CHAPTER 1

Sentiment Analysis: A Fascinating Problem

Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. It represents a large problem space. There are also many names and slightly different tasks, e.g., *sentiment analysis*, *opinion mining*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emotion analysis*, *review mining*, etc. However, they are now all under the umbrella of sentiment analysis or opinion mining. While in industry, the term *sentiment analysis* is more commonly used, but in academia both *sentiment analysis* and *opinion mining* are frequently employed. They basically represent the same field of study. The term *sentiment analysis* perhaps first appeared in (Nasukawa and Yi, 2003), and the term *opinion mining* first appeared in (Dave, Lawrence and Pennock, 2003). However, the research on *sentiments* and *opinions* appeared earlier (Das and Chen, 2001; Morinaga et al., 2002; Pang, Lee and Vaithyanathan, 2002; Tong, 2001; Turney, 2002; Wiebe, 2000). In this book, we use the terms *sentiment analysis* and *opinion mining* interchangeably. To simplify the presentation, throughout this book we will use the term *opinion* to denote opinion, sentiment, evaluation, appraisal, attitude, and emotion. However, these concepts are not equivalent. We will distinguish them when needed. The meaning of opinion itself is still very broad. Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments.

Although linguistics and natural language processing (NLP) have a long history, little research had been done about people's opinions and sentiments before the year 2000. Since then, the field has become a very active research area. There are several reasons for this. First, it has a wide arrange of applications, almost in every domain. The industry surrounding sentiment analysis has also flourished due to the proliferation of commercial applications. This provides a strong motivation for research. Second, it offers many challenging research problems, which had never been studied before. This book will systematically define and discuss these problems, and describe the current state-of-the-art techniques for solving them. Third, for

the first time in human history, we now have a huge volume of opinionated data in the social media on the Web. Without this data, a lot of research would not have been possible. Not surprisingly, the inception and the rapid growth of sentiment analysis coincide with those of the social media. In fact, sentiment analysis is now right at the center of the social media research. Hence, research in sentiment analysis not only has an important impact on NLP, but may also have a profound impact on management sciences, political science, economics, and social sciences as they are all affected by people's opinions. Although the sentiment analysis research mainly started from early 2000, there were some earlier work on interpretation of metaphors, sentiment adjectives, subjectivity, view points, and affects (Hatzivassiloglou and McKeown, 1997; Hearst, 1992; Wiebe, 1990; Wiebe, 1994; Wiebe, Bruce and O'Hara, 1999). This book serves as an up-to-date and comprehensive introductory text, as well as a survey to the subject.

1.1 Sentiment Analysis Applications

Opinions are central to almost all human activities because they are key influencers of our behaviors. Whenever we need to make a decision, we want to know others' opinions. In the real world, businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it, and others' opinions about political candidates before making a voting decision in a political election. In the past, when an individual needed opinions, he/she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. Acquiring public and consumer opinions has long been a huge business itself for marketing, public relations, and political campaign companies.

With the explosive growth of social media (e.g., reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites) on the Web, individuals and organizations are increasingly using the content in these media for 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 because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions 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 opinion text that is not always easily deciphered in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analysis systems are thus needed.

In recent years, we have witnessed that opinionated postings in social media have helped reshape businesses, and sway public sentiments and emotions, which have profoundly impacted on our social and political systems. Such postings have also mobilized masses for political changes such as those happened in some Arab countries in 2011. It has thus become a necessity to collect and study opinions on the Web. Of course, opinionated documents not only exist on the Web (called external data), many organizations also have their internal data, e.g., customer feedback collected from emails and call centers or results from surveys conducted by the organizations.

Due to these applications, industrial activities have flourished in recent years. Sentiment analysis applications have spread to almost every possible domain, from consumer products, services, healthcare, and financial services to social events and political elections. I myself have implemented a sentiment analysis system called *Opinion Parser*, and worked on projects in all these areas in a start-up company. There have been at least 40-60 start-up companies in the space in the USA alone. Many big corporations have also built their own in-house capabilities, e.g., Microsoft, Google, Hewlett-Packard, SAP, and SAS. These practical applications and industrial interests have provided strong motivations for research in sentiment analysis.

