# Word-Level Language Identification of Code-Mixed Tulu-English Data

Poorvi Shetty[1]

[1]JSS Science and Technology University, Mysuru, India

Abstract

Code-mixing, the amalgamation of languages in speech, particularly common in India, generates informal, multilingual content on social media. Analyzing this content for linguistic tasks, notably Language Identification, is crucial. This study focuses on word-level Language Identification in Tulu-English code-mixed words, using diverse embeddings and classifiers. Results show promising accuracy, affirming the viability of the proposed approach, with the best system achieving a weighted average F1 score of 0.799. The study enhances multilingual processing by providing insights into effective language identification in complex linguistic scenarios, with broader implications for communication understanding in multilingual societies. The proposed system ranked 3rd in the shared task.

**Keywords**

language identification, code-mixing, multilingual communication, word embeddings, classifiers, Tulu-English, code-mixed words, multilingual processing

## 1. Introduction

Language Identification (LID) in Natural Language Processing (NLP) refers to the process of determining the natural language in which a given piece of text is written. It involves analyzing various linguistic features and patterns within the text to accurately determine the language it belongs to.

Tulu, along with the state language Kannada is part of the cultural and linguistic landscape of Karnataka, India. Those proficient in Tulu, known as Tuluvas, commonly exhibit fluency in both Tulu and Kannada, encompassing reading, writing, and verbal communication. Moreover, the Tulu language incorporates numerous lexical elements from Kannada. Additionally, the usage of English characters holds prominence among many Tulu speakers, particularly those active on social media platforms. Notably, the commentary contributed by Tulu users in response to Tulu-focused content on social media platforms often manifests as a linguistic amalgamation, involving Tulu, Kannada, and English. This intricate linguistic phenomenon has given rise to a valuable collection of trilingual code-mixed data, an area that has remained relatively unexplored within the realm of research. [1, 2]

This paper delves into the realm of word-level LID within the context of code-mixed Tulu-English (Tu-En) textual compositions. These textual instances have been sourced from com-

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

poorvishetty1202@gmail.com (P. Shetty)
0009-0004-2243-5176 (P. Shetty)

mentary sections of Tulu YouTube videos, consequently facilitating the construction of the Code-mixed Tulu-English Language Identification (CoLI-Tunglish) dataset. This task was part of the Word-level Language Identification in Code-mixed Tulu Texts (CoLI-Tunglish) shared task[3]. A similar shared task CoLI-Kanglish (Kannada and English) was conducted last year [4].

## 2. Related Work

In addressing the challenge of code-mixed language identification, several researchers have contributed innovative approaches. Gundapu and Mamidi [5] introduced Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) for English-Telugu code-mixed data, ultimately finding success with CRFs. Sabty et al. [6] focused on Arabic-English (AR-EN) text and found Segmental Recurrent Neural Networks (SegRNN) to excel in intra-word language identification. Mandal et al. [7] presented supervised learning methods for Bengali-English code-mixed data, utilizing character-based and root phone-based encodings in deep Long Short-Term Memory (LSTM) models.

In the realm of code-mixed language identification, researchers have explored various methodologies. Ojo et al. [8] delved into code-mixed Kannada and English (Kn-En) texts, achieving high accuracy with their CK-Keras model, incorporating pre-trained Word2Vec embeddings. Tonja et al. [9] introduced a Transformer-based model for word-level language identification in code-mixed Kannada-English texts. Uchoi and Kaur [10] combined language-specific morphological dictionary-based approaches with character n-gram language models to achieve precise word classification in English and Punjabi code-mixed sentences.

Researchers have developed versatile approaches to address code-mixed language identification across various languages and contexts. Chittaranjan et al. [11] presented a CRF-based system that incorporates lexical, contextual, character n-gram, and special character features, applicable to multiple languages. Gella et al. [12] tackled language identification in concise code-mixed documents across 28 languages. Sarma et al. [13] addressed word-level language identification in a multilingual context, proposing and evaluating strategies for low-resource languages like Assamese, Bengali, Hindi, and English.

Studies have also explored the effectiveness of BERT and Transformer models in code-mixed language identification. Hidayatullah et al. [14] demonstrated the superiority of fine-tuned IndoBERTweet models, utilizing sub-word language representations for accurate language identification. Shashirekha et al. [15] created the CoLI-Kenglish dataset and employed various models, with the CoLI-ngrams model standing out as superior. Vajrobol [16] utilized transformer-based techniques, fine-tuning the DistilBERT model to discern the language of individual words within code-mixed Kannada-English texts using the Distilka model.

## 3. Existing Dataset

The Code-mixed Tulu-English Language Identification (CoLI-Tunglish) dataset [2] consists of Tulu, Kannada, and English words in Roman script and is grouped into seven major categories, namely, ”Tulu”, “Kannada”, “English”, “Mixed-language”, “Name”, “Location” and “Other”. These

**Table 1**
Classwise distribution within the training set of the dataset provided by Hegde et al.

| Category | Count |
|---|---|
| Name | 8647 |
| Location | 5499 |
| English | 2068 |
| Tulu | 1104 |
| Kannada | 506 |
| Mixed | 403 |
| Other | 369 |

texts are extracted from Tulu YouTube video comments, a rich source of trilingual code-mixed data.

