ICON 2022

Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts

Proceedings of the Shared Task

December 15-18, 2022

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-38-8

Introduction

South Asia is the world's most linguistically diverse region, with over 650 languages. India, in particular, is a multilingual country with a rich language heritage that includes the Dravidian language Kannada. Kannada is the official and administrative language of the state of Karnataka, and has over 40 million native speakers. Many people in this region are comfortable using both English and their native language in daily communication. On social media platforms, multilingual speakers often use code-mixing, which is the mixing of multiple languages and scripts in a single piece of text. Code-mixing can occur at the paragraph, sentence, word, or even sub-word level. However, using non-Roman scripts like Kannada on social media can be difficult, as most keyboard layouts and keypads use the Roman alphabet. As a result, many people prefer to use the Roman script for their social media posts. This poses challenges for natural language processing tasks such as sentiment analysis and emotion detection. In this article, we propose a model for identifying the language of code-mixed text on social media. We focus on Kannada-English code-mixing, and use a combination of deep learning and traditional machine learning techniques to achieve high accuracy in our model.

To address the challenges of code-mixed text in the context of the Kannada-English language pair, we conducted a shared task for identifying the language of code-mixed text on social media. In particular, we have open-sourced a Kannada-English code-mixed dataset for word level language identification of Kannada, English, and mixed-language words written in the Roman script. The task includes classifying each word in the given text into one of six predefined categories: Kannada, English, Kannada-English, Name, Location, and Other. Among the models submitted by participants, the best performing model obtained averaged-weighted and averaged-macro F1 scores of 0.86 and 0.62, respectively.

The results of the shared task reveal the difficulty of the language identification task in code-mixed text. This difficulty is mainly due to the nature of code-mixed texts that do not follow the rules and grammar of any language. This task aims to attract the attention of researchers for word level language identification of different language pairs in code-mixed text.

Organizing Committee

Program Committee Chairs

Fazlourrahman Balouchzahi, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Sabur Butt, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Asha Hegde, Mangalore University, India
Noman Ashraf, Dana-Farber Cancer Institute, Harvard Medical School, United States
Shashirekha Hosahalli Lakshmaiah, Mangalore University, India
Grigori Sidorov, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Alexander Gelbukh, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico

Program Committee

Program Committee

Fazlourrahman Balouchzahi, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Sabur Butt, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Asha Hegde, Mangalore University, India
Noman Ashraf, Dana-Farber Cancer Institute, Harvard Medical School, United States
Shashirekha Hosahalli Lakshmaiah, Mangalore University, India
Grigori Sidorov, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico
Alexander Gelbukh, Instituto Politécnico Nacional (IPN), Center for Computing Research (CIC), Mexico

Table of Contents