Apart from real-life applications, many application-oriented research papers have also been published. For example, in (Liu et al., 2007), a sentiment model was proposed to predict sales performance. In (McGlohon, Glance and Reiter, 2010), reviews were used to rank products and merchants. In (Hong and Skiena, 2010), the relationships between the NFL betting line and public opinions in blogs and Twitter were studied. In (O'Connor et al., 2010), Twitter sentiment was linked with public opinion polls. In (Tumasjan et al., 2010), Twitter sentiment was also applied to predict election results. In (Chen et al., 2010), the authors studied political standpoints. In (Yano and Smith, 2010), a method was reported for predicting comment volumes of political blogs. In (Asur and Huberman, 2010; Joshi et al., 2010; Sadikov, Parameswaran and Venetis, 2009), Twitter data, movie reviews and blogs were used to predict box-office revenues for movies. In (Miller et al., 2011), sentiment flow in social networks was investigated. In (Mohammad and Yang, 2011), sentiments in mails were used to find how genders differed on emotional axes. In (Mohammad, 2011), emotions in novels and fairy tales were tracked. In (Bollen, Mao and Zeng, 2011), Twitter moods were used to

predict the stock market. In (Bar-Haim et al., 2011; Feldman et al., 2011), expert investors in microblogs were identified and sentiment analysis of stocks was performed. In (Zhang and Skiena, 2010), blog and news sentiment was used to study trading strategies. In (Sakunkoo and Sakunkoo, 2009), social influences in online book reviews were studied. In (Groh and Hauffa, 2011), sentiment analysis was used to characterize social relations. A comprehensive sentiment analysis system and some case studies were also reported in (Castellanos et al., 2011). My own group has tracked opinions about movies on Twitter and predicted box-office revenues with very accurate results. We simply used our *Opinion Parser* system to analyze positive and negative opinions about each movie with no additional algorithms.

1.2 Sentiment Analysis Research

As discussed above, pervasive real-life applications are only part of the reason why sentiment analysis is a popular research problem. It is also highly challenging as a NLP research topic, and covers many novel sub-problems as we will see later. Additionally, there was little research before the year 2000 in either NLP or in linguistics. Part of the reason is that before then there was little opinion text available in digital forms. Since the year 2000, the field has grown rapidly to become one of the most active research areas in NLP. It is also widely researched in data mining, Web mining, and information retrieval. In fact, it has spread from computer science to management sciences (Archak, Ghose and Ipeirotis, 2007; Chen and Xie, 2008; Das and Chen, 2007; Dellarocas, Zhang and Awad, 2007; Ghose, Ipeirotis and Sundararajan, 2007; Hu, Pavlou and Zhang, 2006; Park, Lee and Han, 2007).

1.2.1 Different Levels of Analysis

I now give a brief introduction to the main research problems based on the level of granularities of the existing research. In general, sentiment analysis has been investigated mainly at three levels:

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment (Pang, Lee and Vaithyanathan, 2002; Turney, 2002). For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This task is commonly

known as *document-level sentiment classification*. This level of analysis assumes that each document expresses opinions on a single entity (e.g., a single product). Thus, it is not applicable to documents which evaluate or compare multiple entities.

Sentence level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion. Neutral usually means no opinion. This level of analysis is closely related to *subjectivity classification* (Wiebe, Bruce and O'Hara, 1999), which distinguishes sentences (called *objective sentences*) that express factual information from sentences (called *subjective sentences*) that express subjective views and opinions. However, we should note that subjectivity is not equivalent to sentiment as many objective sentences can imply opinions, e.g., "*We bought the car last month and the windshield wiper has fallen off*." Researchers have also analyzed clauses (Wilson, Wiebe and Hwa, 2004), but the clause level is still not enough, e.g., "*Apple is doing very well in this lousy economy*."

