SENTIMENTAL ANALYSIS OF COMMENTS IN SOCIAL MEDIA IN SINHALA - ENGLISH CODE-MIXED LANGUAGE USING SUPERVISED LEARNING TECHNIQUES

P. M. I. U Aththanayaka*, H. M. M. Naleer

[1]*Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka, Sammanthurai*

imesh.udaya@gmail.com

Abstract

Sinhala is a morphologically rich but low resourced language for computer-based natural language processing. Since the introduction of Unicode character set for Sinhala, considerable growth of Sinhala textual web contents can be observed. With the rapid popularity of social media in sri Lanka this growth can be also seen in the social media contents also. Social media comments are frequently code-mixed in Sinhala and English and consists of Singlish terms (Sinhala words written in Roman Script). Therefore, performing sentiment analysis on such document considering only Sinhala would be inaccurate since there may be content in English or Singlish which may contribute to the sentimental value of the content. To overcome this challenge a model will be built through this study using social media comments which will be able to identify the correct language of the terms and perform sentiment analysis considering the whole content. In this study, using YouTube platform 500 code-mixed comments were extracted and they were labeled manually as Positive, Negative or Neutral. After preprocessing steps such as emoji removal, stop word, URL and Special symbol removal, comments were tokenized separately based on their character set, Roman scripts were tokenized into separate list to identify the Singlish words. Roman script token list was stemmed and lemmatized using Natural Language Toolkit (NLTK) library and compared with an English word list to recognize English words, rest of the words are considered Sinhala and transliteration is performed to Sinhala script and Singlish to Sinhala dictionary is created. Singlish words on comments were replaced using the dictionary created. Sinhala words were stemmed based on shallow learning method and English words using NLTK library. After the prepossessing and transliteration stage feature extraction is performed using various techniques and Supervised Machine Learning method such as Random forest, Support vector machine and Multinomial Naïve Bayes were used for classification the Sentiment Analysis. In the Proposed methodology Transliteration was accurate up to 72% and Random Forest classifier gave highest accuracy which is 75%.

Keywords: Sinhala Sentiment Analysis, Code-mixed comments, Social media, transliteration