Word-level Language Identification using Character-level Features

Amaan Ahmad[1], Asha Hegde[2] and Sharal Coelho[2]

[1]Department of Computer Science, Manipal Institute of Technology (MIT), Bangaluru, India
[2]Department of Computer Science, Mangalore University, India

Abstract

The social media contexts poses significant challenges for natural language processing (NLP) and machine learning tasks. Existing language detection tools struggle to identify languages at the word level. Word-level Language Identification (LID) is a critical task in NLP, particularly for handling code-mixed and multilingual text prevalent in social media and digital communication. In this work, we address the challenge of identifying languages at the word level across five diverse languages: Kannada, Malayalam, Telugu, Tamil and Tulu. We employ a feature extraction approach by using Term Frequency-Inverse Document Frequency (TF-IDF) of character n-grams ranging from 1 to 4, which are then fed into classical machine learning model (ExtraTrees Classifier). Our method achieves superior performance, securing better ranks in benchmark evaluations for all five languages. Experimental results demonstrate high accuracy for ExtraTrees.

Keywords

Word Level Language Identification, Machine Learning, ExtraTrees, Dravidian Languages

1. Introduction

Language identification (LID) is a foundational task in natural language processing (NLP) that involves determining the language of a given text segment [1]. While document-level or sentence-level LID has been well-studied, word-level LID presents unique challenges, especially in multilingual and code-mixed scenarios where multiple languages co-exist within a single utterance [2]. Word-level LID aims to tag each word or token with its corresponding language, enabling downstream applications such as machine translation, sentiment analysis, and hate speech detection in diverse linguistic contexts. The importance of word-level LID has grown with the proliferation of social media platforms, where users frequently mix languages such as Hindi-English, Telugu-English [3] Kannada-English, Malayalam-English, and so on[4]. Early approaches to LID relied on rule-based methods or simple statistical models, but recent advancements have incorporated machine learning and deep learning techniques. For instance, character-level features have proven effective due to their ability to capture orthographic patterns unique to languages. Dravidian languages are a well-known language category spoken by more than 250 million people, mainly in South India, Sri Lanka, and other parts of South Asia. Kannada, Telugu, Tamil, Malayalam and Tulu, are the most widely spoken Dravidian languages [5][6][7]. These Dravidian languages are low-resource languages due to the limited availability of digital tools and resources.

To address the aforementioned limitations, we propose a lightweight pipeline for word-level LID that leverages classical machine learning models with robust feature engineering. Our approach uses TF-IDF combined with character n-grams (1-4 grams) to extract discriminative features from words. The features are then input to SVM and DT classifiers, chosen for their efficiency, interpretability, and strong performance on textual data.

The rest of the paper is organized as follows: Section 2 contains Related Work. While Section 3 describes the Methodology, Section 4 gives a description of the Experiments, Results, and Observations followed by Conclusion in Section 5.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
amaanahd1@gmail.com (A. Ahmad); hegdekasha@gmail.com (A. Hegde); sharalmucs@gmail.com (S. Coelho)

2. Related Work

Recently, researchers in the field of code-mixed text have shown growing interest in under-resourced and low-resource languages for various applications, such as sentiment analysis[8], machine translation, and so on [9] [10] [7]. The task of word-level LI has become increasingly important as multilingual and code-mixed content continues to grow, especially on digital platforms like Facebook, YouTube, etc. Researchers have explored various approaches to address word-level LI for languages where extensive corpora and linguistic resources are readily available [5]. However, the challenge of processing languages with limited resources, often referred to as low-resource languages for word-level LI has gained significant attention. Some of the related works for word-level LI are described below:

Thara and Poornachandran [11] have scraped YouTube comments to identify bilingual Malayalam-English code-mixed text. To filter out the comments they have removed English alphabets, numbers, special characters, and emoticons. They used transformer models (CamemBERT, XLMRoBERTa, ELECTRA, and DistilBERT) to predict language tags at the word-level. The results of this study showed that ELECTRA performed better than other models by obtaining F1-score of 0.993. Deka et al. [12] proposed the Bidirectional Encoder Representations for Transformers (BERT) based approach for LI using Kannada-English code-mixed corpus. Their approach achieved 86% weighted average F1-score and a macro average F1-score of 57%. To identify the language of words in code-mixed Kannada texts Yigezu et al. [13] proposed a Bi-LSTM with an attention model that integrates BERT features to enhance word-level LI accuracy. While the existing work on word-level LI in low-resource has made significant advances, there are still some limitations that create opportunities for further research. For instance, code-mixed data often depends on the availability of high-resource languages, which are not always accessible. Processing user-generated text is another challenge due to its variability, including code-mixed nature, spelling or grammatical errors, etc. Additionally, it often lacks context, making it harder to interpret meaning and intent accurately. Additionally, the effectiveness of incorporating linguistic features can vary greatly depending on the specific languages and features used, and finding the optimal combination needs to be explored.