Entity and Aspect level: Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called *feature level* (*feature-based opinion mining and summarization*) (Hu and Liu, 2004). Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of a *sentiment* (positive or negative) and a *target* (of opinion). An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also helps us understand the sentiment analysis problem better. For example, although the sentence "*although the service is not that great, I still love this restaurant*" clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the *restaurant* (emphasized), but negative about its *service* (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. For example, the sentence "*The iPhone's call quality is good, but its battery life is short*" evaluates two aspects, *call quality* and *battery life*, of *iPhone* (entity). The sentiment on iPhone's *call quality* is positive, but the sentiment on its *battery life* is negative. The *call quality* and *battery life* of iPhone are the opinion targets. Based on this level of analysis, a structured summary of opinions about entities and their aspects can be produced, which turns unstructured text to structured data and can be used for all kinds of qualitative and quantitative analyses. Both the document level and sentence level classifications are already

highly challenging. The aspect-level is even more difficult. It consists of several sub-problems, which we will discuss in Chapters 2 and 5.

To make things even more interesting and challenging, there are two types of opinions, i.e., *regular opinions* and *comparative opinions* (Jindal and Liu, 2006b)**.** A regular opinion expresses a sentiment only on an particular entity or an aspect of the entity, e.g., "*Coke tastes very good,*" which expresses a positive sentiment on the aspect *taste* of Coke. A comparative opinion compares multiple entities based on some of their shared aspects, e.g., "*Coke tastes better than Pepsi*," which compares Coke and Pepsi based on their tastes (an aspect) and expresses a preference for Coke (see Chapter 8).

1.2.2 Sentiment Lexicon and Its Issues

Not surprisingly, the most important indicators of sentiments are *sentiment words*, also called *opinion words*. These are words that are commonly used to express positive or negative sentiments. For example, *good*, *wonderful*, and *amazing* are positive sentiment words, and *bad*, *poor*, and *terrible* are negative sentiment words. Apart from individual words, there are also phrases and idioms, e.g., *cost someone an arm and a leg*. Sentiment words and phrases are instrumental to sentiment analysis for obvious reasons. A list of such words and phrases is called a *sentiment lexicon* (or *opinion lexicon*). Over the years, researchers have designed numerous algorithms to compile such lexicons. We will discuss these algorithms in Chapter 6.

Although sentiment words and phrases are important for sentiment analysis, only using them is far from sufficient. The problem is much more complex. In other words, we can say that sentiment lexicon is necessary but not sufficient for sentiment analysis. Below, we highlight several issues:

1. A positive or negative sentiment word may have opposite orientations in different application domains. For example, "suck" usually indicates negative sentiment, e.g., "*This camera sucks*," but it can also imply positive sentiment, e.g., "*This vacuum cleaner really sucks*."
2. A sentence containing sentiment words may not express any sentiment. This phenomenon happens frequently in several types of sentences. Question (interrogative) sentences and conditional sentences are two important types, e.g., "*Can you tell me which Sony camera is good*?" and "*If I can find a good camera in the shop, I will buy it*." Both these sentences contain the sentiment word "*good*", but neither expresses a positive or negative opinion on any specific camera. However, not all conditional sentences or interrogative sentences express no sentiments, e.g., "*Does anyone know how to repair this terrible printer*" and "*If you*

are looking for a good car, get Toyota Camry." We will discuss such sentences in Chapter 4.

3. Sarcastic sentences with or without sentiment words are hard to deal with, e.g., "*What a great car! It stopped working in two days.*" Sarcasms are not so common in consumer reviews about products and services, but are very common in political discussions, which make political opinions hard to deal with. We will discuss such sentences in Chapter 4.
4. Many sentences without sentiment words can also imply opinions. Many of these sentences are actually objective sentences that are used to express some factual information. Again, there are many types of such sentences. Here we just give two examples. The sentence "*This washer uses a lot of water*" implies a negative sentiment about the washer since it uses a lot of resource (water). The sentence "*After sleeping on the mattress for two days, a valley has formed in the middle*" expresses a negative opinion about the mattress. This sentence is objective as it states a fact. All these sentences have no sentiment words.

These issues all present major challenges. In fact, these are just some of the difficult problems. More will be discussed in Chapter 5.

