Word Level Language Identification in Code-mixed Dravidian Languages

Harshitha S Kumar, Sharal Coelho, Asha Hegde, Kavya G and Hosahalli Lakshmaiah Shashirekha

Department of Computer Science, Mangalore University, India

Abstract

In social media posts, it is common to see a mix of languages at word level and identifying the language of the word in these posts is essential for applications like Machine Translation. Word level Language Identification (LI) deals with identification of the language of the word in a sentence or document. The limited availability of labeled data in low-resource Indian languages such as Kannada, Tulu, Tamil, Telugu, etc., for word-level LI hinders the advancements in developing robust models for LI in such languages. In order to meet this requirement "CoLI-Dravidian 2024: Word-level Code-Mixed Language Identification in Dravidian Languages" shared task organized at FIRE-2024 invites researchers to develop models to address the challenges of LI in four different Dravidian languages (Kannada, Tulu, Tamil, and Malayalam). In this paper, we - team MUCSNLPLab, describe the Conditional Random Field (CRF) model trained with text-based features to identify the language of the word in a given sentence. The proposed CRF model obtained Macro F1 scores of 0.608, 0.869, 0.767, and 0.772 for Tamil, Kannada, Malayalam, and Tulu respectively.

Keywords

Natural Language Processing, Dravidian Languages, CRF model, Language Identification, Code-mixed data

1. Introduction

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. While Kannada is one of the scheduled languages of India with 40+ million speakers mainly from Karnataka state. Tulu has a rich cultural and literary heritage and is spoken by a community of about 2.5 million native speakers in coastal regions of southern part of India, predominantly in Karnataka state [1]. Tamil is the official language of Tamil Nadu and Puducherry and it is the oldest language in India. Malayalam language is another most widely spoken language in the southern region of India with nearly 35+ million speakers [2]. Despite their rich linguistic history, these languages are considered low-resource languages, as they lack extensive digital tools, resources, and formal computational language processing infrastructure.

Recently social media has allowed speakers of low-resource languages to create and share content in their native and/or regional language mixed with English on platforms like Twitter, WhatsApp, Facebook, etc., [3]. This combination of multiple languages in a paragraph, sentence, or word is termed code-mixing or code-switching and such data is called code-mixed data that has become a default language of social media which has attracted the attention of researchers in the Natural Language Processing (NLP) domain. Social media platforms have given users the freedom to write text informally, often ignoring the grammar conventions of the specific language used. This has resulted in considerable growth in user-generated texts, which are characterized by code-mixed words like "moviege" (Kannada-English), "trailerg" (Tulu-English), user-defined abbreviations ('ASAP' for "As Soon As Possible"), and the repetition of characters ("toooo goood" for "too good"), and so on. Such contents are difficult to understand due to the informal linguistic arrangement.

Forum for Information Retrieval Evaluation (FIRE), December 12-15, India

hskbangera@gmail.com (H. S. Kumar); sharalmucs@gmail.com (S. Coelho); hegdekasha@gmail.com (A. Hegde); kavyamujk@gmail.com (K. G); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4054/T3-8.pdf

Such contents are difficult to understand due to the informal linguistic arrangement. In order to handle code-mixing and informal content on platforms like social media, chatbots, and real-time systems, the concept of word-level LI is developed. Each word in the text can be classified as belonging to a single language or a group of languages is known as word-level Language Identification (LI). This encouraged NLP researchers to explore the problem of LI at the word-level. Word-level LI is necessary for applications such as Machine Translation (MT), Part-Of-Speech (POS), and Named Entity Recognition (NER) in multilingual contexts. With massive amounts of social media data generated everyday, manual word-level LI is difficult. Additionally, it often lacks context, making it harder to interpret meaning and intent accurately. 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. Hence, this requires automated tools and techniques to identify language of the word [4]. The word-level LI is modeled as a sequence labeling task which involves assigning a label to each element in an input sequence, maintaining the order of the words. This approach allows for better handling of the complexities inherent in code-mixed content, where multiple languages may be present within the same sentence.

