CoLI-Kanglish: Word-Level Language Identification in Code-Mixed Kannada-English Texts Shared Task using the Distilka model

Vajratiya Vajrobol
tiya101@south.du.ac.in
Institute of Informatics and Communication
University of Delhi

Abstract

Due to the intercultural demographic of online users, code-mixed language is often used by them to express themselves on social media. Language support to such users is based on the ability of a system to identify the constituent languages of the code-mixed language. Therefore, the process of language identification that helps in determining the language of individual textual entities from a code-mixed corpus is a current and relevant classification problem. Code-mixed texts are difficult to interpret and analyze from an algorithmic perspective. However, highly complex transformer-based techniques can be used to analyze and identify distinct languages of words in code-mixed texts. Kannada is one of the Dravidian languages which is spoken and written in Karnataka, India. This study aims to identify the language of individual words of texts from a corpus of code-mixed Kannada-English texts using transformer-based techniques. The proposed *Distilka* model was developed by fine-tuning the DistilBERT model using the code-mixed corpus. This model performed best on the official test dataset with a macro-averaged F1-score of 0.62 and weighted precision score of 0.86. The proposed solution ranked first in the shared task.

1 Introduction

Language identification is the process of determining the natural language that a document is written in. Automatic language identification has been widely researched for 50 years (Jauhiainen et al., 2019). However, while recognizing text in different languages might come naturally to a human reader, it is still challenging for the computer. Natural Language Processing (NLP) focuses on teaching computers to comprehend spoken and written language in a manner similar to humans. NLP research is evolving and rapidly expanding. Classification tasks such as text classification, sentiment analysis, name entity recognition, and speech recognition have been solved in the past using algorithms of NLP. Similarly, language identification, is a recent research problem that can be dealt with using NLP techniques (Shanmugalingam et al., 2018)

Focusing on language identification when it comes to a low-resource language or a mixed language, in order to identify the language could be a huge challenge. As we know, India has a rich language culture covering different geographical areas such as Hindi, Bengali, and Kannada. Kannada is spoken mainly in Karnataka, which is a southern state of India (Kumar et al., 2015). People of Karnataka read, write, and speak Kannada, but many find it difficult to use Kannada script to post comments on social media. As a result, Kannada is a low-resource language since social media users typically use Roman script or a combination of Kannada and Roman script. In this shared task, a dataset was created using Kannada YouTube comments named “CoLI-Kanglish”. (Balouchzahi et al., 2022; Shashirekha et al., 2022).

The main objective of this shared task is to create a novel method for language identification in mixed languages that consists of tokens from English, Kannada, mixed languages of Kannada and English, name, location, and other categories. In this study, the dataset is initially prepared by

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 7 - 11
December 15-18, 2022 ©2022 Association for Computational Linguistics

task organizers. Then, the data processing technique is utilized. Finally, we evaluate the model using macro-averages and weighted average scores. The following points are the contribution of our study:

- Exploratory data analysis of the dataset

- Create the Distilka (**Distil**BERT + **Ka**nnada) model based on the Transformer-based DistilBERT.

To the best of our knowledge, this is the first study that applies a DistilBERT-based model to identify mixed language in Kannada and English (CoLI-Kanglish) datasets, and the fact that it presents the best performance is highlighted.

2 Related works

Language identification has been studied for half a decade, and automatic language identification has been proposed rigorously in social media data. In 2014, Barman et al, presented an initial study on automatic language identification using Indian language code mixed in social media communication. The dataset of Bengali, Hindi, and English in Facebook comments. The authors conclude that character n-gram features, contextual information is also important, and information from dictionaries can be useful for Language Identification tasks. Apart from Facebook data, the studies also investigate Twitter data with Support Vector Machine (linear kernel), which contains bilingual tweets written in the most commonly used Iberian languages (i.e., Spanish, Portuguese, Catalan, Basque, and Galician) as well as English language. The study achieved 0.792 for macro-F1 (Pla & Hurtado, 2017).

