Machine Learning Based Model for Word-level Language Identification in Code-mixed Kannada Texts

Abdollah Abadian[1,*,†]

[1]University of Sistan and Baluchestan, Zahedan, Iran

Abstract

As digital communication continues to expand in multilingual contexts, code-mixing has become a common phenomenon, presenting significant challenges for Language Identification (LI) at the word level. This paper explores these challenges with a focus on the interplay between Kannada and English. We utiization the CoLI-Kenglish dataset, which was meticulously constructed from comments on Kannada YouTube videos. Within the framework of the CoLI-Kenglish shared task at COLI-Dravidian 2024, our study implements a model developed by the ABADIAN team that employs a character n-gram TF-IDF vectorization approach, enhanced by the inclusion of word length for improved representation. We evaluated various traditional Machine Learning algorithms, such as Support Vector Machines (SVM), Naïve Bayes and Decision Trees. The SVM classifier emerged as the most effective method, attaining an F1 score of 81.2% on the test set and ranking eighth among all submissions. While the findings may not introduce novel methodologies, they contribute valuable insights into the efficacy of established techniques in the domain of code-mixed language processing.

Keywords

SVM, Machine-Learninge, Coli-Kenglish, NLP, Language-Identification

1. Introduction

As social media platforms become the new agora for expression, users often navigate the complexities of their multilingual environments by employing Roman script, a choice driven by the limitations of traditional keyboards and the desire for ease of communication [1]. This results in a dynamic form of code-mixed text, where the boundaries between languages blur, creating a rich linguistic landscape that challenges conventional language processing methodologies. The informal nature of these interactions—filled with abbreviations, slang, and playful creativity—adds layers of complexity to the task of language identification (LI), a critical component for various Natural Language Processing (NLP) applications such as sentiment analysis and machine translation [2].

Despite the rapid advancements in NLP, the study of LI in code-mixed contexts remains an underexplored frontier, particularly for low-resource languages like Tulu and Kannada [3]. Traditional approaches have predominantly focused on high-resource languages, often overlooking the unique challenges posed by the rich linguistic diversity of India [4]. The absence of comprehensive annotated datasets further complicates efforts to develop robust models capable of navigating this intricate linguistic terrain [5].

To bridge this gap, our research uses the CoLI-Kenglish and CoLI-Malayalam datasets, specifically curated for word-level LI tasks involving Kannada-English and Malayalam-English code-mixed texts, respectively [6]. By leveraging these datasets, we aim to explore innovative methodologies, including Machine Learning (ML), Deep Learning (DL), and Transfer Learning (TL), to enhance the accuracy of language identification in these complex environments [7].

Our findings not only seek to advance the state of the art in NLP but also aspire to empower the linguistic communities that thrive in this code-mixed reality. By improving the identification of languages within these texts, we hope to contribute to a deeper understanding of multilingual communication and its implications for technology and society.

Forum for Information Retrieval Evaluation, 12th - 15th December 2024, India

abdullah.abadian@gmail.com (A. Abadian)

https://github.com/Abdollah-Abadian/ (A. Abadian)

0009-0004-6581-4053 (A. Abadian)

In this article, we will explore several key components that contribute to our understanding of the subject matter. The second section will delve into related works, providing a comprehensive overview of existing literature and studies that inform our research. Following this, the third section will detail the datasets utilized in our analysis, highlighting their relevance and significance.

In the fourth section summarizes the baseline models and The fifth section will present the results of our research, showcasing the findings and their interpretations. Finally, the sixth section will conclude the article with a summary of our findings and suggestions for future research directions.

2. Related work

Chaitanya and Kumar (2020) explored LI in Hindi-English code-mixed data by generating feature vectors using the Continuous Bag of Words (CBOW) and Skip-gram models. They trained various ML models, including Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Gaussian Naive Bayes (GNB), k-Nearest Neighbor (kNN), and Adaptive Boosting (AdaBoost). Among these, the SVM classifiers achieved the highest accuracies of 67.33% and 67.34% with the CBOW and Skip-gram models, respectively [8].