The increasing prevalence of code-mixed content, especially on social media, has made word-level LI a critical challenge. Despite its importance, the development of reliable tools for word-level LI in Dravidian languages continues to face significant challenges, mainly due to the lack of standardized resources and the scarcity of large-scale annotated datasets. While Deep Learning (DL) models, such as neural networks, have demonstrated strong performance in various NLP tasks, they often require vast amounts of labeled data to function effectively. In contrast, traditional Machine Learning (ML) techniques like Conditional Random Fields (CRF), when integrated with well-designed linguistic features, provide a more resource-efficient alternative. These methods offer a promising balance between accuracy and computational feasibility, making them particularly suitable for low-resource languages like the Dravidian languages. To address the challenges of word-level LI in Dravidian languages - Tamil, Kannada, Malayalam, and Tulu, in this paper, we - team MUCSNLPLab, describe the learning models submitted to "CoLI-Dravidian 2024: Word-level Code-Mixed Language Identification in Dravidian Languages" shared task organized at FIRE 2024 [5]. Word-level LI is modeled as a multi-class classification using CRF model trained with text-based features, to identify the language of each word in the given text.

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.

2. Related Work

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 [6]. 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:

Chaitanya et al. [7] proposed word-level LI model in Hindi-English code-mixed data using word embeddings (Continuous Bag of Words (CBOW) and Skip-gram) to effectively capture word semantics and relationships. They achieved 67.33% and 67.34% for CBOW and Skip-gram model respectively using Support Vector Machine (SVM) classifiers. [8] implemented a word-level LI system for code-mixed Malayalam-English and Tamil-English Facebook data and generated character embedding features using skip-gram architecture. They employed 10-fold cross-validation to train and evaluate the SVM model, ensuring the robust performance of the model, and obtained 93% and 95% accuracies for Malayalam-English and Tamil-English text, respectively. Thara and Poornachandran [9] 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

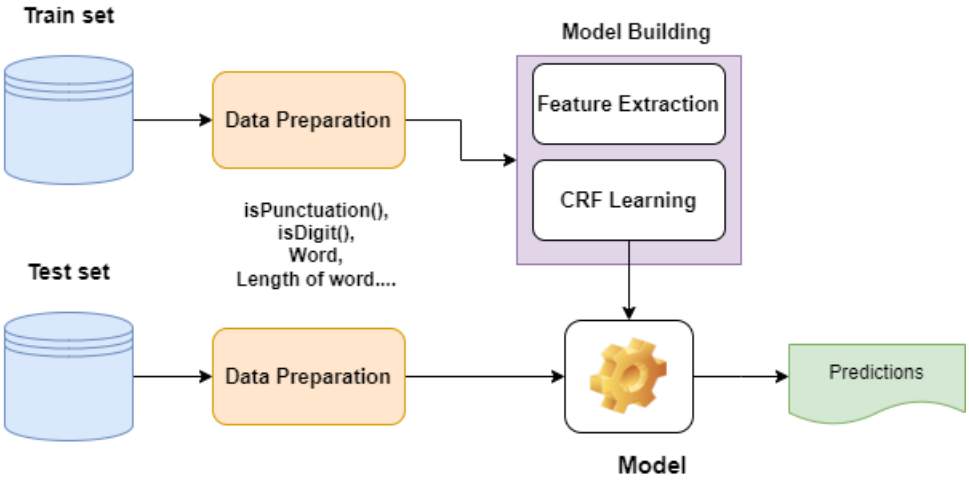

Figure 1: Framework of CRF model

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. [10] 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. [11] proposed a Bi-LSTM with an attention model that integrates BERT features to enhance word-level LI accuracy.

Mandal and Singh [12] proposed bidirectional LSTM network with CRF layer (Bi-LSTM-CRF) for word-level LI in Bengali-English and Hindi-English and obtained accuracies of 93.28% and 93.32% for Bengali-English and Hindi-English, respectively. Gundapu et al. [13] proposed Naive Bayes, Random Forest, CRF and Hidden Markov Model for word-level LI in English-Telugu code-mixed data. Among these models CRF-based model obtained F1-score of 0.91.

