# Chapter 77
# Semantics-Based Document Categorization Employing Semi-Supervised Learning

**Jan Žižka**
*Mendel University in Brno, Czech Republic*

**František Dařena**
*Mendel University in Brno, Czech Republic*

## ABSTRACT

*The automated categorization of unstructured textual documents according to their semantic contents plays important role particularly linked with the ever growing volume of such data originating from the Internet. Having a sufficient number of labeled examples, a suitable supervised machine learning-based classifier can be trained. When no labeling is available, an unsupervised learning method can be applied, however, the missing label information often leads to worse classification results. This chapter demonstrates a method based on semi-supervised learning when a smallish set of manually labeled examples improves the categorization process in comparison with clustering, and the results are comparable with the supervised learning output. For the illustration, a real-world dataset coming from the Internet is used as the input of the supervised, unsupervised, and semi-supervised learning. The results are shown for different number of the starting labeled samples used as "seeds" to automatically label the remaining volume of unlabeled items.*

## INTRODUCTION

Let us imagine a common problem: Having a small set of textual documents as samples of certain semantic categories, where each document is correctly labeled by its category, it is necessary to categorize a very big mountain of remaining unlabeled documents where their category is not known but can be determined by the semantic contents. It is possible by reading those documents; however, it can take a very long time and expenses. Today's possibility is to employ machines, computers, which could do it

DOI: 10.4018/978-1-5225-1759-7.ch077

for us via *machine learning*. If there is a sufficiently high number of labeled samples, a classification algorithm can be inductively trained using the specific labeled samples and then applied to labeling the unlabeled documents – a procedure known as *supervised learning* where a supervisor (teacher) is a process that monitors the training process from the minimization of the classification error point of view. Lowering the error is achieved by gradual modification of particular parameters of the selected algorithm.

When no training samples are available, *unsupervised learning* (clustering) can be applied but the missing information – provided by the labeled samples during the supervised learning – mostly later leads to somehow inferior classification results because that absent feedback between the teacher and learner influences the training process negatively. The missing labels might be compensated by manual labeling but it could be unacceptable due to a very high necessary effort, which is inevitable for large data.

As a certain kind of trade-off, the above mentioned possibility of having a small set of training samples as the starting point looks appealingly because the limited labeling can be performed manually. Then, using it for the right aiming, that "seed" training set can be applied to labeling of the uncategorized documents which may be gradually added to the training set. The classifier can be repeatedly retrained as the training set size is growing and the expected result may be better than in the case of the unsupervised learning because more training information is progressively available. Such an approach is known as *semi-supervised learning*. Naturally, the quality of the semi-supervised learning depends on the ability to correctly label the samples that are supposed to strengthen the training set, which is usually given by a specific application, its data, and the method of the labeling correctness evaluation.

Using real-world data, this chapter's goal is to demonstrate how the semi-supervised learning can work. In addition, the results of the semi-supervised approach are compared with outputs of the unsupervised as well as supervised learning employing the same data. The substance of experiments aimed at the classification of textual documents from their semantic point of view, which was satisfaction or dissatisfaction with hotel services. People who used the hotel services could then express their opinion by writing a not too long review using a www portal with the help of the Internet and a browser in their computer (PC). The opinions were not placed at hotel web-pages; they were published by an agency that enables on-line booking of accommodation. Such opinions can later serve for other potential customers as well as directly for the service providers. The reviews were written freely, with no requested structure, using any natural language, and for the experiments described in the following sections, those textual items were divided into two categories: positive and negative ones. As it is shown, all three training methods provided positive results in accordance with the information that was available either by no, or limited, or full labeling.

## TEXT MINING USING MACHINE LEARNING APPROACH

In accordance with this chapter pointing, and without any exact definition, the concept *text mining* is generally comprehended as a specialized branch of *data mining*. Data mining is the computational process of discovering knowledge in large data sets of any type involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Text mining area focuses on revealing knowledge in large *text data*, namely analyzing text in natural languages, which are (or, for some old languages, were) used by human beings. People use their spoken and written natural languages for communicating pieces of knowledge or information to each other. The current technology enables