Word Level Language Identification in Code-mixed Kannada-English Texts using traditional machine learning algorithms

M. Shahiki Tash, Z. Ahani, A.L. Tonja, M. Gemeda, N. Hussain and **O. Kolesnikova**
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
Corresponding: moein.tash@gmail.com

Abstract

Language Identification at the Word Level in Kannada-English Texts. This paper describes the system paper of CoLI-Kanglish 2022 shared task. The goal of this task is to identify the different languages used in CoLI-Kanglish 2022. This dataset is distributed into different categories including Kannada, English, Mixed-Language, Location, Name, and Others. This Code-Mix was compiled by CoLI-Kanglish 2022 organizers from posts on social media. We use two classification techniques, KNN and SVM, and achieve an F1-score of 0.58 and place third out of nine competitors.

1 Introduction

Nowadays, it is impossible to find somebody who doesn't use social media or smartphones. This prompts us to identify a new difficulty for people who have used social media. The following difficulties are just one example of the many tasks are performed in Natural Language Processing (NLP) to address various issues for instance, fake news(Arif et al., 2022), machine translation detection, sentiment analysis and language identification(Yigezu et al., 2021).

Identification language for mixed languages is the major challenge. Many users want an easy way to construct sentences, or employ habitual expressions. They try to write in a combination of two or three different languages, which leads the creation of Code-Mix data(Balouchzahi et al., 2022b). User-generated content like web articles, tweets, and message boards frequently contain code-mix text, yet majority of the language ID models in use today have been ignored. As observed in these English-Hindi examples, code-mixing entails language changes inside and across constituents.

[NP aapki profile photo] [V P pyari hai]

Your profile photo is lovely

In many areas, such as those where Hindi and English speakers coexist, code-mixing is the norm. As many as 17% of Facebook posts from India are code-mixed (Bali et al., 2014) and 3.5% from tweets (Rijhwani et al., 2017).

Nearly all social media networks where people speak several languages are Code-Mix. For instance, in a nation like India, where there are more than a dozen different languages with various alphabets, you may easily locate a Mix-Code of English and Indian languages if you check the posts on Facebook or Twitter that are linked to garments or related to shopping (Balouchzahi et al., 2022a). Because of their extensive range, code mixes cannot be adequately described in a finite number of words. Code-Mixing may contain a variety of words including words that combine the alphabets of two languages that identify an area, a person or a place, and different situations.

We will now discuss the classification system we utilized in this paper and also TF-IDF vectorizer.One of the effective supervised machine learning techniques that we may utilize for both regression and data classification is called the Support Vector Machine (SVM). Finding the hyperplane in an N-dimensional space that clearly classifies the data points is the objective of an SVM. (Ekbal and Bandyopadhyay, 2008). This means that the decision boundry line between the data points that fall into a category and those that do not is drawn clearly by the algorithm. Almost all data that is encoded as a vector is suitable for this technique. If it create a good vector from our data, we can use SVM to find good results (Tonja et al., 2022). Although KNN can be used just like SVM for both classification and regression issues, it is the primary application in classification. This algorithm stores all the data and can classify a new data point based on similarities.

This method places the new instance into the column that is more comparable to the available categories and makes the assumption that the new data is linked to the available items (Nongmeika-

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 25 - 28
December 15-18, 2022 ©2022 Association for Computational Linguistics

pam et al., 2017). As it encounters new data, this algorithm simply stores the data set during training and then classifies it into a group that is roughly similar to the present data. The TF-IDF statistic gives keywords that can be used to identify or categorize particular documents by demonstrating the relevance of certain keywords to a given set of documents (Gautam and Kumar, 2013).

2 Task description and Datasets

Language Identification (LI) is the process of automatically recognizing the languages used in a text. Kannada is one of the Dravidian languages that make up India's rich linguistic legacy and is used as the official language of the state of Karnataka. Karnataka residents can read, write, and speak Kannada, yet many find it challenging to use the language while posting messages or comments on social media.

Language identification is the process of automatically recognizing the languages used in a given text because code-mixing is one of the most challenging subjects in Natural Language Processing (NLP). The goal of the current investigation is to identify the language of the words.

As part of this work, we must determine which words are of English, Kannada, and mixed languages. The CoLI-Kenglish dataset consists of Kannada and English words written in Roman script and is divided into six main categories: "Kannada," "English," "Mixed-language," "Name," "Location," and "Other." Participants are asked to submit their methods in the Kanglish shared task, which requires that each word be recognized and categorized in one of these categories (Hosahalli Lakshmaiah et al., 2022).

3 Related Work

Language identification in social media texts is difficult because of things like social media content that has been code-mixed, using one alphabet to write in two languages at this point. Chakravarthi et al. (2021) proposed a code that combines Dravidian data in Kannada, Malayalam, and Tamil. Bohra et al. (2018) extend a Twitter data collection that include Hinglish data. They provided primary experiment findings with an accuracy of 0.71 using classifiers Support Vector Machine (SVM) and Random Forest (RF) with n-grams and lexicon-based features (Chakravarthi et al., 2020b,a). Sentiment Analysis for Dravidian Languages in Code-Mixed Data was a shared task in Dravidian-Code-Mix-FIRE2020 established by Kanwar et al. (2020). Researchers had submitted a variety of models, and they used the under sampling technique from Tomek (1976) to train some machine learning classifiers with various syntax-based n-gram features. The linear regression classifier with word and char n-gram features produced positive results with average weighted F1-scores of 0.71 and 0.62.