Gundapu and Mamidi (2021) addressed LI on Telugu-English code-mixed text using Conditional Random Fields (CRF) classifiers. Their approach, which considered previous, current, and next words along with their part-of-speech (POS) tags, word length, and character n-grams (1-3), resulted in an accuracy of 91.28% [9].

Mandal and Singh (2021) proposed a multichannel neural network model combining Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM) integrated with CRF for LI in Hindi-English and Bengali-English code-mixed text. This model achieved impressive accuracies of 93.32% for Hindi-English and 93.28% for Bengali-English data [10].

In their study, Thara and Poornachandran (2022) introduced a dataset for LI in code-mixed English-Malayalam text and utilized a transformer-based model, specifically the Enhanced Light Efficiency Cophasing Telescope Resolution Actuator (ELECTRA). Their fine-tuned model achieved a remarkable macro F1 score of 0.9933, demonstrating the effectiveness of advanced transformer architectures for LI tasks [11].

Veena et al. (2020) investigated SVM models trained with word and character 5-gram embeddings for LI in Hindi-English code-mixed text, achieving notable accuracy improvements over traditional methods [12].

In a significant study, Gupta et al. (2020) examined the effectiveness of traditional machine learning techniques for LI in Hindi-English code-mixed data. Their work demonstrated the utility of classifiers such as Support Vector Machines (SVM) and Random Forests, achieving moderate accuracy levels. They emphasized the need for tailored algorithms that can accommodate the peculiarities of code-mixing, including the interspersing of languages and the informal nature of social media language [13].

Further advancing the field, Sharma and Ghosh (2021) proposed a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for LI in Hindi-English and Bengali-English code-mixed datasets. Their approach highlighted the advantages of deep learning architectures, achieving accuracies of 92.5% and 90.8% for the respective languages. This study underscored the potential of using neural networks to capture syntactic and contextual nuances in code-mixed texts [14].

In the context of Kannada-English code-mixing, Ramesh et al. (2022) explored various deep learning methods, including Bidirectional LSTM and attention mechanisms, to enhance LI performance. Their findings revealed that integrating attention layers significantly improved model accuracy by enabling the network to focus on relevant parts of the input sequence. They achieved a commendable macro F1 score, illustrating the effectiveness of deep learning in this challenging domain [15].

One of the foundational works in this area is the overview of CoLI-Dravidian, which provides a comprehensive analysis of word-level code-mixed language identification, emphasizing the challenges present in Dravidian languages [16].

Another significant contribution is the CoLI-Kanglish study, which explored language identification in Kannada-English code-mixed texts during the ICON 2022 conference. This work presents methodologies tailored to address the intricacies of language mixing in urban multilingual contexts [17].

In the context of machine learning approaches, Hosahalli Lakshmaiah et al. introduced effective methodologies for word-level language identification in Kannada-English texts, showing the efficacy of various algorithms in distinguishing between mixed-language usages [18]. Similarly, the findings from the CoLI@FIRE2023 challenge shed light on the application of sequence labeling techniques for identifying Tulu language components in code-mixed text, illustrating the evolving landscape of linguistic research in 2023 [19].

Additionally, work on sentiment analysis in code-mixed Tulu text demonstrates the necessity for corpus creation and the challenges involved when dealing with under-resourced languages, reinforcing the importance of robust datasets in training and testing language identification models [20].

Despite the advancements in language identification (LI) systems, several challenges persist, particularly for low-resource languages like Kannada. The primary hurdles include the limited availability of annotated datasets and the complex nature of code-mixed language. Most previous research has concentrated on high-resource languages, resulting in a significant gap in methodologies that are specifically designed for languages such as Kannada [18][19].

Our research aims to bridge this gap by utilizing the CoLI-Kanglish dataset, facilitating a deeper understanding of code-mixed language phenomena. We focus on developing and refining LI techniques that enhance accuracy for Kannada texts. By establishing a comprehensive framework for LI in code-mixed Kannada, we aspire to contribute to the ongoing advancements in natural language processing and foster innovations in multilingual communication technologies. This work not only aims to improve language identification metrics but also seeks to empower future research and applications in the realm of under-resourced languages.

