Chapter 6

A SURVEY OF TEXT CLASSIFICATION ALGORITHMS

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY
charu@us.ibm.com

ChengXiang Zhai
University of Illinois at Urbana-Champaign
Urbana, IL
czhai@cs.uiuc.edu

Abstract
The problem of classification has been widely studied in the data mining, machine learning, database, and information retrieval communities with applications in a number of diverse domains, such as target marketing, medical diagnosis, news group filtering, and document organization. In this paper we will provide a survey of a wide variety of text classification algorithms.

Keywords: Text Classification

1. Introduction
The problem of classification has been widely studied in the database, data mining, and information retrieval communities. The problem of classification is defined as follows. We have a set of training records \( \mathcal{D} = \{X_1, \ldots, X_N\} \), such that each record is labeled with a class value drawn from a set of \( k \) different discrete values indexed by \( \{1 \ldots k\} \). The training data is used in order to construct a classification model, which relates the features in the underlying record to one of the class labels. For a given test instance for which the class is unknown, the training model
is used to predict a class label for this instance. In the *hard version* of the classification problem, a particular label is explicitly assigned to the instance, whereas in the *soft version* of the classification problem, a probability value is assigned to the test instance. Other variations of the classification problem allow ranking of different class choices for a test instance, or allow the assignment of multiple labels [52] to a test instance.

The classification problem assumes categorical values for the labels, though it is also possible to use continuous values as labels. The latter is referred to as the regression modeling problem. The problem of text classification is closely related to that of classification of records with set-valued features [28]; however, this model assumes that only information about the presence or absence of words is used in a document. In reality, the frequency of words also plays a helpful role in the classification process, and the typical domain-size of text data (the entire lexicon size) is much greater than a typical set-valued classification problem. A broad survey of a wide variety of classification methods may be found in [42, 62], and a survey which is specific to the text domain may be found in [111]. A relative evaluation of different kinds of text classification methods may be found in [132]. A number of the techniques discussed in this chapter have also been converted into software and are publicly available through multiple toolkits such as the *BOW* toolkit [93], Mallot [96], WEKA ¹, and LingPipe ².

The problem of text classification finds applications in a wide variety of domains in text mining. Some examples of domains in which text classification is commonly used are as follows:

- **News filtering and Organization:** Most of the news services today are electronic in nature in which a large volume of news articles are created very single day by the organizations. In such cases, it is difficult to organize the news articles manually. Therefore, automated methods can be very useful for news categorization in a variety of web portals [78]. This application is also referred to as *text filtering*.

- **Document Organization and Retrieval:** The above application is generally useful for many applications beyond news filtering and organization. A variety of supervised methods may be used for document organization in many domains. These include large digital libraries of documents, web collections, scientific literature,

²[http://alias-i.com/lingpipe/]