BFCAI at CoLI-Tunglish@FIRE 2023: Machine Learning Based Model for Word-level Language Identification in Code-mixed Tulu Texts

Ahmed M. Fetouh, Hamada Nayel

Computer Science Department, Faculty of Computers and Artificial Intelligence, Benha University, Egypt

Abstract

This paper describes the model submitted by the BFCAI team for the CoLI-Tunglish shared task held at FIRE 2023. The proposed model used a character n-gram TF-IDF vectorization as a representation scheme. TF-IDF has been enhanced using word length, then applied several Machine Learning algorithms namely; Support Vector Machines, Stochastic Gradient Descent, K-Nearest Neighbors and Multi-layer Perceptron. SVM outperformed all other classifiers. All submissions are considered and ranked based on the macro average F1 score. SVM reported an F1 score of 81.2% on the test set and achieved the second rank among all other submissions.

Keywords

Language Identification, Natural Language Processing, Machine Learning, Character N-Gram, Text Classification.

1. Introduction

Automatically recognizing the languages present in a document, known as Language Identification (LI) [1]. It's a critical task in various text processing pipelines. LI involves text classification, where texts are assigned to specific language categories. The rise of social media platforms have led to a large amount of text data, including code-mixed content [2]. Code mixing represent the texts that contain multiple language. It become common in social media platforms, where users often combine their tongue language with other languages to express their opinions online. LI can be represented and solved using Natural Language Processing (NLP) approaches. Dealing with diverse languages found in social media documents poses a significant challenge in NLP. State-of-the-art NLP methods employ word embedding and n-gram-based models at the character or word level for tasks like LI [3]. However, accurately identifying languages in code-mixed texts from social media remains a difficult problem for NLP.

There are several challenges to LI in code-mixed text. One challenge is that the words from different languages may be mixed, making it challenge to distinguish the boundaries between the languages. The fact that a word might have several meanings depending on the language presents another challenge, which can make it difficult to distinguish the

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

ahmed192081@fci.bu.edu.eg (A. M. Fetouh); hamada.ali@fci.bu.edu.eg (H. Nayel)

language of the word. Additionally, the use of slang, informal language and abbreviations can also make it difficult for LI systems to accurately identify the language of the text [4].

Tulu is a regional language and Kannada is the language of the Indian state of Karnataka. Tuluvas, who are native speakers of Tulu, typically know both Tulu and Kannada languages fluently. Furthermore, many Kannada terms are used in the Tulu language. English is also widely known by many Tulu speakers, especially those who using the social media platforms.

Tulu songs, videos, movies, comedy programs and skits are popular on social media. The comments posted by Tulu users of Tulu programs on social media are often a code-mix of Tulu, Kannada, and English. This is because many Tuluvas face difficulties in using the Kannada script to post messages or comments on social media due to the technological limitations of key- boards/keypads on computers/smartphones. Additionally, the complexity of framing words with consonant conjuncts makes it challenging to type Tulu using the Kannada alphabet. For this reason, many people only use the Roman alphabet or a combination of Kannada and Roman script to post comments on social media [5].

The primary goal of this collaborative task is to develop a new approach for LI in mixed languages. The task entails dealing with tokens from various categories, including English, Kannada, Mixed-language, Tulu, names, locations, symbol, and other [6, 7]. To address this issue, the paper employs a variety of machine learning models as well as the TF-IDF representation schema. Furthermore, the word length is used to improve the word representation. This allows the algorithms to more accurately identify the language of a new word.

Category	Description
Kannada	Kannada words written in Roman script
English	Pure English words
Tulu	Tulu words written in Roman script
Mixed-language	Combination of Kannada, Tulu and/or English words in Roman script
Name	Words that indicate name of person (including Indian names)
Location	Words that indicate locations
Other	Words not belonging to any of the above categories and words of other languages

Table 1
Description of the CoLI-Tunglish dataset.

2. Dataset

The CoLI-Tunglish dataset [5] for the shared task CoLI-Tunglish [7] contains English and Kannada words written in Roman script. The data is divided into eight categories: **Tulu**, **Kannada**, **English**, **Mixed-language**, **Name**, **sym**, **Location** and **Other**. Description of the dataset are shown in Table 1. The dataset is divided into training, development, and test set composed of 21,727, 3,582, 10,506 samples respectively. The distribution of training set tokens across the labels is illustrated in Figure 1.