1.2.3 Natural Language Processing Issues

Finally, we must not forget sentiment analysis is a NLP problem. It touches every aspect of NLP, e.g., coreference resolution, negation handling, and word sense disambiguation, which add more difficulties since these are not solved problems in NLP. However, it is also useful to realize that sentiment analysis is a highly restricted NLP problem because the system does not need to fully understand the semantics of each sentence or document but only needs to understand some aspects of it, i.e., positive or negative sentiments and their target entities or topics. In this sense, sentiment analysis offers a great platform for NLP researchers to make tangible progresses on all fronts of NLP with the potential of making a huge practical impact. In this book, I will describe the core problems and the current state-of-the-art algorithms. I hope to use this book to attract researchers from other areas of NLP to join force to make a concerted effort to solve the problem.

Prior to this book, there were a multi-author volume "*Computing Attitude and Affect in Text: Theory and Applications*" edited by Shanahan, Qu, and Wiebe (2006), and also a survey article/book by Pang and Lee (2008). Both books have excellent contents. However, they were published relatively early in the development of the field. Since then, there have been significant advancements due to much more active research in the past 5 years.

Researchers now also have a much better understanding of the whole spectrum of the problem, its structure, and core issues. Numerous new (formal) models and methods have been proposed. The research has not only deepened but also broadened significantly. Earlier research in the field mainly focused on classifying the sentiment or subjectivity expressed in documents or sentences, which is insufficient for most real-life applications. Practical applications often demand more in-depth and fine-grained analysis. Due to the maturity of the field, the book is also written in a structured form in the sense that the problem is now better defined and different research directions are unified around the definition.

1.3 Opinion Spam Detection

A key feature of social media is that it enables anyone from anywhere in the world to freely express his/her views and opinions without disclosing his/her true identify and without the fear of undesirable consequences. These opinions are thus highly valuable. However, this anonymity also comes with a price. It allows people with hidden agendas or malicious intentions to easily game the system to give people the impression that they are independent members of the public and post fake opinions to promote or to discredit target products, services, organizations, or individuals without disclosing their true intentions, or the person or organization that they are secretly working for. Such individuals are called *opinion spammers* and their activities are called *opinion spamming* (Jindal and Liu, 2008; Jindal and Liu, 2007).

Opinion spamming has become a major issue. Apart from individuals who give fake opinions in reviews and forum discussions, there are also commercial companies that are in the business of writing fake reviews and bogus blogs for their clients. Several high profile cases of fake reviews have been reported in the news. It is important to detect such spamming activities to ensure that the opinions on the Web are a trusted source of valuable information. Unlike extraction of positive and negative opinions, opinion spam detection is not just a NLP problem as it involves the analysis of people's posting behaviors. It is thus also a data mining problem. Chapter 10 will discuss the current state-of-the-art detection techniques.

1.4 What's Ahead

In this book, we explore this fascinating topic. Although the book deals with

the natural language text, which is often called *unstructured data*, I take a structured approach to writing this book. The next chapter will formally define the problem, which allows us to see a structure of the problem. From the definition, we will see the key tasks of sentiment analysis. In the subsequent chapters, existing techniques for performing the tasks are described. Due to my research, consulting, and start-up experiences, the book not only discusses key research concepts but also looks at the technology from an application point of view in order to help practitioners in the field. However, I must apologize that when I talk about industrial systems, I cannot reveal the names of companies or their systems, partially because of my consulting/business agreements and partially because of the fact that the sentiment analysis market moves rapidly and the companies that I know of may have changed or improved their algorithms when you read this book. I do not want to create problems for them and for me.

Although I try to cover all major ideas and techniques in this book, it has become an impossible task. In the past decade, a huge number of research papers (probably more than 1000) have been published on the topic. Although most papers appeared in NLP conferences and journals, many papers have also been published in data mining, Web mining, machine learning, information retrieval, e-commerce, management sciences, and many other fields. It is thus almost impossible to write a book that covers the ideas in every published paper. I am sorry if your good ideas or techniques are overlooked. However, a major advantage of publishing this book in the synthesis lecture series of Morgan & Claypool is that the authors can always add new or updated materials to the book because the printing is on demand. So if you find that some important ideas are not discussed, please do not hesitate to let me know and I will be very happy to include.

Finally, background knowledge in the following areas will be very helpful in reading this book: natural language processing (Indurkhya and Damerau, 2010; Manning and Schutze, 1999), machine learning (Bishop, 2006; Mitchell, 1997), data mining (Liu, 2006 and 2011), and information retrieval (Manning, Raghavan and Schutze, 2008).

