Corpus-based Construction of Sentiment Lexicon to Identify Extreme Opinions by Supervised and Unsupervised Machine learning Methods

Título Corpus-based Construction of Sentiment Lexicon to Identify Extreme Opinions by Supervised and Unsupervised Machine learning Methods

Autor/a Sattam Al Matarneh Mohammad

Directores Pablo Gamallo Otero

Tipo Tese doutoral

Data de lectura 30/11/2018

Lugar de lectura Universidade de Santiago de Compostela

Abstract

Studies in sentiment analysis and opinion mining focused on many aspects related to opin- ions, particularly polarity classification by making use of positive, negative or neutral values. However, most studies overlooked the identification of extreme opinions (very negative and very positive opinions) in spite of their vast significance in many applications. This doctoral thesis describes a strategy to build sentiment lexicons from corpora, namely lexicons lexicons adapted to extreme values. This strategy has been used to build some lexicons and to know its effectiveness in determining the polarity of opinions. First, we will construct a domain- specific lexicon from a corpus of movie reviews. Polarity words of the lexicon are assigned weights standing for different degrees of positiveness and negativeness. This lexicon is will be combined into a sentiment analysis system to evaluate its performance in the task of sentiment classification. Second, two lexicons will be built of extremely negative and positive words from labeled corpora. We will integrate the lexicons that have been built into classifiers, whether super- vised or unsupervised classifier. We will use a supervised classifier, more precisely, Support Vector Machine (SVM) with some linguistic features such as a bag of words, word embed- ding, polarity lexicons, and set of textual features, in order to identify extreme opinions and provide a comprehensive analysis of the relative importance of each set of features. We also will compare our lexicons with four well-known sentiment lexicons. For this purpose, an indirect evaluation is carried out. The lexicons will be integrated into supervised sentiment classifiers, and their performance is evaluated in two sentiment classification tasks to identify i) the most negative vs. not most negative opinions, and ii) the most positive vs. not most positive. Moreover, a set of textual features is integrated into the classifiers to analyze how these textual features improve the lexicon performance. On the other hand, we also tested the efficiency of our lexicons in determining extreme opinions through the use of unsupervised classifiers. Our classification algorithm is based on a fundamental word-matching scheme to carry out unsupervised sentiment analysis.

LIGAZÓNS

Teseo

DESCARGAS

Referencia BibTex

Descargar versión completa

PROGRAMAS CIENTÍFICOS

Tecnoloxías da Linguaxe Natural