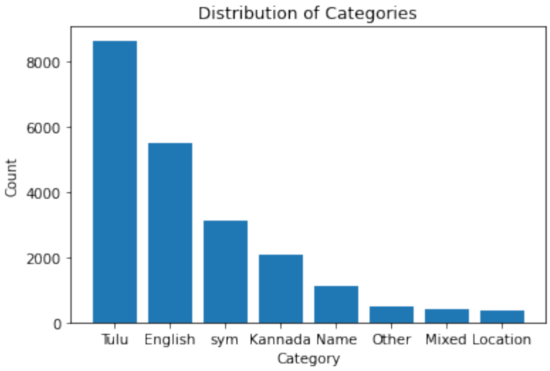

Figure 1: Distribution of Categories

This code-mixed dataset allows for developing systems that can handle English-Kannada LI at the word level. The diversity of categories enables models to distinguish between names, locations, mixed-language words, and other tokens. Overall, the CoLI-Tunglish dataset provides a challenging benchmark for evaluating language identification systems on Romanized English-Kannada code-mixed data.

3. Methodology

In this section, a thorough explanation of the methodology we employed will be presented, which involves utilizing TF-IDF vectorization. Furthermore, we perform a comparison of the results acquired from several machine learning models.

3.1. Feature Extraction

In this study, TF-IDF was used with character n-gram features to represent each word as a feature vector. The n-gram range is set to (1, 4), which means that we will consider character unigram, bigram, trigram, and 4-gram. TF-IDF is a numerical statistic that reflects the importance of a term (in this case, an n-gram) in a document within a collection of documents. It combines the ideas of how frequently a word appears in a document (term frequency/TF) and how important a word is. Based on the total number of documents in which it appears (inverse document frequency/IDF) is calculated by the following formula:

$$w_{dt} = tf_{dt} \cdot \log\left(\frac{N+1}{df_t+1}\right) \tag{1}$$

Where, w_{dt} is the weight of word d in vector t, tf_{dt} is the count of word d in document t, N is the total number of words, and df_t is the count of word d in all words.

By calculating the TF-IDF values for each n-gram, we obtain a numerical feature matrix that represents the textual data effectively [8].

3.2. Feature Enhancement

By augmenting the dataset to incorporate the additional attribute of "word length", as depicted in Table 3, we capture additional information about the words beyond their textual content. This feature provides valuable insights into the relationship between word length and language classification. The combination of the TF-IDF and word length enable a comprehensive analysis of the dataset. This leads to increased accuracy and a better understanding of the core patterns in language identification.

Words	Language	Word Length
Oo	English	2
anna	Kannada	4
ninna	Tulu	5
pukuli	Tulu	6
naddh	Tulu	5
korpa	Tulu	5
.	sym	1
shivam	Name	6
music	English	5
movie	English	5

Table 2
Words, Language, and Word Length

The feature of word length can be instrumental in differentiating between languages by providing valuable insights into the unique word structures and patterns of each language. When analyzing the average word length, along with other statistical measures such as the distribution of word lengths, we can effectively distinguish one language from another.

In the case of the provided data [5], as show in Table 3 we can observe that English has an average word length of 4.710, while Kannada has an average word length of 5.968 and Tulu has an average word length of 5.458. These variations in average word length indicate differences in the linguistic structures of these languages. Additionally, the average word length for sym is 1.0, suggesting that it likely represents a language or category with very short words (dot). Furthermore, the average word length for Name is 6.006, indicating that it might be a category associated with personal names or proper nouns. The Mixed category has an average word length of 7.469, suggesting that it contains a combination of multiple languages or texts with longer words. Finally, the average word length for Location is 6.878, indicating that it might represent a category related to geographical locations.

3.3. Classification

For the classification task, we employed a variety of machine learning algorithms to evaluate the effectiveness of our method. These algorithms

Language	Average Word Length
English	4.710
Kannada	5.968
Tulu	5.458
sym	1.0
Name	6.006
Mixed	7.469
Location	6.878
Other	5.144

Table 3
Average Word Length for Different Languages

are Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), K-Nearest Neighbors (KNN), and Multi-layer Perceptron (MLP) classifiers.

- **SVM**: is a powerful supervised learning algorithm that can be used for dialect classification [9]. It finds an optimal hyperplane in a high-dimensional space to classify data points into different classes. SVMs are effective but can be computationally expensive.

- **SGD**: is a well-known optimization algorithm for training machine learning models in NLP tasks such as LI [10]. It works by iteratively updating the model's parameters in the direction of the negative gradient of the loss function. The gradient is calculated using a portion of the training data, called a mini-batch. This makes SGD computationally efficient, even for large datasets.

- **KNN**: is a non-parametric, supervised machine learning algorithm that can be used for both classification and regression tasks. It works with the aid of finding the k most comparable instances in the training set to a new instance, Following that, the new instance is assigned to the class of the majority of those k instances. KNN is a versatile and powerful algorithm that may be used for a range of NLP tasks such as rumor detection [11].

- **MLP**: is artificial neural network composed of multiple layers of interconnected nodes (neurons). It uses forward propagation to compute outputs based on weighted sums and activation functions. MLPs can learn complex non-linear relationships between input and output data, making them suitable for various tasks such as LI [12]. However, they require careful tuning of hyper parameters and can be prone to over-fitting if not properly regularized.

