th rank.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

Team IRLab@IITBHU used a two-step process for LI. They leveraged the mBERT model to obtain word embeddings and then applied a softmax activation function to obtain language predictions for each word in the code-mixed Tulu text. By fine-tuning mBERT on the shared task dataset and tuning hyperparameters for the Bi-LSTM layer, their model achieved a macro F1 score of 0.602 and placed 5th rank in the shared task.

7. Results and Discussion

The best macro F1 score achieved by each team along with the macro F1 scores of three baselines shown in Table 5 provides a comprehensive comparison of the performances of the submitted models in the shared task against the baselines. This comparison reveals that the four teams achieved better macro F1 scores than the baseline models. The highest macro F1 score of 0.813 highlights the challenging nature of the shared task. Further, among the three baselines (RF, DT, and SVM) trained with character n-grams in the range (1, 3), the RF classifier achieved a better macro F1 score of 0.744 for LI in code-mixed Tulu text.

Most of the teams employed a variety of ML models (SVM, LR, RF, kNN, MLP, MNB, DT, and One vs Rest) for LI in code-mixed Tulu text. In addition, participants also explored boosting classifiers (Stochastic Gradient Descent, Adaboost, and Gradient Boost) to enhance the performance of the classifiers. Among all the participants, only one team implemented the mBERT model based on the TL approach. Further, ML models proposed by the participants are commonly trained with TF-IDF of character n-grams and two submissions used the pre-trained models for feature extraction. The TL-based model utilizes the features fine-tuned with mBERT model to train the Bi-LSTM classifier. The proposed models and the features used by the participating teams reveal the lack of computational tools in processing code-mixed Tulu text. The team that utilized an ML classifier trained on TF-IDF of character sequences, coupled with a feature selection method, outperformed the other models, including the mBERT model. This result underscores the significance of tailored feature engineering and selection strategies.

It is noteworthy that most participating teams opted for language-independent features (TF-IDF of character n-grams) rather than exploring the potential of very few available pre-trained models. Surprisingly, no specific methods such as sub-word level representation, normalization, or character-level representation are explored by the participants in this shared task to directly address the challenges posed by code-mixed texts. This indicates a gap in leveraging specialized techniques for handling linguistic variations in multilingual data.

8. Conclusion

LI serves as a crucial initial step for numerous NLP tasks, but is often neglected in low-resource languages. The recent technological advancements have led to a significant surge in the volume of text data in low-resource languages, particularly on social media platforms where code-mixed content - a blend of local/regional languages and English, is quite common. The combination of more than one language at word-level necessitates word-level LI in code-mixed texts. The primary objective of the CoLI-Tunglish shared task was to promote word-level LI in code-mixed Tulu texts. This task attracted considerable interest initially, with 14 teams expressing their

Table 5
Results of CoLI-Tunglish shared task

Rank	Team Name	Weighted			Macro		
		Precision	Recall	F1 score	Precision	Recall	F1 score
1	SATLAB	0.898	0.901	0.898	0.851	0.783	**0.813**
2	BFCAI	0.899	0.902	0.899	0.859	0.777	0.812
3	Poorvi	0.891	0.893	0.891	0.821	0.781	0.799
4	MUCS	0.874	0.876	0.873	0.807	0.743	0.770
-	RF-Baseline	0.859	0.861	0.854	0.841	0.693	0.744
-	DT-Baseline	0.828	0.832	0.830	0.701	0.691	0.696
-	SVM-Baseline	0.816	0.821	0.807	0.793	0.593	0.639
5	IRLab@IITBHU	0.843	0.857	0.838	0.740	0.571	0.602

intent to participate, ultimately resulting in the submission of ten distinct runs from five different teams. Most of the teams have explored ML models trained with TF-IDF of character n-grams, for different ranges of “n”. This underscores the limited availability of resources for the Tulu language.

An ML model of stacking of ML classifiers trained with character n-grams emerged as the top performer, achieving a notable macro F1 score of 0.813. This outcome reveals the significance of effective feature engineering and highlights the substantial difficulty of the task, given the complexities introduced by code-mixing in Tulu texts. 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. Word-level LI for other Dravidian languages including Tamil, Telugu and Malayalam, will be addressed in future.

References

[1] A. Hegde, H. L. Shashirekha, A. K. Madasamy, B. R. Chakravarthi, A Study of Machine Translation Models for Kannada-Tulu, in: Congress on Intelligent Systems, Springer Nature Singapore Singapore, 2022, pp. 145–161.

[2] S. B. Steever, The Dravidian Languages, Routledge, 2019.

[3] K. Padmanabha, A Comparative Study of Tulu Dialects, Mangalore, 1990.

[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] B. R. Chakravarthi, R. Priyadharshini, V. Muralidaran, N. Jose, S. Suryawanshi, E. Sherly, J. P. McCrae, Dravidiancodemix: Sentiment Analysis and Offensive Language Identification Dataset for Dravidian Languages in Code-Mixed Text, in: Language Resources and Evaluation, Springer, 2022, pp. 765–806.

