Kurzzusammenfassung

Mit zunehmender Popularität nutzergenerierter Inhalte im Internet steigt auch die Relevanz der automatischen Auswertung solcher Informationen. Die Sentiment-Polaritäts-Klassifikation, eine der Disziplinen der Sentiment-Analyse, analysiert und klassifiziert Dokumente anhand der in ihnen geäußerten Meinungen. Während sich die Forschung vorwiegend auf die Anwendung von Standardverfahren des maschinellen Lernens konzentriert, liegt der Fokus dieser Arbeit auf einem neuen Ansatz, welcher bereits erfolgreich auf andere Teildisziplinen der automatischen Textanalyse angewandt wurde. Hierbei kommen Sentiment-Klassifkatoren auf Basis statistischer Datenkompressionsverfahren und Statistiken über Buchstabensequenzen einer variablen bzw. festen Länge n – sogenannter character-n-grams – zum Einsatz.

Neben einem auf Prediction-by-Partial-Matching (PPM) basierenden Klassifikator wird auch das p^2-Measure vorgestellt. Dabei handelt es sich um ein neues, vom PPM-Algorithmus abstrahiertes informationstheoretisches Maß. Durch die Kombination mit Gewichtungsverfahren und Feature-Selektions-Metriken übertrifft das p^2-Measure durchweg die deutlich anspruchsvolleren, wortbasierten Support-Vector-Machines.

Im Verlauf dieser Arbeit wird das Potential des p^2-Measure und der character-n-gram-basierten Ansätze im Detail untersucht. Neben unterschiedlichen Quell- und Zieldomänen werden auch potentielle Vorteile auf fremdsprachlichen Datensätzen betrachtet. Zudem wird untersucht, in welchem Maße das p^2-Measure neben der Polarität auch die Stärke einer Meinung und darauf aufbauend die Bewertung eines Dokumentes vorhersagen kann. Die Ergebnisse dieser Arbeit zeigen, dass das p^2-Measure eine ernstzunehmende Alternative zu den wortbasierten Standardmethoden darstellt und insbesondere bei verrauschten oder fremdsprachlichen Daten Vorteile bieten kann.

Abstract

With growing availability and popularity of user-generated content, automatic analysis and aggregation of such information becomes increasingly important. Sentiment polarity classification, one of the main tasks in sentiment analysis, aims to analyze and classify documents according to opinions stated therein. Existing work has mainly focused on standard machine learning techniques. Below, we investigate a novel approach that has proven successful in conventional text classification tasks such as authorship attribution or topic categorization.

Our thesis examines classifiers based on adaptive statistical data compression models or more general based on statistics about variable or fixed length character sequences, i.e. character n-grams. We define a classifier using the prediction by partial matching (PPM) compression algorithm and introduce the p^2-Measure as a simple abstraction of PPM, motivated in information theory. By coupling the p^2-Measure with feature weighting and feature selection schemes, it consistently outperforms the far more sophisticated SVM.

In the course of this work, we analyze advantages of the p^2-Measure and character n-gram based approaches in detail. Besides the transfer performance between different source and target domains, namely cross-domain sentiment analysis, we are also interested in potential benefits of our method on foreign language datasets. Moreover, we will investigate to which extend the p^2-Measure can be used to determine not only the polarity but also the strength and even the original rating of a document. Altogether, our results show that the p^2-Measure is a serious alternative to the word-based standard approach and that it is especially suitable for noisy or foreign language datasets.