Language Identification at the Word Level in Code-Mixed Texts Using Character Sequence and Word Embedding

O. E. Ojo[1, a], **A. Gelbukh**[1, b], **H. Calvo**[1, c], **A. Feldman**[2, d],
O. O. Adebanji[1, e], **J. Armenta-Segura**[1, f]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico
[2]Montclair State University, USA
{[a]olumideoea, [e]olaronke.oluwayemisi}@gmail.com, [d]feldmana@montclair.edu,
{[b]gelbukh,[c]hcalvo,[f]jarmentas2022}@cic.ipn.mx

Abstract

People often switch languages in conversations or written communication in order to communicate thoughts on social media platforms. The languages in texts of this type, also known as code-mixed texts, can be mixed at the sentence, word, or even sub-word level. In this paper, we address the problem of identifying language at the word level in code-mixed texts using a sequence of characters and word embedding. We feed machine learning and deep neural networks with a range of character-based and word-based text features as input. The data for this experiment was created by combining YouTube video comments from code-mixed Kannada and English (Kn-En) texts. The texts were pre-processed, split into words, and categorized as "Kannada", "English", "Mixed-Language", "Name", "Location", and "Other". The proposed techniques were able to learn from these features and were able to effectively identify the language of the words in the dataset. The proposed CK-Keras model with pre-trained Word2Vec embedding was our best-performing system, as it outperformed other methods when evaluated by the F1 scores.

1 Introduction

Language Identification (LI), the process of automatically recognizing the language(s) present in a given document, is a key component of several text processing pipelines. The main task in LI is text classification, which is the act of assigning text to categories represented by a finite number of labels. As social media facilitate rapid information exchange and generate large amounts of text, dealing with the many different languages in documents obtained from social media users is one of the challenges in natural language processing (NLP). Code mixing is the practice of using multiple languages at once, changing the lexical and grammatical aspects of informal communication, especially on social media. Social media users around the world often combine multiple languages to express their opinions online. NLP, a branch of artificial intelligence, plays an important role in understanding the language in which humans write and speak. In some circumstances, it is very difficult to identify languages in texts collected from social media, but computational methods can be applied to automatically recognize these languages. State-of-the-art methods (Balouchzahi and Shashirekha, 2020) (Hosahalli Lakshmaiah et al., 2022) (Ojo et al., 2021) in NLP tasks apply word embedding and n-gram-based models at the character or word level for different tasks, including LI.

There are several countries where different natural languages are spoken. Language diversity has a great impact on people in marriage, sports, business, medicine, and education. In India, officially known as the Republic of India, citizens have the ability to read, write, and speak various languages. The country is one of the largest by population, and citizens can communicate in a variety of languages including Kannada and English. Code mixing, which is the practice of switching between two or more languages, is widespread in multilingual societies like India. People frequently communicate in more than one language and not always in the language in which they are addressed. The multi-language texts of users in these multilingual societies are often difficult to understand and analyze. With several understudied low-resource languages, the challenge of LI in code-mixed text is far from being solved. With this motivation, our task was to develop and evaluate models that can correctly classify labels in code-mixed Kannada and English texts.

Machine learning algorithms (Hosahalli Lakshmaiah et al., 2022) (Sidorov et al., 2014) (Ojo et al., 2020) (Kolesnikova and Gelbukh, 2010), as well as deep learning algorithms (Balouchzahi et al., 2021) (Hoang et al., 2022) (Tonja et al., 2022) (Ojo et al., 2022), have been used in many sequence

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 1 - 6

labeling tasks in NLP. To identify language at the word level in Kannada-English code-mixed text, we proposed two machine learning techniques submitted to the "CoLI-Kanglish: Word-Level Language Identification in Code-mixed Kannada-English Texts" shared task (Balouchzahi et al., 2022), namely CK-Multiplex and CK-Keras. The developed models were applied to the CoLI-Kenglish dataset where the language of each word was detected and classified into a specified category.

