11 Opinion Mining and Sentiment Analysis

In Chap. 9, we studied the extraction of structured data from Web pages. The Web also contains a huge amount of information in unstructured texts. Analyzing these texts is of great importance as well and perhaps even more important than extracting structured data because of the sheer volume of valuable information of almost any imaginable type contained in text. In this chapter, we only focus on mining opinions which indicate positive or negative sentiments. The task is technically challenging and practically very useful. For example, businesses always want to find public or consumer opinions about their products and services. Potential customers also want to know the opinions of existing users before they use a service or purchase a product.

This area of study is called opinion mining or sentiment analysis. It analyzes people’s opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes. Opinions are important because they are key influencers of our behaviors. Our beliefs and perceptions of reality, and the choices we make, are to a considerable degree conditioned on how others see and evaluate the world. For this reason, when we need to make a decision we often seek out the opinions of others. This is true not only for individuals but also for organizations.

With the explosive growth of social media (i.e., reviews, forum discussions, blogs, and social networks) on the Web, individuals and organizations are increasingly using the content in these media for their decision making. Nowadays, if one wants to buy a consumer product, one is no longer limited to asking one’s friends and family for opinions as in the past because there are many user reviews of products on the Web. For an organization, it may no longer be necessary to conduct opinion polls, surveys, and focus groups in order to gather public opinions about its products and services because there is an abundance of such information publicly available. However, finding and monitoring opinion sites on the Web and distilling the information contained in them remains a formidable task because of the proliferation of diverse sites. Each site typically contains a huge volume of opinionated text that is not always easily deciphered in long forum postings and blogs. The average human reader will have difficulty identifying relevant sites and accurately summarizing the information.
and opinions contained in them. Moreover, it is also known that human analysis and evaluation of text information is subject to considerable biases, e.g., people often pay greater attention to opinions that are consistent with their own preferences. People also have difficulty, owing to their mental and physical limitations, producing consistent results when the amount of information to be processed is large. Automated opinion mining and summarization systems are thus needed, as subjective biases and mental limitations can be overcome with an objective opinion analysis system.

In the past decade, a considerable amount of research has been done in academia [70, 91]. There are also numerous commercial companies that provide opinion mining services. In this chapter, we first define the opinion mining problem. From the definition, we will see the key technical issues that need to be addressed. We then describe various key mining tasks that have been studied in the research literature and their representative techniques. After that, we discuss the related issue of opinion spam detection. Opinion spam refers to dishonest opinions or reviews that try to promote or demote some target products or services. Detecting such spam opinions is critical for practical applications of opinion mining.

11.1 The Problem of Opinion Mining

In this first section, we define an abstraction of the opinion mining problem. It enables us to see a structure from the complex and intimidating unstructured text. Moreover, for most opinion-based applications, it is essential to analyze a collection of opinions rather than only one because one opinion represents only the view of a single person, which is usually not sufficient for action. This indicates that some form of summary of opinions is needed [37]. The abstraction should facilitate this summarization.

11.1.1 Problem Definitions

We use the following review segment on iPhone to introduce the problem (an id number is associated with each sentence for easy reference):

“(1) I bought an iPhone a few days ago. (2) It was such a nice phone. (3) The touch screen was really cool. (4) The voice quality was clear too. (5) However, my mother was mad with me as I did not tell her before I bought it. (6) She also thought the phone was too expensive, and wanted me to return it to the shop. …”

The question is: what we want to mine or extract from this review? The first thing that we notice is that there are several opinions in this review.