4 Methods

In this study, we employed standard machine learning algorithms for language identification. For this task, we used two different classifiers, including (i) support vector machines and (ii) k-nearest neighbors. We also used N-gram TF-IDF word and character features for vectorization. On each of these classifiers and this vectorization, we make a comment. For this task, we submit 5 runs, and the outcome varies each time.

4.1 Feature Engineering:

For this model, we used TF-IDF Vectorizer from the Sklearn module to extract char n-grams in the range of distinct pre-processed text data that are ready as word frames (1, 2). In Table 1 we lists the quantity of tasks, test sets, data sets, and category and tag values.

Table 1: Code-mixing language categories with test- and training-set counts

Task	Category	Tag	Number of Test-set	Number of Train-set
Task1	Kannada	kn	4585	14847
	English	en		
	Mixed-language	Kn-en		
	Name	name		
	Location	location		
	Other	other		

4.2 Model Construction:

These two classifiers were employed for the training set of data. For this challenge, we had 5 Runs. K-nearest neighbors (KNN) was employed for the first four runs. In order to vectorization, we just employed TF-IDF while using alternative parameters for support vector machines. Table 2 contains a list of all the parameters we utilized for each Run.

Table 2: Parameters that used in KNN,SVM,TF-IDF

Name of classifier/ vectorizer	**Parameter1**	**Parameter2**	**Parameter3**	**Parameter4**
KNN	n_neighbors=6	metric= ’manhattan’	p=2	weights= ’distance’
SVM	C=1.0	kernel=’poly’	degree=3	gamma= ’scale’
TF-IDF	analyzer=’char_wb’	ngram_range=(1,2)	min_df=0	norm= ’l1’

5 Experiments and Results

We demonstrate our experiment with text data that was gathered from YouTube. Each language pair’s word should be categorized into one of the six groups shown in Table 1. The suggested method w e used 14847 data for training and 4585 data for testing and we applied The purpose of the weighted average F1-score is assessment. We have displayed the number of errors made by the KNN algorithm during four runs in Table 3. Additionally, Table 4 displays the number of mistakes produced by the SVM algorithm in a single run. It is important to note that TF-IDF is used by both algorithms. As seen in Table 3, the weighted average F1-score increased and was able to rise in each run by modifying the KNN’s parameters.

Table 3: Results with using KNN classifier

	Weighted			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
RUN1	0.78	0.79	0.77	0.63	0.43	0.47
RUN2	0.8	0.8	0.79	0.61	0.5	0.53
RUN3	0.83	0.83	0.83	0.65	0.53	0.56
RUN4	0.83	0.83	0.83	0.64	0.56	0.58

Table 4: Results with using SVM classifier

	Weighted			Macro		
	Precision	Recall	F1-score	Precision	Recall	F1-score
RUN5	0.78	0.79	0.77	0.63	0.43	0.47

6 Conclusion

This study shows how different languages may be identified in code-mix data using a classifier that uses two algorithms, KNN and SVM. The first technique produces better results, with the best weighted average F1-score 0.58.

Acknowledgement

The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852, 20220859, and 20221627 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

Muhammad Arif, Atnafu Lambebo Tonja, Iqra Ameer, Olga Kolesnikova, Alexander Gelbukh, Grigori Sidorov, and AG Meque. 2022. Cic at checkthat! 2022: multi-class and cross-lingual fake news detection. *Working Notes of CLEF*.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. “i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

Fazlourrahman Balouchzahi, Sabur Butt, Asha Hagde, Noman Ashraf, Shashirekha Hosahalli Lakshmaiah, Grigori Sidorov, and Alexander Gelbukh. 2022a. 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, Sabur Butt, Grigori Sidorov, and Alexander Gelbukh. 2022b. CIC@LT-EDI-ACL2022: Are transformers the only hope? hope speech detection for Spanish and English comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 206–211, Dublin, Ireland. Association for Computational Linguistics.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020a. A sentiment analysis dataset for code-mixed malayalam-english. *arXiv preprint arXiv:2006.00210*.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John P. McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on*

Speech and Language Technologies for Dravidian Languages, pages 133–145, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020b. Overview of the track on sentiment analysis for dravidian languages in code-mixed text. In *Forum for information retrieval evaluation*, pages 21–24.

Asif Ekbal and Sivaji Bandyopadhyay. 2008. Bengali named entity recognition using support vector machine. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Jyoti Gautam and Ela Kumar. 2013. An integrated and improved approach to terms weighting in text classification. *International Journal of Computer Science Issues (IJCSI)*, 10(1):310.

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*.

Nikita Kanwar, Megha Agarwal, and Rajesh Kumar Mundotiya. 2020. Pits@ dravidian-codemix-fire2020: Traditional approach to noisy code-mixed sentiment analysis. In *FIRE (Working Notes)*, pages 541–547.

Kishorjit Nongmeikapam, Wahengbam Kumar, and Mithlesh Prasad Singh. 2017. Exploring an efficient handwritten Manipuri meetei-mayek character recognition using gradient feature extractor and cosine distance based multiclass k-nearest neighbor classifier. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 328–337, Kolkata, India. NLP Association of India.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1971–1982.

Ivan Tomek. 1976. Two modifications of cnn.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Mohammed Arif Khan, 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 *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 58–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94.