The next section highlights the background and previous research on code mixing in social media texts. The dataset used to investigate the code mixing between English and Kannada is then introduced in Section 3. The methods applied to recognize word-level languages are discussed in Section 4. The language detection experimental results are shown in Section 5, and the conclusions and future work are presented in Section 6.

2 Background and Related Work

A number of machine learning and neural networks have been used to tackle and improve various NLP tasks, including the classification of code-mixed languages. Code mixing, according to (Muysken et al., 2000), refers to a situation in which words and grammar from two or more distinct languages are combined in a single sentence. In addition, code mixing is used while speaking two different languages at the same time. It suggests that all lexical and grammatical components indicate the act of switching languages, and that code mixing is most common in informal settings and occurs when the conversants use both languages concurrently.

(Shekhar et al., 2020) offered a method that was applied to the Facebook, Twitter, and WhatsApp dataset to identify the language of the text that has been mixed with Hindi and English. Certain sub-classes of the quantum LSTM network model have been shown to be able to accurately learn and predict language in a text on social media. The obtained results pave the way for further use of machine learning methods in quantum dynamics without relying on the precise form of the Hamiltonian.

To identify the language of Twitter data, (Ansari et al., 2021) conducted an extensive experiment using transfer learning and fine-tuning of BERT models. For language pre-training and word-level language classification, the study uses a data set consisting of code-mixed texts in Hindi, English, and Urdu. The findings demonstrate that pre-trained representations on code-mixed data perform better than their monolingual counterparts.

(Yasir et al., 2021) addresses the issue of mixed-script identification for a dataset that comprises Roman Urdu, Hindi, Saraiki, Bengali and English. RNN and word vectorization were used to train the language identification model. Furthermore, numerous model architectures were optimized, such as long short-term memory (LSTM), bidirectional LSTM, gated recurrent unit (GRU), and bidirectional gated recurrent unit (BGRU), and experimentation yielded a very good performance score. The study also looked at multilingual challenges including Roman words fused with English letters, generative spellings, and phonetic typing.

For a code-mixed text in English-Bodo-Assamese, (Kalita et al., 2021) was able to identify the language of the text at the word level. Several classification methods were applied to analyze and predict the language of text collected from Facebook. The n-gram and dictionary-based features were used to train the models on the code-mixed corpus and yielded different accuracies for the word-level language detection task.

For word-level language detection in code-mixed text, (Chittaranjan et al., 2014) developed a CRF-based system. Their method can be replicated on different languages since it takes advantage of lexical, contextual, character n-gram, and special character features. The experimental results show that the CRF-based technique performs consistently across language pairs when its performance is compared to other datasets.

To identify language boundaries at the word level, (Dutta, 2022) conducted a study using chat message datasets in mixed English-Bengali and English-Hindi languages. The author introduced a code-mixing index to evaluate the level of mixing in the corpora and evaluated the performance of the system to multiple languages.

(Jhamtani et al., 2014) proposed several techniques to learn the sequence of characters that are frequently swapped for others in standard transliterations. The authors demonstrated how these algorithms can do better than others in identifying Hindi words that correlate with the transliterated words supplied. Their distinctive experimental model for word-level language identification considers the language and part of speech of nearby words. The experimental findings indicate that the proposed

model performs better in terms of accuracy than the previous methods.

To help machine learning (ML) classifiers tackle the issue of offensive language identification (OLI) in code-mixed and multi-script texts, (Balouchzahi et al.) proposed the use of relevant features of syllables and character n-grams. Three pairs of Dravidian languages, Malayalam-English, Tamil-English, and Kannada-English, were used to evaluate the performance of the proposed models. Syllable and character n-gram features performed well for code-mixed and multi-script text analysis, as shown by the results of ML classifiers.

(Mandal and Singh, 2018) developed a unique architecture for code-mixed data language tagging that uses multichannel neural networks that mix CNN and LSTM for code-mixed data word level language identification. This architecture incorporates context information. The multichannel neural network performed well in the language identification task when used with a Bi-LSTM-CRF context capture module.

