Sentiment Analysis Theory and Applications

25. August 2018

1 Introduction

What is sentiment analysis

In this project the goal is to extract useful information out of news articels gathered from different news portals. One idea is to compare the sentiment of articels from different sources.

In general there are three types of approaches for sentiment analysis:

- Kowledge based techniques, ...
- Statistical methods, which use machine learning methods for fitting models which return the sentiment of a sentence or document
- Hybrid approaches

A problem of sentiment analyis is that often even humans do not agree over the sentiment of a text, so it is difficult to evaluate the performance of an algorithm.

2 Knowledge based techniques

3 Statistical methods

Machine learning methods

3.1 Binary SPAM classification by support vector machines

An exercise from the statistical learning lab module deals with SPAM filtering by using support vector machines. For that purpose, training data in the form of emails is given. Each mail is labeled as either SPAM or HAM (no SPAM).

First, feature encoding is performed. That means, transforming the training data in a way that it can be used by a classification algorithm. In this case, the "bag-of-words"model is used. Each word, that occurs in the training data set is represented as a binary variable. Each email text is encoded in a way that if a word occurs in the text, the corresponding variable has the value 1, otherwise it has the value 0. This way, the text is transformed into a set of binary variables.

On this transformed dataset, a support vector machine alorithm can be trained to predict wether a text is SPAM or HAM.

Disadvantages from my point of view:

- Needs labeled training data
- Needs clear distinction (SPAM or HAM). I don't know if that is the case in sentiment analysis

4 Hybrid approaches

Ontologies, semantic networks

5 Miscellaneous

Maybe first just use bag-of-words model, which disregards context, grammar and even word order. According to wikipedia, most sentiment classification approaches do that.

Maybe predict source from text.

5.1 Subjectivity/objectivity identification