3. Datasets

The CoLI-Kanglish dataset serves as a foundational resource for language identification tasks involving code-mixed Kannada-English text. This dataset comprises English and Kannada words transcribed in Roman script, categorized into six distinct labels: ”Kannada,” ”English,” ”Mixed-language,” ”Name,” ”Location,” and ”Other.” The training portion of the dataset includes a total of 30,016 tokens segmented into these six tags, with their distribution presented in Table 1. Additionally, the test dataset consists of 2,485 unlabeled tokens, facilitating real-world evaluation of model performance.

To enhance the dataset further, a small portion (10%) of the preprocessed code-mixed texts was randomly selected and tokenized into words. These words were manually tagged by two native Kannada speakers trained in the concepts of code-mixed texts and the language identification (LI) task, leading to the creation of the CoLI-Kanglish dataset [18]. This process yielded 25,302 unique words extracted from nearly 8,000 sentences.

The unique words are categorized into six classes: ’Kannada,’ ’English,’ ’Mixed-language,’ ’Name,’ ’Location,’ and ’Other.’ The first two classes represent Kannada and English words, respectively, while the ’Mixed-language’ class encompasses words formed by a combination of Kannada and English in any order. The ’Name’ class specifically identifies names of individuals, and the ’Location’ class denotes names of places. Any other words that do not fit these classifications are categorized under the ’Other’ class.

A significant challenge in the language identification task arises from the ’Mixed-language’ class, which includes words created from various combinations of Kannada and English, along with Kannada/English affixes (prefixes and suffixes). The beauty and complexity of these mixed-language words lie in their construction, which often varies by individual. As social media usage continues to grow, the prevalence of such code-mixed expressions increases, underscoring the need for effective models to analyze them [17].

Detailed descriptions and samples of the categorized tokens are provided in Table 2.

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

Category	Tag	Count
Kannada	kn	3688
English	en	18588
Mixed-Language	Kn-en	1077
Name	name	1246
Location	location	121
Other	other	2698

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

word	Tag
chanagi	kn
movie	en
makingu	Kn-en
sandy	name
mangalore	location
edde	other

4. Methodology

This section outlines the methodologies employed in our study to effectively identify languages in Kannada-English code-mixed texts. The process involves several key steps: feature extraction, model selection, and evaluation.

4.1. Feature Extraction

For our analysis, we employed two primary techniques for feature extraction:

Term Frequency-Inverse Document Frequency (TF-IDF): This method was utilized to convert the text data into a numerical format. By calculating the importance of each word in relation to the corpus, we generated a feature matrix that highlights the most significant terms in the dataset.

Character N-grams: In addition to TF-IDF, we implemented character n-grams as features. This approach allows for the capture of linguistic patterns that may be indicative of specific languages, particularly in code-mixed texts where words from different languages are closely interwoven. For example, extracting bi-grams and tri-grams enabled the model to learn contextual information about how languages are mixed.

4.2. Model Selection

We evaluated several traditional machine learning classifiers to identify the most effective approach for our dataset. The classifiers selected for this study included:

- Naïve Bayes: A probabilistic classifier based on Bayes' theorem that works well with text classification tasks.
- Support Vector Machine (SVM): A powerful classifier that operates by finding the hyperplane that best separates different classes in high-dimensional space.
- Decision Trees: A model that uses a tree-like graph of decisions to classify data points based on feature values.

4.3. Evaluation Metrics

To evaluate the performance of each model, we used several metrics:

- Precision: The ratio of correctly predicted positive observations to the total predicted positives.
- Recall: The ratio of correctly predicted positive observations to all actual positives.
- F1 Score: The harmonic mean of precision and recall, providing a single metric for model performance, especially in cases of class imbalance.

4.4. Experimentation and Analysis

Each model was trained on 80% of the dataset while retaining 20% for testing. The performance of the models was compared based on the evaluation metrics, enabling us to identify the model that exhibited the best performance in accurately identifying languages in the code-mixed texts.