In the same year, 2017, Transformers were introduced. The paper "Attention is all you need," describing attention mechanisms, provides context for any position in the input sequence. (Vaswani et al., 2017). Furthermore, if the input data is a natural language sentence, the transformer does not have to process one word at a time. This allows for more parallelization than a recurrent neural network and therefore reduces training times. In 2018, BERT (Bidirectional Encoder Representations from Transformers) from transformer-based technique, was developed with large amounts of pre-trained data and the ability to capture context, so BERT has become a well-known architecture since then (Devlin et al., 2018). Due to some constraints of BERT, DistilBERT has emerged to optimize the training by reducing the size of BERT and increasing the speed of BERT,while trying to retain as much performance as possible. Moreover, DistilBERT is 40% smaller than the original BERT base model, is 60% faster than it, and retains 97% of its functionality.

DistilBERT has been used in a variety of text classification tasks, including language identification. There are several papers using DistilBERT for text classification tasks, for example. Bambroo & Awasthi, 2021 proposed a fine-tuned DistilBERT on legal-domain specific corpora and discovered that this model outperformed other algorithms while also being faster at the task of legal document classification. Another study is to carry out a word-level language identification (WLLI) of Malayalam-English code-mixed data from YouTube. According to the study, DistilBERT produced the highest precision score with 91.74% in Hindi and English pairs (Thara & Poornachandran, 2021).

3 Experiments

3.1. Datasets

The CoLI-Kanglish dataset includes English and Kannada words written in Roman script and is divided into six labels: "Kannada," "English," "Mixed-language," "Name," "Location," and "Other. The CoLI-Kanglish (train dataset) contains 14,847 tokens, and there are 6 tags. Table 1 shows the number of entries in each category. In addition, the test dataset consists of 4,585 tokens without labels. The example of the dataset can be found in Table 2.

Category	Tag	Count
Kannada	kn	6,526
English	en	4,469
Mixed-Language	kn-en	1,379
Name	name	708
Location	location	102
Other	other	1,663

Table 1. The description and samples of tokens in CoLI-Kanglish Dataset

Word	Tag
hegilla	kn
staying	en
aparictarannu	en-kn
kamal	name
bangalore	location
mamao	other

Table 2. The example of CoLI-Kanglish Dataset in each tag

3.2 DistilBERT

A distilled version of BERT that is smaller, quicker, less expensive, and lighter was proposed by Sanh et al., 2019. DistilBERT is a BERT base-trained transformer model that is compact, quick, affordable, and light. It runs 60% faster with 40% fewer parameters than BERT-base-uncased while maintaining over 95% of BERT's performance as measured by the GLUE language understanding benchmark. When compared to other models, DistilBERT produces the quickest results with 106 seconds (Bambroo & Awasthi, 2021).

3.3 Distilka model

Distilka is the fine-tuned model in the mix-language Kannada and English identification tasks by using DistilBERT-based categorization with 6 labels such as "Kannada," "English", "Mixed-language", "Name", "Location", and "Other". This model can be downloaded from the Hugging Face Hub (https://huggingface.co/tiya1012/distilka_applied), and the framework of this study is illustrated in Fig 1.

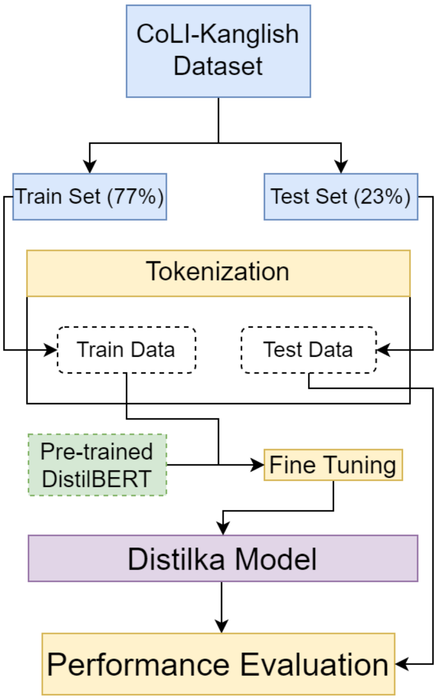

Fig 1. Framework of Model Development for Language Identification using CoLI-Kanglish Dataset

