Machine Learning Approach for Sentiment Polarity Detection in Different Languages

Jelena Graovac

University of Belgrade, Faculty of Mathematics, Belgrade, Serbia

e-mail: jgraovac@matf.bg.ac.rs

Abstract

Sentiment Polarity Detection (SPD) is a challenging task that combines Natural Language Processing (NLP) and text mining techniques to automatically classify text documents into "positive" and "negative" categories regarding sentiment orientation. The proposed technique is based on the byte-level n-gram frequency statistics method for text representation, and Support Vector Machine (SVM) - Machine Learning (ML) algorithm for categorization process. It does not require any morphological analysis of texts, any preprocessing steps, or any prior information about document content or language. We avoid the necessity for use of taggers, parsers, feature selection, or other language-dependent and non-trivial NLP tools. Proposed approach fully relies on the power of ML algorithm based on strong mathematical foundations. For driving experiments we used seven publicly available movie review benchmarks in English, Spanish, Arabic, French, Turkish, Czech languages and Serbian. Despite their simplicity and broad applicability, experimental results confirm that the presented technique is comparable with the best ranked previously published techniques, when applied to movie reviews datasets.

Keywords: sentiment polarity detection; movie reviews; n-grams; SVM