The 2nd Faculty Annual Research Session
Faculty of Applied Science
University of Vavuniya, Sri Lanka
September 15, 2021

A sentiment analysis dataset for code-mixed Sinhala-English language

Uthpala, K.
Department of Physical Science
University of Vavuniya
2015asp74@vau.jfn.ac.lk

Thirukumaran, S.
Department of Physical Science
University of Vavuniya
thiruvks@gmail.com

ABSTRACT

Communication is an essential part of human life, and technology has been involved in several ways much wider than ever before. Single language often does not reflect on how people communicate naturally. Many parts of the world are multilingual on a societal level because most people using two or more languages. In multilingual settings, people tend to mix words from all languages they use, which is introduced as code-mixing. Moreover, there is an increasing demand for sentiment analysis of text from social media, mostly code-mixed. It can be helpful in various decision-making processes. Systems trained for monolingual data fail for code-mixed data due to the complexity of mixing at different levels. However, available resources for code mix data to create a model are very few. Only a few datasets are available for some popular languages such as Hindi-English, and Spanish-English. There are no resources available for Sinhala-English code-mixed language, and still, researchers have not given their attention to sentiment analysis for Sinhala-English mixed language. To overcome this, we presented a Sinhala-English, sentiment-labelled corpus using comments from YouTube videos for sentiment analysis of code-mixed text in the Sinhala-English language. Comments from social media do not follow strict grammar rules, and those are very noisy; we preprocessed to clean the comments. Then we use an annotation setup to label and create Sinhala-English dataset with 7,832 comments for sentiment analysis. At least two annotators annotate each comment, and altogether nine voluntary annotators contribute to the annotation process. The whole data have been categorized into three classes; positive, negative, and neutral. We use five machine learning algorithms on the newly created Sinhala-English code-mixed language dataset to show the insight of the dataset. All the used classification algorithms achieved significant accuracy for the created Sinhala-English dataset. The logistic regression algorithm showed the higher macro average score for precision, recall, and F1 score for the presented dataset.

Keywords: Code-mixed language, Sentiment analysis, Sinhala-English.