Chapter 10
Automatic Categorization of Reviews and Opinions of Internet E-Shopping Customers

Jan Žižka
Mendel University in Brno, Czech Republic

Vadim Rukavitsyn
Mendel University in Brno, Czech Republic

ABSTRACT

E-shopping customers, blog authors, reviewers, and other web contributors can express their opinions of a purchased item, film, book, and so forth. Typically, various opinions are centered around one topic (e.g., a commodity, film, etc.). From the Business Intelligence viewpoint, such entries are very valuable; however, they are difficult to automatically process because they are in a natural language. Human beings can distinguish the various opinions. Because of the very large data volumes, could a machine do the same? The suggested method uses the machine-learning (ML) based approach to this classification problem, demonstrating via real-world data that a machine can learn from examples relatively well. The classification accuracy is better than 70%; it is not perfect because of typical problems associated with processing unstructured textual items in natural languages. The data characteristics and experimental results are shown.

DOI: 10.4018/978-1-4666-1861-9.ch010

INTRODUCTION

Business intelligence (BI) is a specialization connected with technological applications to the automatic support of making decisions and the competition ability in economics. BI deals with the online analytical processing (OLAP), predictions, and data analyses, taking knowledge and information from data. Today, it is typical that there are big volumes of very different data which usually contain very useful, however, hidden information. Revealing this information by traditional analytical and modeling methods is sometimes very difficult not only because of big volumes of data but also, for example, because of non-homogenous nature of this data. Information in databases could be general or special: numerical, nominal, binary, acoustic, image, video, including text in natural languages (Abney, 2008). This heterogeneous nature introduces certain problems, for example, in creating mathematical models which is often impossible without excessive simplifications. As a good example, a reader can imagine textual comments of clients about some commodities, their purchasing and selling, and so like. Human experts can get information from natural-language data connected with solving a special problem and finding in the data certain rules or other forms of the previously hidden knowledge: data mining (Berry & Linoff, 2004). Modern computer science provides an array of algorithms in the area of artificial intelligence, for example, *machine learning* (Alpaydin, 2010; Mitchell, 1997). One possibility is using a collection of text databases as labeled training samples to teach a machine what and how to do in a specific situation when new unknown but more-or-less similar data items appear in the future: classification or prediction (Hastie, Tibshirani & Friedman, 2009; Srivastava & Sahami, 2009). This method is called *supervised learning* (Vapnik, 2000). Having knowledge obtained by training a machine can suggest a solution based on a certain similarity to one or more generalized cases known from the past times (Theodoridis & Konstantinos, 2009). Such a machine learning-based approach tries to emulate behavior of human experts (Rukavitsyn & Žižka, 2010).

DATA DESCRIPTION

To investigate possibilities of automatic data-mining from customer comments that are written in a quite free, unstructured form using natural language (Berry & Kogan, 2010; Konchady, 2006), the authors collected some publicly accessible textual data from the Internet web-site *amazon.com*. The main intention was to get comments about various consumer goods with at least 100 different opinions per each goods item provided by purchasers. The customer reviews describe their experiences that are good, bad, or something between. It is possible to apply also a certain scale as a kind of classification, or *rating*: from one star (the worst experience) up to five stars (the best one). The reviews are expected to explain reasons of their ratings which are usually relatively short, tens or hundreds of words. Typically, the language is English, however, with many mistypings, grammar errors, and so forth. In addition, the used English is really very "international", and the customers are not only people whose native language is one of existing English languages that can more or less differ in grammar and vocabulary. Also, a reader of reviews can sometimes see non-standard interjections and onomatopoeic words.

The nine different commodities the reviews of which were used in the research are shown in Table 1. Interestingly, the average customer rating is very typically closer to five stars which means that customers were probably mostly satisfied.

For the experiments described further, the data were prepared using a simple, standard approach (Sebastiani, 2002). For each data-set, all its words created a corresponding dictionary. The dictionary did not include numbers, punctuation, and special