CHAPTER 2

The Problem of Sentiment Analysis

In this chapter, we define an abstraction of the sentiment analysis or opinion mining problem. From a research point of view, this abstraction gives us a statement of the problem and enables us to see a rich set of inter-related sub-problems which make up the sentiment analysis problem. It is often said that if we cannot structure a problem, we probably do not understand the problem. The objective of the definitions is thus to abstract a structure from the complex and intimidating unstructured natural language text. They also serve as a common framework to unify various existing research directions, and to enable researchers to design more robust and accurate solution techniques by exploiting the inter-relationships of the sub-problems. From a practical application point of view, the definitions let practitioners see what sub-problems need to be solved in a practical system, how they are related, and what output should be produced.

Unlike factual information, opinions and sentiments have an important characteristic, namely, they are subjective. It is thus important to examine a collection of opinions from many people rather than only a single opinion from one person because such an opinion represents only the subjective view of that single person, which is usually not sufficient for application. Due to a large collection of opinions on the Web, some form of summary of opinions is needed (Hu and Liu, 2004). The problem definitions state what kind of summary may be desired. Along with the problem definitions, the chapter will also discuss several related concepts such as subjectivity and emotion.

Note that throughout this chapter and also the whole book, I mainly use reviews and sentences from reviews as examples to introduce ideas and to define key concepts, but the ideas and the resulting definitions are general and applicable to all forms of formal and informal opinion text such as news articles, tweets (Twitter postings), forum discussions, blogs, and Facebook postings. Since product reviews are highly focused and opinion rich, they allow us to see different issues more clearly than from other forms of opinion text. Conceptually, there is no difference between them. The differences are mainly superficial and in the degree of difficulty in dealing with them. For example, Twitter postings (tweets) are short (at most 140 characters) and informal, and use many Internet slangs and emoticons. Twitter postings are, in fact, easier to analyze due to the length limit because

the authors are usually straight to the point. Thus, it is often easier to achieve high sentiment analysis accuracy. Reviews are also easier because they are highly focused with little irrelevant information. Forum discussions are perhaps the hardest to deal with because the users there can discuss anything and also interact with one another. In terms of the degree of difficulty, there is also the dimension of different application domains. Opinions about products and services are usually easier to analyze. Social and political discussions are much harder due to complex topic and sentiment expressions, sarcasms and ironies.

2.1 Problem Definitions

As mentioned at the beginning of Chapter 1, sentiment analysis mainly studies opinions which express or imply positive or negative sentiments. This section thus defines the problem in this context.

2.1.1 Opinion Defintion

We use the following review about a Canon camera to introduce the problem (an id number is associated with each sentence for easy reference):

> Posted by: John Smith Date: September 10, 2011
>
> "(1) *I bought a Canon G12 camera six months ago*. (2) *I simply love it*. (3) *The picture quality is amazing*. (4) *The battery life is also long*. (5) *However, my wife thinks it is too heavy for her*."

From this review, we notice a few important points:

1. The review has a number of opinions, both positive and negative, about Canon G12 camera. Sentence (2) expresses a positive opinion about the Canon camera as a whole. Sentence (3) expresses a positive opinion about its picture equality. Sentence (4) expresses a positive opinion about its battery life. Sentence (5) expresses a negative opinion about the weight of the camera. From these opinions, we can make the following important observation:

   Observation: An opinion consists of two key components: a target g and a sentiment s on the target, i.e.,

   $$(g, s),$$

   where g can be any entity or aspect of the entity about which an opinion has been expressed, and s is a positive, negative, or neutral sentiment, or a numeric rating score expressing the strength/intensity

of the sentiment (e.g., 1 to 5 stars). Positive, negative and neutral are called *sentiment* (or *opinion*) *orientations* (or *polarities*).

For example, the target of the opinion in sentence (2) is *Canon G12*, and the target of the opinion in sentence (3) is the *picture quality of Canon G12*. Target is also called *topic* in the literature.

