Bilingual sentiment analysis on Malaysian social media using vader and normalisation heuristics

ABSTRACT

This research addresses a number of important issues involved in performing Sentiment Analysis (SA) on Malaysian Social Media (SM), including an analysis of bilingual or mixed language, choice of sentiment lexicon, normalisation heuristics, and the use of public datasets. This work is the first to quantify the level of language mixing in informal Malaysian text. Analysis of the 2M tweet Malaya dataset revealed a significant level of English sentiment content in Malaysian social media (13.5%), demonstrating the neccessity of a bilingual approach to Malaysian Sentiment Analysis. Significant patterns in noisy Malaysian SM text were identified and heuristics for normalising them were devised. The popular and effective English lexicon-based SA system VADER (Valence Aware Dictionary and sEntiment Reasoner) was translated to Malay using automatic and manual methods, with the combination of English and Malay VADER yielding a bilingual SA system. A subset of the Malaya dataset was both corrected and extended from two to three classes in order to properly test the bilingual SA system. Bilingual VADER with normalisation heuristics was able to achieve an impressive level of performance on a three-class problem (accuracy=0.71, mean F1=0.72), as compared to Malay VADER alone and several popular machine learning-based algorithms.