## 4. Data Preprocessing

The undertaken methodologies within this study encompassed initial preprocessing steps, including the conversion of the provided text to lowercase, followed by its representation as strings for further analysis. The following embedding techniques were employed and tested to encapsulate the inherent linguistic characteristics of the text data:

Bag-of-Words (BoW) is a basic and widely used text representation technique in NLP. It treats each document (or piece of text) as a "bag" of individual words, disregarding the order and structure of the words. The basic idea is to create a vocabulary of all unique words in the entire corpus (collection of documents). For each document, a vector is created where each dimension corresponds to a word from the vocabulary, and the value in each dimension represents the frequency of that word in the document. BoW is simple and efficient but does not capture word order or context.

Character n-grams are a more fine-grained technique that represents text by breaking it down into chunks of characters, rather than words. An n-gram is a contiguous sequence of n characters in a string. The character n-grams technique was trialled across varying n-gram intervals, specifically (1,2), (1,3), and (1,4), i.e., we are considering all possible combinations of character sequences with lengths ranging from 1 to 2 characters, 1 to 3 characters, and 1 to 4 characters. Character n-grams capture subword information.

## 5. Classifiers

A comprehensive array of models was applied in this study to address the task at hand. The Scikit-Learn library was employed for model implementation, and default parameters were utilized. The utilization of this diverse set of models aimed at exploring a wide spectrum of possibilities and capturing nuanced patterns within the code-mixed data. The descriptions of the models used is as follows:

RandomForest is an ensemble learning method that builds a forest of decision trees and combines their predictions to improve accuracy and reduce overfitting in classification and regression tasks. Multinomial Naive Bayes is a classification algorithm commonly used for text and document classification tasks. It's based on the Bayes' theorem and assumes that features are conditionally independent. Logistic Regression is a simple linear classification algorithm used for binary classification problems. It models the probability of a binary outcome. Linear Support Vector Classifier is a linear machine learning model used for binary classification. It aims to find a hyperplane that best separates the data into two classes. A Decision Tree is a tree-like model that makes decisions by recursively splitting the dataset based on the most significant feature at each node. KNN is a non-parametric and instance-based algorithm used for classification and regression. It classifies data points based on the majority class of their k-nearest neighbors.

AdaBoost is an ensemble learning technique that combines multiple weak learners (usually decision trees) to create a strong classifier. OneVsRest classifier was used, a multi-class classification strategy where a separate binary Logistic Regression classifier is trained for each class to handle multi-class classification problems. Gradient Boosting is an ensemble method that builds an additive model by training weak learners sequentially, where each new learner corrects the errors made by the previous one.

Stacking classifier was used, which combines multiple base models (LinearSVC, RandomForest, KNN) with a meta-learner (Logistic Regression) to improve overall model performance. The Voting Classifier combines the predictions of multiple classifiers (e.g., LR, RF, and SVC) using majority voting or weighted voting to make a final decision. Bagging (Bootstrap Aggregating) is an ensemble technique that trains multiple instances of the same base model (KNN) on bootstrapped samples of the data and combines their predictions.

## 6. Methodology

After performing data preprocessing, each of the models mentioned in the previous section was trained separately using the various word embeddings discussed. Table 2 (refer to Table 2) displays the weighted average F1 scores of the classifiers when combined with different combinations of word embeddings, including Bag of Words (BoW) and Character n-grams with varying n-gram ranges. The evaluation of the models was based on their performance in terms of the weighted average F1 score. This particular metric is well-suited for assessing multi-class classification models because it accounts for class imbalances, provides a comprehensive measure of overall performance across all classes, and considers practical considerations such as the significance of individual classes. This metric delivers a balanced evaluation that combines both precision and recall, making it a valuable tool for selecting models and evaluating their performance in real-world applications.

## 7. Results

Out of all the models, CountVectorizer with an n-gram range of (1,4), coupled with LinearSVC classifier was the most effective configuration observed for the language identification task. This combination adeptly captures linguistic nuances and establishes clear decision boundaries,