3 Dataset

The words in the CoLI-Kenglish dataset, provided by the shared task organizers, were written in Kannada, English, or a combination of the two languages and are classified into six main groups: "Kannada", "English", "Mixed-language", "Name", "Location", and "Other". The data was scraped from Kannada YouTube video comments and preprocessed according to (Hosahalli Lakshmaiah et al., 2022). The unstructured texts with incomplete sentences and shortened words were code-mixed between the two languages. Two native Kannada speakers carefully tagged 19,432 unique words extracted from more than 7,000 sentences to create the CoLI-Kenglish dataset. Table 1 contains a description of the data. The test dataset includes words of unknown language. This single-label classification only allows one language to be assigned to each word, and the languages can be either "Kannada" or "English" or "Mixed-language" or "Name" or "Location" or "Other". Table 2 shows the percentage of words in each category.

4 Methodology

In different text classification tasks, numerous algorithms have been proposed and yielded promising results. The models predicted the categories of the words in the vocabulary based on the feature responses received from the vector representation of the text. Data cleansing, word segmentation, and tokenization are typically the pre-processing steps applied to the raw input text and used to train the models. The text representation transmits the pre-processed text in the form of N-gram (Cavnar et al., 1994), Bag-Of-Words (BOW) (Zhang et al., 2010), Term Frequency-Inverse Document Frequency (TF-IDF) (Peng et al., 2014), and Word2Vec representations (Mikolov et al., 2013) that the models can understand while minimizing information loss.

Word2Vec model (Mikolov et al., 2013) generates word vectors for semantic meanings using local contexts. A word vector is a fixed-length real-value vector that is used to represent any word in the corpus. Word2Vec employs two critical models: CBOW and Skip-gram. The first way entails guessing the term that is being used on the assumption that its context is understood. When the word in use is known, the latter predicts the context. The Word2Vec training approach helps the system learn vector representations of words using the structure of the neural network. The proposed techniques implement systems based on well-researched methodologies such as character ngram and Word2Vec embedding.

4.1 CK-Multiplex

The initial model used for the text classification task by the CK-Multiplex is the Random Forest Classifier (RFC). Subsequently, the multilayer perceptron (MLP) classification model was used, which has shown positive results in terms of performance.

- **Random Forest Classifier (RFC)**

  The random forest classifier is one of the supervised learning algorithms that blends ensemble learning techniques with the decision tree architecture. It can manage large data sets and can automatically balance data sets when one class is more frequent than others. RFC does not require feature scaling since it employs a rule-based approach rather than distance calculation, and non-linear factors have no impact on its performance. It is extremely stable, robust to outliers, and has a lower noise impact.

- **Multi-Layer Perceptron (MLP)**

  The multi-layer perceptron (MLP) is a feed-forward neural network that learns the associ-

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

Table 1: Description of the CoLI-Kenglish dataset

Category	Tag	% of words
Kannada	kn	43.9%
English	en	30.1%
Mixed-language	kn-en	9.3%
Name	name	4.8%
Location	location	0.7%
Other	other	11.2%

Table 2: Percentage of words per category

ations between linear and non-linear data. It has one input layer with one node (or neuron) for each input, one output layer with one node for each output, and any number of hidden layers, each with any number of nodes. The multi-layer perception uses sigmoid activation functions at each node. What makes the MLP model so potent is its ability to learn the representation in the training data, as well as its capacity to learn any mapping function and being shown to be a universal approximation method.

Character n-grams were used as a very effective feature set in both the RFC and MLP models. Character n-grams can identify a word's morphological structure, in contrast to word n-grams, which can only recognize a word and its potential neighbors. Characters n-grams are much more effective in spotting patterns than word n-grams when identifying language in text. The results of the ngram model for each language category are obtained and recorded in Table 3.

