A performance comparison of feature extraction methods for sentiment analysis

Abstract

Sentiment analysis is the task of classifying documents according to their sentiment polarity. Before classification of sentiment documents, plain text documents need to be transformed into workable data for the system. This step is known as feature extraction. Feature extraction produces text representations that are enriched with information in order to have better classification results. The experiment in this work aims to investigate the effects of applying different sets of features extracted and to discuss the behavior of the features in sentiment analysis. These features extraction methods include unigrams, bigrams, trigrams, Part-Of-Speech (POS) and Sentiwordnet methods. The unigrams, part-of-speech and Sentiwordnet features are word based features, whereas bigrams and trigrams are phrase-based features. From the results of the experiment obtained, phrase based features are more effective for sentiment analysis as the accuracies produced are much higher than word based features. This might be due to the fact that word based features disregards the sentence structure and sequence of original text and thus distorting the original meaning of the text. Bigrams and trigrams features retain some sequence of the sentences thus contributing to better representations of the text.