3. 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 training CRF model with text based features like next words, previous words, etc. The framework of CRF model is shown in Figure 1 and the steps involved in the framework are described in the following subsection.

3.1. Data Preparation and Feature Extraction

The data preparation phase necessitates a careful examination of the provided data and arranging the data to fit the selected feature extraction technique to extract text features that can be used to train the CRF model.

In this work, the CRF model utilizes text features such as word length, previous words, next words, digits, and punctuations to capture the dependencies and relationships between the words in the given sequence. These text features show an essential role in determining the performance of a CRF model. The text features for words in the sample sentences are shown in Table 1 and few of these features are described below:

- **A word**: the current word.
- **Word length**: refers to the number of characters in a word.
- **Local context**: two preceding and two succeeding words.

- **Is current word digit**: checks whether the word contains only numerical characters. This text feature helps in identifying tokens belonging to 'Number' or 'Other' classes.
- **Is current word punctuation**: checks whether the word contains only punctuations like comma (,), full stop (.), asterisk (*), etc. This helps in identifying tokens belonging to the 'sym' category.

These text features are used to assist the learning models in making more accurate predictions.

Table 1
Text features used in proposed CRF model

Example 1- (from Tulu Dataset): Yedde concept bhojanna Ethe wunduwe **jaasti** madime aandala Facebook love Joru		
Example 2- (from Kannada Dataset): nam athra short moviedu **conceptuu** ede produce mathira		
Text Features	**Example 1**	**Example 2**
A word	jaasti	conceptuu
Length of word	6	**9**
Is the word at the beginning of the sentence	False	False
Is the word at the end of the sentence	False	False
Is current word digit	False	False
Is current word is written in lower case	True	True
Is current word punctuation	False	produce
Previous word-2	wunduwe	moviedu
Previous word-3	Ethe	short
Next word+2	madime	ede
Next word+3	aandala	produce

3.2. Model Construction

CRF model is a probabilistic model often used for sequence prediction tasks and performs well in modeling the conditional probability distribution of tags such as NER or POS tagging. It grasps the dependencies between tags by taking into account both previous and subsequent observations and learns from the text features to predict the sequence of labels for the input sequence. Due to its ability in capturing relationships between words, CRF models have drawn much attention in labeling sequence data.

In this work, CRFSuite is employed for implementing CRF, which is a wrapper built using the sklearn_crfsuite library, providing a scikit-learn compatible estimator for CRF. CRFSuite is a wrapper implemented using sklearn_crfsuite[1] library, which is a scikit-learn compatible estimator for CRF implementation. This library simplifies the classifier construction process by wrapping the transformation of textual features into feature vectors and training the CRF classifier.

4. Experiments

The experiments are carried out on the datasets provided by the shared task for word-level LI in four code-mixed Dravidian languages–Tulu-English [1], Kannada-English [14][15], Tamil-English, and Malayalam-English [16]. The data provided in all the four languages are in Romanized script. Table 2 displays the train and validation set's tag-wise distribution for the given four code-mixed datasets. Few of the tags in the annotated data are given below:

- en/mal/tam: English / Malayalam / Tamil
- mixed: Code-mixed words
- sym/SYM: Symbol
- tmen: Tamil-English

Table 2
Label distribution of train and validation set in given four datasets

Tulu		Malayalam		Tamil		Kannada	
Tulu	11649/1251	**MALAYALAM**	11233/1175	**tm**	7064/1000	**en**	17668/1109
English	7480/742	**ENGLISH**	5492/538	**en**	2763/496	**sym**	3730/334
sym	4243/422	**SYM**	2777/294	**tmen**	1255/144	**kn**	3623/637
Kannada	2950/273	**OTHER**	2108/179	**sym**	1211/183	**other**	2573/53
Name	1501/135	**NAME**	1941/169	**name**	1149/160	**name**	1223/158
Other	638/85	**MIXED**	762/63	**Other**	76/1	**mixed**	1077/180
Mixed	543/57	**NUMBER**	568/77	**Location**	11/0	**location**	121/13
Location	519/41	**PLACE**	114/9	-		-	

