Language identification at word level in Sinhala-English code-mixed social media text

Kasthuri Shanmugalingam
Department of Computational Mathematics
University of Moratuwa, Sri Lanka
s.shanshiya@gmail.com

Sagara Sumathipala
Department of Computational Mathematics
University of Moratuwa, Sri Lanka
sagaras@uom.lk

Abstract

Automatic analyzing and extracting useful information from the noisy social media content are currently getting attention from the research community. It is common to find people easily mixing their native language along with the English language to express their thoughts in social media, using Unicode characters or the Unicode characters written in Roman Scripts. Thus these types of noisy code-mixed text are characterized by a high percentage of spelling mistakes with phonetic typing, wordplay, creative spelling, abbreviations, Meta tags, and so on. Identification of languages at word level become a necessary part for analyzing the noisy content in social media. It would be used as an intimidate language identifier for chatbot application by using the native languages. For this study we used Sinhala-English code-mixed text from social media. Natural Language Processing (NLP) and Machine Learning (ML) technologies are used to identify the language tags at the word level. A novel approach proposed for this system implemented is machine learning classifier based on features such as Sinhala Unicode characters written in Roman scripts, dictionaries, and term frequency. Different machine learning classifiers such as Support Vector Machines (SVM), Naive Bayes, Logistic Regression, Random Forest and Decision Trees were used in the evaluation process. Among them, the highest accuracy of 90.5% was obtained when using Random Forest classifier.

Keywords: *Code-mixing, Language identification, Machine learning, Natural Language Processing (NLP)*