**Table 2**
Weighted average F1 score of models on the Development set

| Model | BoW | (1, 2) n-grams | (1, 3) n-grams | (1, 4) n-grams |
|---|---|---|---|---|
| Multinomial NB | 0.60 | 0.66 | 0.74 | 0.76 |
| Random Forest | 0.73 | 0.85 | 0.86 | 0.86 |
| Logistic Regression | 0.61 | 0.78 | 0.84 | 0.85 |
| Linear SVC | 0.73 | 0.77 | 0.84 | **0.87** |
| Decision Tree | 0.73 | 0.83 | 0.82 | 0.82 |
| KNN | 0.63 | 0.81 | 0.81 | 0.80 |
| AdaBoost | 0.40 | 0.46 | 0.51 | 0.51 |
| One Vs Rest | 0.59 | 0.77 | 0.84 | 0.85 |
| Gradient Boost | 0.53 | 0.75 | 0.76 | 0.75 |
| Stacking | 0.73 | 0.86 | 0.83 | 0.86 |
| Voting | 0.72 | 0.85 | 0.85 | 0.85 |
| Bagging | 0.63 | 0.82 | 0.81 | 0.81 |

showcasing superior accuracy and precision in distinguishing languages. With the development dataset, the model gives a weighted average F1 score of 0.87. The weighted average F1 score with this set-up for the test dataset was 0.799. This was the third best score in the CoLi-Tunglish shared task.

## 8. Conclusion

This study addressed the task of language identification within code-mixed Tulu-English words, prevalent in multilingual communication. Through the utilization of diverse word embeddings and classifiers, significant progress was made in effectively meeting this challenge. Notably, character n-grams in the range 1 to 4 with LinearSVC classifier demonstrated exceptional performance, yielding the highest weighted average F1 score compared to the other embeddings-model that were evaluated, highlighting the critical role of appropriate selection in achieving accurate language identification. Further exploration could involve refining embeddings and considering ensemble strategies to advance the accuracy and resilience of code-mixed language identification systems.

## References

[1] N. H. Hebbar, Tulu Language - Its Script and Dialects, https://www.mangaloretoday.com/opinion/Tulu-Language-Its-Script-and-Dialects.html, -. [Accessed 07-10-2023].

[2] 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.

[3] 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.

[4] F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, H. Shashirekha, G. Sidorov, A. Gelbukh, Overview of coli-kanglish: Word level language identification in code-mixedkannada-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.

[5] 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, Association for Computational Linguistics, Hong Kong, 2018. URL: https://aclanthology.org/Y18-1021.

[6] C. Sabty, I. Mesabah, Özlem Çetinoğlu, S. Abdennadher, Language identification of intra-word code-switching for arabic–english, Array 12 (2021) 100104. URL: https://www.sciencedirect.com/science/article/pii/S2590005621000473. doi:https://doi.org/10.1016/j.array.2021.100104.

[7] S. Mandal, S. D. Das, D. Das, Language Identification of Bengali-English Code-Mixed data using Character & Phonetic based LSTM Models, 2018. URL: http://arxiv.org/abs/1803.03859. doi:10.48550/arXiv.1803.03859, arXiv:1803.03859 [cs] version: 1.

[8] O. E. Ojo, A. Gelbukh, H. Calvo, A. Feldman, O. O. Adebanji, J. Armenta-Segura, Language Identification at the Word Level in Code-Mixed Texts Using Character Sequence and Word Embedding, 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. 1–6. URL: https://aclanthology.org/2022.icon-wlli.1.

[9] A. L. Tonja, M. G. Yigezu, O. Kolesnikova, M. S. Tash, G. Sidorov, A. Gelbuk, Transformer-based Model for Word Level Language Identification in Code-mixed Kannada-English Texts, 2022. URL: http://arxiv.org/abs/2211.14459. doi:10.48550/arXiv.2211.14459, arXiv:2211.14459 [cs].

[10] E. Uchoi, M. Kaur, Language Identification of English and Punjabi, Eur. Chem. Bull. (2023) 4119–4123. doi:10.48047/ecb/2023.12.si6.367.

[11] G. Chittaranjan, Y. Vyas, K. Bali, M. Choudhury, Word-level Language Identification using CRF: Code-switching Shared Task Report of MSR India System, in: Proceedings of the First Workshop on Computational Approaches to Code Switching, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 73–79. URL: https://aclanthology.org/W14-3908. doi:10.3115/v1/W14-3908.

[12] S. Gella, K. Bali, M. Choudhury, "ye word kis lang ka hai bhai?" Testing the Limits of Word level Language Identification, in: Proceedings of the 11th International Conference on Natural Language Processing, NLP Association of India, Goa, India, 2014, pp. 368–377. URL: https://aclanthology.org/W14-5151.

[13] N. Sarma, S. R. Singh, D. Goswami, Word level language identification in assamese-bengali-hindi-english code-mixed social media text, in: 2018 International Conference on Asian Language Processing (IALP), 2018, pp. 261–266. doi:10.1109/IALP.2018.8629104.

[14] A. F. Hidayatullah, R. A. Apong, D. T. C. Lai, A. Qazi, Corpus creation and language identification for code-mixed Indonesian-Javanese-English Tweets, PeerJ Com-

puter Science 9 (2023). URL: https://www.readcube.com/articles/10.7717%2Fpeerj-cs.1312. doi:10.7717/peerj-cs.1312.

[15] 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, 2022. URL: http://arxiv.org/abs/2211.09847. doi:10.48550/arXiv.2211.09847, arXiv:2211.09847 [cs].

[16] 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.