

Master Project

Automated Bipolar Sentiment Text Classification in HR Domain

Peyman Mohajerin Esfahani, Delft Center for Systems and Control, TU Delft
P.MohajerinEsfahani@tudelft.nl

Marco Loog, Pattern Recognition Laboratory, TU Delft
M.Loog@tudelft.nl

Context

With the rapidly growing number of online reviews, comments and ratings on products and services, human analysis of these expressions becomes inconvenient and time consuming. Fast and broad analysis on these expressions demands generic automated approaches that can be found in machine learning. One of the attributes expressions possess that is of great interest, is the sentiment. The possible sentiments an expression can have, can be brought down to two classes; either positive or negative (bipolar sentiment classification). The automated sentiment classification of these textual opinion expressions yields possible applications in the field of recommender systems, business intelligence and comparison systems.

Several machine learning techniques have proven to be successful in classifying movie reviews, twitter messages, product reviews and hotel reviews. Popular approaches are Support Vector Machines and Nave Bayes techniques. As feature space (TF-IDF weighted) word counts are used with possible feature space extension such as n-grams and part of speech. Typical accuracy percentages of popular bipolar sentiment classifiers in these domains are in the range of 80%-90%.

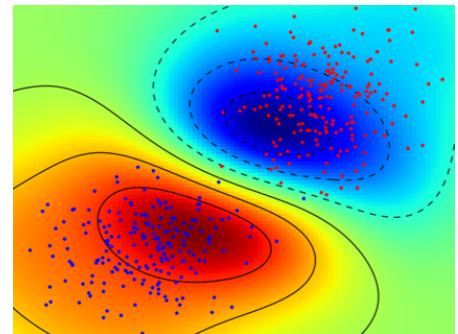


Figure 1: Classification problems in Machine Learning

Project tasks

This master thesis project is aimed at researching the possibilities of performing automated bipolar sentiment classification in a new domain; human resources (HR). In collaboration with the company Focus Orange, a company specialized in HR advisory and tools, the possibilities of classifiers in the HR domain will be researched. The classifier should correctly classify the opinion expressions of employees obtained from a company survey. The project ultimately tries to fully answer the following research questions:

1. How well do popular bipolar sentiment classifiers perform when trained and tested on textual expressions of employees within one company?
2. How well do these classifiers perform when trained on data from a set of companies and tested on another company (cross company performance)?
3. Can performance in the HR domain be improved by using data from other domains such as online user reviews?

⇒ **The project will be conducted at Focus Orange, a consulting company developing data-analytics tools to help clients improving their investment in human capital.**