4.2 CK-Keras

The system architecture to distinguish Kannada from English at word level is built on Long Short-Term Memory (LSTM) neural network and Word2Vec embedding. LSTM networks have been at the cutting edge of sequence-to-sequence learning (Chang and Lin, 2014) (Adebanji et al., 2022). Order dependency in sequence prediction tasks can be learned with LSTM neural networks that also contain internal states that can encode context input. The LSTM network architecture can handle text as a long word or character string and incorporates feedback loops to help keep information over time. LSTM can encode internal text structures such as word dependencies and is perfectly suited for language identification. It is used for various NLP tasks such as time series, machine translation, and many others. Words were trained in the embedding layer of the LSTM model with a sequence length of 30 and a batch size of 64 in CK-Keras and then transferred to the next level with the embedding layer. The length of the sequence defines the features of the dataset.

Model	Language	Prec.	Recall	F1
RFC	en	0.80	0.84	0.82
	en-kn	0.85	0.56	0.68
	kn	0.71	0.93	0.81
	location	1.00	0.07	0.12
	name	0.73	0.23	0.35
	other	0.65	0.20	0.31
MLP	en	0.76	0.78	0.77
	en-kn	0.72	0.68	0.70
	kn	0.78	0.79	0.79
	location	0.44	0.13	0.21
	name	0.44	0.36	0.39
	other	0.47	0.48	0.48

Table 3: Performance scores for the CK-Multiplex Model on the language categories

5 Results

After applying our language identification models to the dataset, we were able to classify the words in the test set according to the categories that the

Model	Features	W.A. F1-score	M.A. F1-score	Accuracy
RFC	Character n-grams	0.71	0.51	0.74
MLP	Character n-grams	0.71	0.54	0.72
LSTM	Word2Vec embedding	0.72	0.56	0.77

Table 4: Comparison of the F1 and accuracy scores of the CK-Multiplex and CK-Keras Models (W.A. - Weighted Average, M. A. - Macro Average)

models had learned from. The languages were identified using Random Forest Classifiers and Multi-Layer Perceptron baseline models and the results are encouraging. We also used LSTM's deep learning model to learn a better feature from the text. LSTMs were further trained using random initialized word embeddings. The systems were evaluated using accuracy, recall, and F1 scores, and the results obtained are shown in Tables 3 and 4.

6 Conclusion and Future Works

In this study, a preliminary investigation was conducted to determine the language used in code mixing during interaction on social media. The vocabulary and grammar of code-mixed texts are often adapted from multiple languages, and new structures are frequently developed based on the language and usage habits of its users. We tackle the problem of language identification at the word level in code-mixed social media text containing English and Kannada languages. We use a two-step classification approach for the word-level language identification task. The embedding of character, sub-word, and word-level information can assist in the learning of meaningful correlations in words from many different languages. The LSTM recurrent neural network and Word2Vec embedding approach achieved the highest F1 score among the proposed models. In the future, we plan to use more deep learning models and text from other languages. In order to extract useful information from code-mixed texts and make code-switching systems better understand reviews, comments, inquiries, sentiments, etc., it is necessary to adequately detect and process the language of these texts.

Acknowledgements

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20211884, 20220859, and 20220553, of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Olaronke Oluwayemisi Adebanji, Irina Gelbukh, Hiram Calvo, and Olumide Ebenezer Ojo. 2022. Sequential models for sentiment analysis: A comparative study. In *Mexican International Conference on Artificial Intelligence*, pages 227–235. Springer.

Mohd Zeeshan Ansari, MM Sufyan Beg, Tanvir Ahmad, Mohd Jazib Khan, and Ghazali Wasim. 2021. Language identification of hindi-english tweets using code-mixed bert. In *2021 IEEE 20th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 248–252. IEEE.

Fazlourrahman Balouchzahi, BK Aparna, and HL Shashirekha. 2021. Mucs@ dravidianlangtech-eacl2021: Cooli-code-mixing offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329.

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. In *19th International Conference on Natural Language Processing Proceedings*.

Fazlourrahman Balouchzahi and HL Shashirekha. 2020. Mucs@ dravidian-codemix-fire2020: Sacosentimentsanalysis for codemix text. In *FIRE (Working Notes)*, pages 495–502.