3. Task Description

Language Identification (LI) is the process of determining the language present in a given text and serves as a foundational step for many applications, including sentiment analysis, machine translation, information retrieval, and natural language understanding. In multilingual India, particularly among young people, social media posts often contain code-mixed text that blends local languages with English across different levels. This creates substantial challenges for LI, especially when code-mixing occurs within a single word.

Dravidian languages, spoken widely in southern India, are under-resourced despite their rich morphological complexity. They face additional technological hurdles, especially regarding script representation in digital spaces, which often leads users to adopt Roman or hybrid scripts for online communication. While this widespread code-mixing offers a wealth of linguistic data for research, it remains relatively unexplored. To tackle the challenge of word-level LI for Dravidian languages, we are organizing a shared task that provides datasets for five languages (Kannada, Tamil, Malayalam, Telugu, and Tulu) encouraging the development of advanced LI models.

4. Methodology

In the proposed methodology, the word-level LI task is modeled as a sequence labeling problem where the goal is to assign a label to each word in a sequence. It is achieved by by training ML models. The framework of proposed model is shown in Figure 1 and the steps involved in the framework are described in the following subsection.

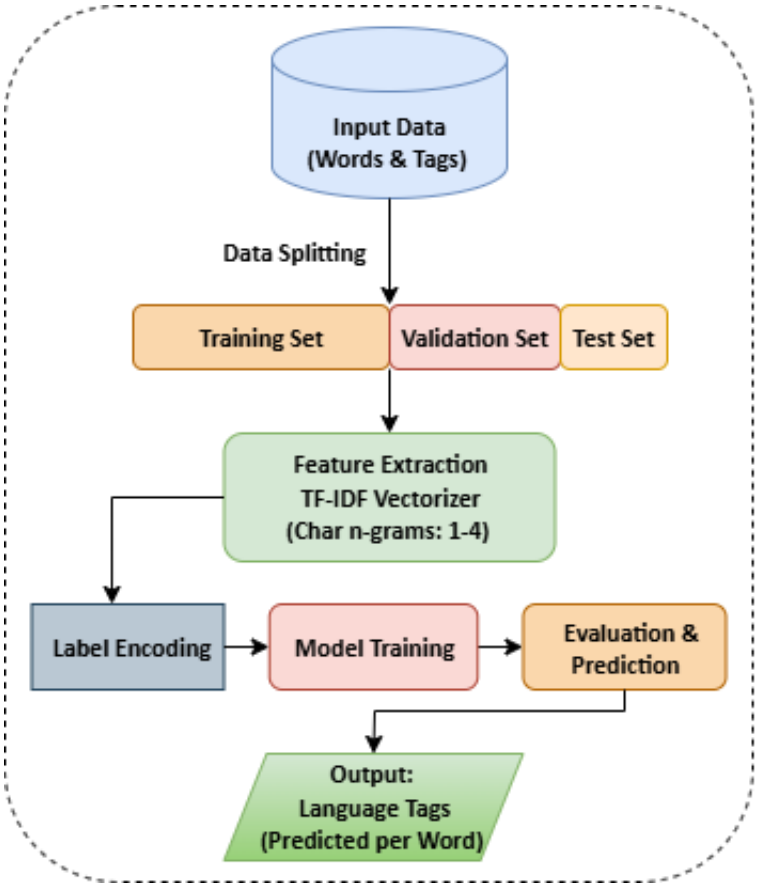

Figure 1: Framework of Proposed model

involves supervised machine learning using character-level features and multiple classifiers for comparison.