Table 3
Sample words and corresponding labels

Language	Word	English Translation	Word	English Translation
Kannada	idu	This	aytu	Okay
Kannada-English	camerada	Camera's	conceptu	Concept
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
Tulu	deppunu	To take	barevodu	Should Write
Tulu-English	comedyg	For Comedy	startapuni	Started
Location	Mysore	-	Kerala	-
Name	Appu	-	Ashwini	-
Symbol	. * ?	-	-	-

Table 4
The hyperparameters and their values used for CRF models

Parameter	Run 1	Run 2	Run 3
Optimizer	lbfgs	l2sgd	lbfgs
c1	0.065	-	0
c2	0.002	0.002	0
max iteration	200	200	200
lbfgs - Limited memory broyden fletcher goldfarb shanno **l2sgd - Stochastic gradient descent**			

Table 3 contains sample words, their English translations, and the corresponding labels/tags from the given dataset. Using four Dravidian language datasets three different experiments are carried out by fine-tuning the hyperparameters and their values to train the CRF models for the given word-level LI task are shown in Table 4.

4.1. Results and Observations

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 5 and the results in the table reveals that M_F1 are considerably low for Tamil language for Run1 compared to other languages because limited Tamil-English data restricts the model's ability to learn patterns and nuances in the language resulting in lower performance. For Kannada language, Run2 experiment showed poor performance whereas Run1 achieved better outcomes. For all Run2 experiments, the models showed poor performance, because L2-SGD is specially designed to work well with large datasets since the provided datasets are limiting

[1]https://pypi.org/project/sklearn-crfsuite/

the diversity of the training samples and making it difficult for the model to learn meaningful patterns effectively.

Table 5
Performance of the proposed CRF models

Language	Submissions	M_F1	M_Pr	M_Re	W_F1	W_Pr	W_Re	Accuracy
Tamil	Run1	0.232	0.218	0.268	0.434	0.362	0.576	0.576
	Run2	0.232	0.218	0.268	0.434	0.362	0.576	0.576
	Run3	**0.608**	**0.627**	**0.596**	**0.859**	**0.856**	**0.864**	**0.864**
Kannada	**Run1**	**0.869**	**0.901**	**0.847**	**0.929**	**0.931**	**0.929**	**0.929**
	Run2	0.247	0.225	0.286	0.488	0.409	0.624	0.624
	Run3	0.827	0.863	0.800	0.902	0.902	0.904	0.904
Malayalam	**Run1**	**0.767**	**0.897**	**0.706**	**0.869**	**0.879**	**0.879**	**0.879**
	Run2	0.767	0.897	0.706	0.869	0.879	0.879	0.879
	Run3	0.764	0.868	0.718	0.877	0.879	0.884	0.884
Tulu	**Run1**	**0.772**	**0.867**	**0.719**	**0.874**	**0.880**	**0.879**	**0.879**
	Run2	0.755	0.839	0.707	0.860	0.863	0.865	0.865
	Run3	0.755	0.839	0.707	0.860	0.863	0.865	0.865
Macro precision (M_Pr), Macro Recall (M_Re), Weighted F1-Score (W_F1), Weighted precision (W_Pr), Weighted Recall (W_Re)								

Considering the performance across all evaluation metrics in both macro and weighted forms, CRF models demonstrate reliable performance across languages. The results indicate that Run1 model has exhibited a better M_F1 for Kannada, Malayalam, and Tulu and the third experiment (Run3) gives better results for Tamil Language. The Run1 model acquired Macro F1 scores of 0.869, 0.767, and 0.772 for Kannada, Malayalam, and Tulu Languages by securing 7^{th}, 9^{th}, and 7^{th} ranks respectively. This is demonstrated by the fact that the model performs more or less consistently with accuracies ranging from 60% - 86% across four code-mixed languages.