Fazlourrahman Balouchzahi, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov, and Alexander Gelbukh. A comparative study of syllables and character level n-grams for dravidian multi-script and code-mixed offensive language identification. *Journal of Intelligent & Fuzzy Systems*, (Preprint):1–11.

William B Cavnar, John M Trenkle, et al. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and*

information retrieval, volume 161175. Las Vegas, NV.

Joseph Chee Chang and Chu-Cheng Lin. 2014. Recurrent-neural-network for language detection on twitter code-switching corpus. *arXiv preprint arXiv:1412.4314*.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Aparna Dutta. 2022. Word-level language identification using subword embeddings for code-mixed bangla-english social media data. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 76–82.

Thang Ta Hoang, Olumide Ebenezer Ojo, Olaronke Oluwayemisi Adebanji, Hiram Calvo, and Alexander Gelbukh. 2022. The combination of bert and data oversampling for answer type prediction. In *CEUR Workshop Proceedings*, volume 3119. CEUR-WS.

Shashirekha Hosahalli Lakshmaiah, Fazlourrahman Balouchzahi, Anusha Mudoor Devadas, and Grigori Sidorov. 2022. CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts. *acta polytechnica hungarica*.

Harsh Jhamtani, Suleep Kumar Bhogi, and Vaskar Raychoudhury. 2014. Word-level language identification in bi-lingual code-switched texts. In *Proceedings of the 28th Pacific Asia Conference on language, information and computing*, pages 348–357.

Nayan Jyoti Kalita, Ankita Goyal Agarwala, and Jayprakash Das. 2021. Word level language identification on code-mixed english-bodo text. In *IOP Conference Series: Materials Science and Engineering*, volume 1020, page 012027. IOP Publishing.

Olga Kolesnikova and Alexander Gelbukh. 2010. Supervised machine learning for predicting the meaning of verb-noun combinations in spanish. In *Mexican International Conference on Artificial Intelligence*, pages 196–207. Springer.

Soumil Mandal and Anil Kumar Singh. 2018. Language identification in code-mixed data using multichannel neural networks and context capture. *arXiv preprint arXiv:1808.07118*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pieter Muysken et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

OE Ojo, A Gelbukh, H Calvo, and OO Adebanji. 2021. Performance study of n-grams in the analysis of sentiments. *Journal of the Nigerian Society of Physical Sciences*, pages 477–483.

Olumide E Ojo, Alexander Gelbukh, Hiram Calvo, Olaronke O Adebanji, and Grigori Sidorov. 2020. Sentiment detection in economics texts. In *Mexican International Conference on Artificial Intelligence*, pages 271–281. Springer.

Olumide Ebenezer Ojo, Thang Ta Hoang, Alexander Gelbukh, Hiram Calvo, Grigori Sidorov, and Olaronke Oluwayemisi Adebanji. 2022. Automatic hate speech detection using cnn model and word embedding. *Computación y Sistemas*, 26(2).

Tao Peng, Lu Liu, and Wanli Zuo. 2014. Pu text classification enhanced by term frequency–inverse document frequency-improved weighting. *Concurrency and computation: practice and experience*, 26(3):728–741.

Shashi Shekhar, Dilip Kumar Sharma, and MM Sufyan Beg. 2020. Language identification framework in code-mixed social media text based on quantum lstm—the word belongs to which language? *Modern Physics Letters B*, 34(06):2050086.

Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernández. 2014. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications*, 41(3):853–860.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Muhammad Arif, Abdul Gafar Manuel Meque, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2022. Cic nlp at smm4h 2022: a bert-based approach for classification of social media forum posts. In *Mining for Health Applications, Workshop & Shared Task (# SMM4H 2022)*, page 58.

Muhammad Yasir, Li Chen, Amna Khatoon, Muhammad Amir Malik, and Fazeel Abid. 2021. Mixed script identification using automated dnn hyperparameter optimization. *Computational intelligence and neuroscience*, 2021.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics*, 1(1):43–52.