4.1. Feature Extraction Using TF-IDF Vectorization

We employed a TF-IDF vectorizer configured for character-level analysis with an n-gram range of (1, 4) to effectively capture subword patterns, such as character sequences distinctive to specific languages. The vectorizer was fitted on the training data to acquire the vocabulary and convert it into a sparse matrix representation. Subsequently, the test data was transformed using the fitted vectorizer to maintain consistency and prevent data leakage. For generating final predictions on unseen data, the 'Word' column is transformed utilizing the same fitted vectorizer.

4.2. Label Encoding

The LabelEncoder is used to convert categorical language tags into numerical labels. The target_names considered as the list of unique language classes, and all_labels as their corresponding indices for multi-class handling.

5. Experiments and Results

5.1. Dataset

The datasets comprising code-mixed and monolingual texts for the five languages (Kannada, Malayalam, Telugu, Tamile and Tulu) is utilized. Each dataset includes word-level annotations. Table 1 contains sample words, their English translations, and the corresponding labels/tags from the given dataset.

Table 1
Sample words and corresponding labels

Language	Word	English Translation	Word	English Translation
Malayalam	oru	One	arakum	No One
Malayalam-English	backil	At the back	traileril	In trailer
Tamil	yennada	What	ippo	Now
Tamil-English	trailerum	Trailer	collegela	At college
Kannada	idu	This	aytu	Okay
Kannada-English	Moviedu	Movie's	conceptu	Concept
Location	Mangalore	-	Karnataka	-
Name	Appu	-	Aravind	-
Symbol	. * ?	-	-	-

5.2. Results

The performance of the classifier is evaluated based on Macro F1-Score (M_F1). Macro scores are preferred for evaluating the performance across all classes without bias. The performances of the proposed CRF models on Test sets are shown in Table 2.

We evaluate our method on datasets for five languages: Kannada, Malayalam, Telugu, Tamile and Tulu, achieving better ranks compared to baselines in benchmark tasks. This work emphasizes simplicity and effectiveness, making it suitable for resource-limited environments while outperforming more complex models in speed and accuracy on the given datasets.

Table 2
Performance and ranking in each Language

Language	W.Precision	W.Recall	W.F1	M.Precision	M.Recall	M.F1	Acc	Rank
Kannada	0.9450	0.9441	0.9418	0.9517	0.8599	0.8987	0.9441	8
Malayalam	0.8714	0.8738	0.8654	0.8765	0.7553	0.7938	0.8738	4
Tamil	0.8965	0.8984	0.8936	0.8780	0.6634	0.7084	0.8984	3
Telugu	0.9234	0.9291	0.9214	0.7946	0.7428	0.7572	0.9291	3
Tulu	0.8930	0.8937	0.8889	0.8742	0.7481	0.7925	0.8937	6

6. Conclusion

In this paper, we describe that TF-IDF with character n-grams and classical ML model can achieve better resulta in word-level LID for multiple languages, addressing key limitations of prior methods. The "CoLI-Dravidian@2025: Word-level Code-Mixed Language Identification in Dravidian Languages" shared task at FIRE 2025. By training the ML models with character n-grams features, the proposed model obtained Macro F1 score of 0.7084 for Tamil and 0.7572 for Telugu resepectively. For both languages we secured 3rd rank. Future work could integrate this with deep learning approaches.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat GPT-4 in order to:Grammar and spelling check. Using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's control.

References

[1] D. Nguyen, A. S. Doğruöz, Word level language identification in online multilingual communication, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 857–862.

[2] K. Shanmugalingam, S. Sumathipala, C. Premachandra, Word level language identification of code mixing text in social media using nlp, in: 2018 3rd international conference on information technology research (ICITR), IEEE, 2018, pp. 1–5.

[3] 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.

[4] H. Jhamtani, S. K. Bhogi, V. Raychoudhury, Word-level language identification in bi-lingual code-switched texts, in: Proceedings of the 28th Pacific Asia Conference on language, information and computing, 2014, pp. 348–357.

[5] 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.

[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, 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.

[8] S. Coelho, A. Hegde, P. Lamani, H. L. Shashirekha, et al., Mucsd@ dravidianlangtech2023: Predicting sentiment in social media text using machine learning techniques, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 282–287.

[9] 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.

[10] 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.

[11] S. Thara, P. Poornachandran, Transformer Based Language Identification for Malayalam-English Code-Mixed Text, IEEE Access 9 (2021) 118837–118850.

[12] 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.

[13] 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.