Text Pre-processing techniques for twitter sentiment analysis

Abstract :

Pre-processing is the first step in text classification, and choosing right pre-processing techniques can improve classification effectiveness.

We experimentally compare 16 commonly used pre-processing techniques on a Twitter dataset for Sentiment Analysis, employing four popular machine learning algorithms, namely, Linear SVC, Bernoulli Naïve Bayes, Logistic Regression, and Convolutional Neural Networks. We evaluate the pre-processing techniques on their resulting classification accuracy and number of features they produce.

We find that techniques like lemmatization, stemming, removing numbers, and replacing contractions, improve accuracy, replace Repetitions of Punctuation, Replace Negations with Antonyms, Remove Punctuation, Handling Capitalized Words, Lowercase, Replace Elongated Words.

We also find some statistics for the dataset like Total Unique words before and after pre-process, Total Elongated words, Total multi Exclamation, Total Slangs and Abbreviations found.