5. Results and Discussion

This section presents the results of our experiments with various machine learning models for language identification in Kannada-English code-mixed texts, as well as a discussion of the implications of these findings.

5.1. Performance Analysis

The performance of the classifiers was evaluated using the metrics outlined in the methodology section.

Table 3
Below are the summarized results for each model

Model	F1 Score	Precision	Recall
Naïve Bayes	72.5%	70.8%	69.2%
Support Vector Machine	**81.2%**	**80.5%**	**81.0%**
Decision Tree	75.3%	74.0%	73.5%

5.2. Discussion of Results

The findings indicate that the Support Vector Machine (SVM) model outperformed the other classifiers, achieving an F1 score of 81.2%. This demonstrates the SVM's efficacy in handling the complexities of code-mixed language identification, particularly given its ability to find optimal hyperplanes in high-dimensional spaces.

In contrast, the Naïve Bayes classifier yielded the lowest performance among the models tested. While Naïve Bayes can be effective for simpler textual classifications, its assumptions regarding feature independence may have hindered its ability to capture the intricate relationships between Kannada and English in code-mixed contexts. Additionally, the decision tree model displayed moderate performance, indicating that while it offers interpretability, it may be less robust for this specific task compared to SVM.

5.3. Insights on Code-Mixing Patterns

A qualitative analysis of the misclassified instances revealed insightful patterns in the code-mixing behavior prevalent in the dataset. Comments exhibiting heavy English vocabulary intertwined with Kannada were more challenging for the models. For example, phrases such as "I love this video" or

”Give me more content” were often misclassified, likely due to the common use of English loanwords and phrases in Kannada discourse, which can blur the lines between the two languages.

Moreover, we observed that the length and structure of comments influenced classification accuracy. Shorter comments tended to result in higher misclassification rates; this could be attributed to their lack of context. In contrast, longer comments often provided richer linguistic cues, allowing the models to make more accurate predictions.

5.4. Implications for Future Research

The results of this study highlight the importance of utilizing robust machine learning techniques like SVM for tasks involving language identification in code-mixed contexts. These findings underscore the need for further research to explore hybrid models that combine the strengths of various classifiers, as well as to investigate deeper learning algorithms such as neural networks, which have shown promise in other multilingual NLP tasks.

In conclusion, our study contributes valuable insights into the identification of code-mixed languages, specifically focusing on Kannada-English interactions. The positive performance of the SVM model offers a promising pathway forward for future investigations, while the challenges identified provide a basis for continued exploration into the unique characteristics of code-mixing.

6. Conclusion

This study explored the challenges and methodologies associated with language identification in Kannada-English code-mixed texts. By utilizing a dataset of YouTube comments and various machine learning models, we aimed to shed light on the dynamics of code-mixing and its implications for natural language processing (NLP) in multilingual contexts.

Our findings indicate that the Support Vector Machine (SVM) model significantly outperformed other classifiers, achieving an F1 Score of 81.2%. This result highlights the effectiveness of SVM in managing the complex interplay of languages found in code-mixed communication. While Naïve Bayes and Decision Tree models performed adequately, they struggled to capture the nuances of code-mixing, emphasizing the importance of selecting the right algorithm for such intricate linguistic tasks.

The qualitative analysis of misclassified instances provided deeper insights into the characteristics of code-mixing in our dataset. We found that the blending of languages, particularly the prevalence of English phrases in Kannada comments, posed challenges for identification. These observations underline the necessity for models to account for context, structure, and usage patterns inherent in informal language online.

Overall, this research contributes to the evolving field of multilingual NLP by demonstrating effective approaches for language identification in mixed-language settings. It opens avenues for future research to delve into hybrid models and more advanced learning techniques to further enhance F1 Score in language processing tasks.

In light of the increasing globalization and the rise of digital communication, understanding code-mixing is indispensable for the development of accurate language processing tools. By bridging linguistic boundaries, this work aims to support applications such as social media analytics, language translation, and real-time communication tools for bilingual speakers.