4 Experiments and Results

4.1 Experimental setting

The notebook ran on Google Colab Pro with Python 3 installed, and the model was fine-tuned. We set the learning rate at 2e-5, the maximum sequence length at 512, and the gradient accumulation steps at 1 and batch size was set at 6 as shown in Table 3. The optimal results were obtained through a comparative study which is shown in Table 4.

Parameters	Values
Maximum sequence length	512
Learning rate	2e-5
Accumulated gradient steps	1
Batch Size	6

Table 3. Model Hyperparameters

4.2 Results and Discussion

As we can see from Table 4, it represents the weighted and macro precision and F1-score. In this shared task, ranking will be finalized using macro F1 score. Since the Distilka model was trained and learned from the DistilBERT-based-cased, its macro F1-score is 0.62. There are several factors that contribute to the best performance of DistilBERT; for instance, DistilBERT is pretrained on the same data as BERT, which is BookCorpus, a dataset consisting of 11,038 unpublished books and English Wikipedia. Although this dataset contains Kannada language, it has been written in English. Furthermore, the model can learn better if more datasets have been trained. In the fine-tuning period, language tag data is used based on the ability of the language model to improve the performance of the downstream tasks. This enables DistilBERT to achieve cutting-edge results in written Kannada-English language benchmarks.

Model	Precision (weighted)	F1-score (weighted)	F1-score (Macro)
Distilka	0.87	0.86	0.62

Table 4. Results of Distilka model

5 Conclusion

In this paper, we describe our proposed method for the shared task of word-level language identification in code-mixed Kannada-English dataset. On comparing with the performance metrics of other solutions based on DistilBERT, that were developed for this shared task, it was found out that the Distilka model performed significantly better with a macro-averaged F1-score of 0.62. The proposed model secured the first rank in the shared task. In future work, we will try to adjust the parameters of the new model in order to improve its performance significantly. Future work will include further application of language identification tasks to several low-resource languages.

References

Purbid Bambroo and Aditi Awasthi. 2021. LegalDB: Long DistilBERT for Legal Document Classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4. IEEE.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code Mixing: A Challenge for Language Identification in the Language of Social Media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65.

K. M. Anil Kumar, N. Rajasimha, Manovikas Reddy, A. Rajanarayana, and Kewal Nadgir. 2015. Analysis of users' Sentiments from Kannada Web Documents. *Procedia Computer Science*, 54:247–256.

Ferran Pla and Lluís-F. Hurtado. 2017. Language identification of multilingual posts from Twitter: a case study. *Knowledge and Information Systems*, 51(3):965–989.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.

Kasthuri Shanmugalingam, Sagara Sumathipala, and Chinthaka Premachandra. 2018. Word Level Language Identification of Code Mixing Text in Social Media using NLP. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–5. IEEE.

Fazlourrahman Balouchzahi, Sabur Butt, Asha Hagde, Noman Ashraf, Shashirekha Hosahalli Lakshmaiah, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022. *19th International Conference on Natural Language Processing Proceedings*.

H L Shashirekha, F Balouchzahi, M D Anusha, and G Sidorov. 2022. CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts.

S. Thara and Prabaharan Poornachandran. 2021. Transformer Based Language Identification for Malayalam-English Code-Mixed Text. *IEEE Access*, 9:118837–118850.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2022. Towards Offensive Language Identification for Tamil Code-Mixed YouTube Comments and Posts. *SN Computer Science*, 3(1):94.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need.