[6] F. Balouchzahi, H. L. Shashirekha, G. Sidorov, A. Gelbukh, A Comparative Study of Sylla-

bles and Character Level N-grams for Dravidian Multi-script and Code-Mixed Offensive Language Identification, in: Journal of Intelligent & Fuzzy Systems, IOS Press, 2022, pp. 1–11.

[7] 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, in: Acta Polytechnica Hungarica, 2022, pp. 123–141.

[8] A. Hegde, H. L. Shashirekha, Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic/Transphobic Content in Code-mixed Dravidian Languages, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2022.

[9] I. Jadhav, A. Kanade, V. Waghmare, S. S. Chandok, A. Jarali, Code-Mixed Hinglish to English Language Translation Framework, in: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 684–688. doi:10.1109/ICSCDS53736.2022.9760834.

[10] K. Akhil, R. Rajimol, V. Anoop, Parts-of-Speech Tagging for Malayalam using Deep Learning Techniques, in: International Journal of Information Technology, Springer, 2020, pp. 741–748.

[11] S. Thara, P. Poornachandran, Social Media Text Analytics of Malayalam–English Code-Mixed using Deep Learning, in: Journal of big Data, Springer, 2022, p. 45.

[12] F. Balouchzahi, H. Shashirekha, LA-SACo: A Study of Learning Approaches for Sentiments Analysis in Code-Mixing Texts, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 109–118.

[13] S. Ghosh, A. Priyankar, A. Ekbal, P. Bhattacharyya, Multitasking of Sentiment Detection and Emotion Recognition in Code-Mixed Hinglish Data, 2023, pp. 110–182.

[14] A. Hegde, H. L. Shashirekha, Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2022.

[15] A. Hegde, S. Coelho, A. E. Dashti, H. Shashirekha, MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, 2022, pp. 312–316.

[16] A. Hegde, M. D. Anusha, H. L. Shashirekha, Ensemble Based Machine Learning Models for Hate Speech and Offensive Content Identification, in: Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org, 2021.

[17] F. Balouchzahi, G. Sidorov, A. Gelbukh, PolyHope: Two-Level Hope Speech Detection from Tweets, in: Expert Systems with Applications, 2023, p. 120078.

[18] A. Hande, R. Priyadharshini, A. Sampath, K. P. Thamburaj, P. Chandran, B. R. Chakravarthi, Hope Speech Detection in Under-Resourced Kannada Language, in: arXiv preprint arXiv:2108.04616, 2021.

[19] 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.

[20] S. Mandal, A. K. Singh, Language Identification in Code-Mixed Data using Multichannel

Neural Networks and Context Capture, in: W-NUT 2018, 2018, p. 116.

[21] 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.

[22] S. Thara, P. Poornachandran, Transformer Based Language Identification for Malayalam-English Code-Mixed Text, in: IEEE Access, IEEE, 2021, pp. 118837–118850.

[23] 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: Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts, 2022, p. 38.

[24] S. Gundapu, R. Mamidi, Word Level Language Identification in English Telugu Code Mixed Data, in: arXiv preprint arXiv:2010.04482, 2020.

[25] S. Mandal, A. K. Singh, Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture, in: Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Association for Computational Linguistics, 2018, pp. 116–120.

[26] P. Veena, M. Anand Kumar, K. Soman, Character Embedding for Language Identification in Hindi-English Code-mixed Social Media Text, in: Computación y Sistemas, Instituto Politécnico Nacional, Centro de Investigación en Computación, 2018, pp. 65–74.

[27] 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 7–11. URL: https://aclanthology.org/2022.icon-wlli.2.

[28] A. Lambebo Tonja, M. Gemeda Yigezu, O. Kolesnikova, M. Shahiki 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 18–24. URL: https://aclanthology.org/2022.icon-wlli.4.

[29] P. Deka, N. Jyoti Kalita, S. Kumar 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 12–17. URL: https://aclanthology.org/2022.icon-wlli.3.

[30] M. Gemeda Yigezu, A. Lambebo Tonja, O. Kolesnikova, M. Shahiki 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 29–33. URL: https://aclanthology.org/2022.icon-wlli.6.

[31] M. Shahiki Tash, Z. Ahani, A. Tonja, M. Gemeda, N. Hussain, O. Kolesnikova, Word

Level Language Identification in Code-Mixed Kannada-English Texts using Traditional Machine Learning Algorithms, 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 25–28. URL: https://aclanthology.org/2022.icon-wlli.5.

[32] S. Ismail, M. K. Gallab, H. Nayel, BoNC: Bag of N-Characters Model for Word Level Language Identification, 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, Association for Computational Linguistics, IIIT Delhi, New Delhi, India, 2022, pp. 34–37. URL: https://aclanthology.org/2022.icon-wlli.7.