Table 6
Few misclassified samples from the validation set obtained by proposed model

Word	Actual Label	Predicted Label	Reason
Nataka	Tulu, Kannada	Tulu, Kannada, Mixed,	The word "Nataka" is classified as 'tulu', and 'Kannada', because it appears in different tags, leading to ambiguity. Inconsistent annotations in the training data further confuse the model.
nana	English, Kannada, Tulu	Kannada, Tulu	The annotation includes multiple labels—English, Kannada, and Tulu—indicating mixed language markers. The predictions also vary, with labels like Kannada and Tulu, showing a similar mix
scene	English	tmen	In the training set, the word 'scene' appears in various code-mixed forms like 'scenela' and 'sceneda,' which are labeled as 'tmen.' As a result, the model predicted 'scene' as 'tmen.
A10	Other	Number	The misclassification of 'A10' as a number may be due to imbalanced training data, with the 'Other' label underrepresented, and the word containing a number.

The performance variation could be due to i) Unique grammatical structures, syntax, and morphology, which may not be captured well by the model, ii) Imbalanced or limited training data, leading to poor classification, and (iii) Wrong annotation issues. Additionally, other than Romanized words, Arabic words found in Malayalam datasets posed significant challenge. In the given dataset, the words like 'good', 'message', 'super' are written as "goood", "msg", and 'superrruu', etc., such nonstandard usage are acceptable by the social media users but is an issue for word-level LI model. The Table 6 shows few misclassification samples, highlighting instances where imbalanced data, mixed-language annotations,

and numeric content in words contributed to prediction errors.

5. Conclusion

In this paper, we describe a CRF model submitted to "CoLI-Dravidian 2024: Word-level Code-Mixed Language Identification in Dravidian Languages" shared task at FIRE 2024, for word-level LI in four code-mixed Dravidian languages (Kannada, Tulu, Tamil, and Malayalam). By training the CRF model with text-based features, the proposed model obtained Macro F1 score of 0.608 for Tamil and secured 3rd rank. Efficient techniques will be explored in the future to handle the imbalanced dataset and improve the performance of the proposed models.

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] 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.

[2] S. Coelho, A. Hegde, G. Kavya, H. L. Shashirekha, Mucs@ dravidianlangtech2023: Malayalam Fake News Detection using Machine Learning Approach, in: Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, 2023, pp. 288–292.

[3] S. Thara, P. Poornachandran, Code-Mixing: A Brief Survey, in: 2018 International conference on advances in computing, communications and informatics (ICACCI), IEEE, 2018, pp. 2382–2388.

[4] S. Coelho, A. Hegde, P. Lamani, G. Kavya, H. L. Shashirekha, 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.

[5] A. Hegde, F. Balouchzahi, S. Butt, S. Coelho, K. G, H. S Kumar, S. D, S. Hosahalli Lakshmaiah, A. Agrawal, Overview of CoLI-Dravidian: Word-Level Code-Mixed Language Identification in Dravidian Languages, in: Forum for Information Retrieval Evaluation FIRE - 2024, 2024.

[6] 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.

[7] 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.

[8] P. Veena, M. A. Kumar, K. Soman, An Effective Way of Word-Level Language Identification for Code-Mixed Facebook Comments using Word-Embedding via Character-Embedding, in: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE, 2017, pp. 1552–1556.

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

[10] 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.

[11] 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.
[12] S. Mandal, A. K. Singh, Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture, arXiv preprint arXiv:1808.07118 (2018).
[13] S. Gundapu, R. Mamidi, Word Level Language Identification in English Telugu Code Mixed Data, arXiv preprint arXiv:2010.04482 (2020).
[14] S. Hosahalli Lakshmaiah, F. Balouchzahi, A. Mudoor Devadas, G. Sidorov, CoLI-Machine Learning Approaches for Code-Mixed Language Identification at the Word Level in Kannada-English Texts, acta polytechnica hungarica (2022).
[15] F. Balouchzahi, S. Butt, A. Hegde, N. Ashraf, S. Hosahalli Lakshmaiah, G. Sidorov, A. Gelbukh, Overview of CoLI-Kanglish: Word Level Language Identification in Code-Mixed Kannada-English Texts at ICON 2022, in: 19th International Conference on Natural Language Processing Proceedings, 2022.
[16] 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, FIRE '23, Association for Computing Machinery, New York, NY, USA, 2024, p. 25–26. URL: https://doi.org/10.1145/3632754.3633075. doi:10.1145/3632754.3633075.