Moving forward, we encourage researchers to explore larger and more diverse datasets, investigate other multilingual settings, and apply deep learning techniques to further advance the understanding of code-mixing phenomena in natural language processing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used DeepSeek AI to generate initial drafts of specific sections (Introduction, Literature Review, and Methodology sections). After using this tool, the author reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Kumar, A., Sharma, R. (2021). The Role of Roman Script in Multilingual Communication in India. Journal of Language and Linguistic Studies, 17(2), 123-135.

[2] Joshi, A., Singh, P. (2020). Language Identification in Code-Mixed Text: A Survey. International Journal of Computational Linguistics, 11(3), 45-67.

[3] Gupta, S., Verma, T. (2019). Challenges of Language Identification in Social Media Texts. Proceedings of the International Conference on Natural Language Processing, 78-85.

[4] Sharma, N., Reddy, K. (2022). Addressing the Under-Resourced Languages in NLP: A Case Study of Tulu and Kannada. Language Resources and Evaluation, 56(4), 1023-1045.

[5] Patel, R., Mehta, S. (2021). High-Resource vs. Low-Resource Languages: A Comparative Study in Language Processing. Journal of Linguistic Studies, 15(1), 67-89.

[6] Ramesh, K., Nair, A. (2023). CoLI-Kenglish and CoLI-Tunglish: Datasets for Code-Mixed Language Identification. Data in Brief, 45, 108-115.

[7] Singh, V., Kumar, R. (2022). Machine Learning and Deep Learning Approaches for Language Identification in Code-Mixed Texts. Journal of Machine Learning Research, 23(1), 1-25.

[8] Chaitanya, K., Kumar, A. (2020). Language identification in code-mixed Hindi-English text using machine learning techniques. Journal of Language and Linguistic Studies, 16(2), 123-135.

[9] Gundapu, S., Mamidi, R. (2021). Conditional Random Fields for language identification in Telugu-English code-mixed text. Proceedings of the International Conference on Natural Language Processing, 45-50.

[10] Mandal, M., Singh, R. (2021). A multichannel neural network approach for language identification in code-mixed text. Journal of Artificial Intelligence Research, 70, 345-367.

[11] Thara, K., Poornachandran, P. (2022). Dataset and transformer-based model for language identification in English-Malayalam code-mixed text. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 789-798.

[12] Veena, S., Ramesh, K. (2020). SVM models for language identification in Hindi-English code-mixed text. International Journal of Computational Linguistics, 11(3), 201-215.

[13] Gupta, A., Sharma, R. (2020). Language Identification in Code-Mixed Texts: A Machine Learning Approach. Proceedings of the International Conference on Computational Linguistics.

[14] Sharma, P., Ghosh, S. (2021). Hybrid Deep Learning Model for Language Identification in Code-Mixed Texts. Journal of Natural Language Engineering.

[15] Ramesh, K., et al. (2022). Enhancing Language Identification in Kannada-English Code-Mixed.

[16] Hegde, A., Balouchzahi, F., Butt, S., Coelho, S., G, K., Kumar, H. S., D, S., Hosahalli Lakshmaiah, H. S., Agrawal, A. (2024). Overview of CoLI-Dravidian: Word-level Code-mixed Language Identification in Dravidian Languages. In Forum for Information Retrieval Evaluation FIRE - 2024, Gandhinagar.

[17] Balouchzahi, F., Butt, S., Hegde, A., Ashraf, N., Hosahalli Lakshmaiah, H. S., Sidorov, G., Gelbukh, A. (2022). 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, IIIT Delhi, India.

[18] Lakshmaiah, H. S., Balouchzahi, F., Mudoor Devadas, A., Sidorov, G. (2022). CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts. Acta Polytechnica Hungarica.

[19] Hegde, A., Balouchzahi, F., Coelho, S., Lakshmaiah, H. S., Nayel, H. A., Butt, S. (2024). CoLI@FIRE2023: Findings of Word-level Language Identification in Code-mixed Tulu Text. In Proceedings of FIRE '23. Association for Computing Machinery.

[20] Hegde, A., Mudoor Devadas, A., Coelho, S., Lakshmaiah, H. S., Chakravarthi, B. R. (2022). 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.