3.4. Evaluation Metrics

Macro-averaged and weighted-averaged scores have been used to evaluate the task. However final ranking will be based on macro-averaged F1 score [7].

4. Experiments and Results

In this task, we explore the performance of various machine learning models on the CoLI-Tunglish LI dataset [5]. As shown in Table 4, we implemented four popular algorithms SVM, KNN, SGD and MLP. Our goal was to determine which model accurately classify the mixed-language tokens in the development set.

Algorithm	Macro avg			Weighted avg		
	Precision	Recall	F1 score	Precision	Recall	F1 score
SVM(Linear)	0.87	0.80	0.83	0.90	0.90	0.90
MLP(10 nodes)	0.85	0.80	0.82	0.90	0.90	0.90
SGD	0.86	0.67	0.72	0.87	0.87	0.86
KNN	0.78	0.66	0.69	0.82	0.83	0.82

Table 4
Comparison of machine learning algorithms scores on the development set

A variety of evaluation metrics were considered to properly assess model performance on this challenging mixed-language task. Precision, recall, F1 score, macro average and weighted average were all measured.

The results indicate that SVM achieved the highest performance with a weighted average F1 score of 0.90 and a macro average F1 score of 0.83 on the development set.

We've got additionally compared our work with the top-ranked groups for the CoLI-Tunglish shared task. The results shown in Table 5 Demonstrate that SVM-based submission achieved the second highest F1 score among all teams. This comparison highlights the effectiveness of our approach in language identification in code-mixed Tulu texts. Our team submitted three runs using three different ML models, namely SGD, MLP, and SVM. The results of all runs can be observed in Table 6.

The source code of the proposed model is available at GitHub [1].

Team Name	Run Name	Precision	Recall	F1 score
SATLAB	Run2	0.851	0.783	0.813
BFCAI(Ours)	Run3	0.859	0.777	0.812
Poorvi	Run1	0.821	0.781	0.799
MUCS	Run1	0.807	0.743	0.770
IRLab@IITBHU	Run1	0.740	0.571	0.602

Table 5
Comparison of macro average scores with top ranked teams in Code-mixed Tulu Texts Shared Task

[1] https://github.com/Ahmedmegahed72/CoLI-Tunglish

Team Name	Run Name	Precision	Recall	F1 score
BFCAI	Run1 (SGD)	0.833	0.776	0.801
BFCAI	Run2 (MLP)	0.867	0.695	0.745
BFCAI	Run3 (SVM)	0.859	0.777	0.812

Table 6
Run-wise Rank List for BFCAI's different runs on the test set

5. Conclusion

In this study, we described our system submitted to the CoLI-Tunglish shared task on word-level language identification in code-mixed Tulu texts. We explored a variety of machine-learning approaches and found that a character n-gram TF-IDF based feature representation and word length combined with an SVM classifier achieved the best performance for this task. In future work, we plan to discover the usage of pre-trained models to improve the performance of classification. We believe that transfer learning can be a valuable tool for this task.

References

[1] T. Jauhiainen, M. Lui, M. Zampieri, T. Baldwin, K. Lindén, Automatic language identification in texts: A survey, Journal of Artificial Intelligence Research 65 (2019) 675–782.

[2] S. Dowlagar, R. Mamidi, A survey of recent neural network models on code-mixed indian hate speech data, in: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE '21, Association for Computing Machinery, New York, NY, USA, 2022, p. 67–74. URL: https://doi.org/10.1145/3503162.3503168. doi:10.1145/3503162.3503168.

[3] F. Balouchzahi, H. Shashirekha, Mucs@ dravidian-codemix-fire2020: Saco-sentimentsanalysis for codemix text., in: FIRE (Working Notes), 2020, pp. 495–502.

[4] U. Barman, A. Das, J. Wagner, J. Foster, 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, 2014, pp. 13–23.

[5] 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.

[6] 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.

[7] 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.

[8] S. Selva Birunda, R. Kanniga Devi, A review on word embedding techniques for text classification, Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020 (2021) 267–281.

[9] H. A. Nayel, H. Shashirekha, Mangalore-university@ inli-fire-2017: Indian native language identification using support vector machines and ensemble approach., in: FIRE (Working Notes), 2017, pp. 106–109.

[10] D. Kosmajac, V. Keselj, Dalteam@ inli-fire-2017: Native language identification using svm with sgd training., in: FIRE (Working Notes), 2017, pp. 118–122.

[11] N. Ashraf, H. Nayel, M. Taha, A comparative study of machine learning approaches for rumors detection in covid-19 tweets, in: 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2022, pp. 384–387. doi:10.1109/MIUCC55081.2022.9781707.

[12] A. Bhola, K. N. Reddy, M. J. Kumar, A. Tiwari, Language identification using multi-layer perceptron, in: 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 2023, pp. 1018–1022. doi:10.1109/CISES58720.2023.10183574.