2. This review has opinions from two persons, which are called *opinion sources* or *opinion holders* (Kim and Hovy, 2004; Wiebe, Wilson and Cardie, 2005). The holder of the opinions in sentences (2), (3), and (4) is the author of the review ("John Smith"), but for sentence (5), it is the wife of the author.
3. The date of the review is September 10, 2011. This date is important in practice because one often wants to know how opinions change with time and opinion trends.

We are now ready to define opinion as a quadruple.

Definition (Opinion): An *opinion* is a quadruple,

$$(g, s, h, t),$$

where g is the opinion (or sentiment) target, s is the sentiment about the target, h is the opinion holder and t is the time when the opinion was expressed.

This definition, although quite concise, may not be easy to use in practice especially in the domain of online reviews of products, services, and brands because the full description of the target can be complex and may not even appear in the same sentence. For example, in sentence (3), the opinion target is actually "picture quality of Canon G12", but the sentence mentioned only "picture quality". In this case, the opinion target is not just "picture quality" because without knowing that the sentence is evaluating the picture quality of the Canon G12 camera, the opinion in sentence (3) alone is of little use. In practice, the target can often be decomposed and described in a structured manner with multiple levels, which greatly facilitate both mining of opinions and later use of the mined opinion results. For example, "picture quality of Canon G12" can be decomposed into an entity and an attribute of the entity and represented as a pair,

(Cannon-G12, picture-quality)

Let us use the term *entity* to denote the target object that has been evaluated. Entity can be defined as follows (Hu and Liu, 2004; Liu, 2006 and 2011).

Definition (**entity**)**:** An *entity* e is a product, service, topic, issue, person, organization, or event. It is described with a pair, e: (T, W), where T is a hierarchy of *parts*, *sub-parts*, and so on, and W is a set of *attributes* of e.

Each part or sub-part also has its own set of attributes.

Example 1: A particular model of camera is an entity, e.g., Canon G12. It has a set of attributes, e.g., *picture quality*, *size*, and *weight*, and a set of parts, e.g., *lens*, *viewfinder*, and *battery*. *Battery* also has its own set of attributes, e.g., *battery life* and *battery weight*. A topic can be an entity too, e.g., *tax increase*, with its parts "*tax increase for the poor*," "*tax increase for the middle class*" and "*tax increase for the rich*."

This definition essentially describes a hierarchical decomposition of entity based on the *part-of* relation. The root node is the name of the entity, e.g., Canon G12 in the above review. All the other nodes are parts and sub-parts, etc. An opinion can be expressed on any node and any attribute of the node.

Example 2: In our example review above, sentence (2) expresses a positive opinion about the entity Canon G12 camera as a whole. Sentence (3) expresses a positive opinion on the attribute of picture quality of the camera. Clearly, one can also express opinions about parts or components of the camera.

This entity as a hierarchy of any number of levels needs a nested relation to represent it, which is often too complex for applications. The main reason is that since NLP is a very difficult task, recognizing parts and attributes of an entity at different levels of details is extremely hard. Most applications also do not need such a complex analysis. Thus, we simplify the hierarchy to two levels and use the term *aspects* to denote both parts and attributes. In the simplified tree, the root node is still the entity itself, but the second level (also the leaf level) nodes are different aspects of the entity. This simplified framework is what is typically used in practical sentiment analysis systems.

Note that in the research literature, entities are also called *objects*, and aspects are also called *features* (as in product features). However, features here can confuse with features used in machine learning, where a feature means a data attribute. To avoid confusion, aspects have become more popular in recent years. Note that some researchers also use the terms *facets*, *attributes* and *topics*, and in specific applications, entities and aspects may also be called other names based on the application domain conventions.

After decomposing the opinion target, we can redefine an opinion (Hu and Liu, 2004; Liu, 2010).

Definition (opinion): An *opinion* is a quintuple,

$$(e_i, a_{ij}, s_{ijkl}, h_k, t_l),$$

where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or

neutral, or expressed with different strength/intensity levels, e.g., 1 to 5 stars as used by most review sits on the Web. When an opinion is on the entity itself as a whole, the special aspect GENERAL is used to denote it. Here, e_i and a_{ij} together represent the opinion target.

Some important remarks about this definition are in order:

1. In this definition, we purposely use subscripts to emphasize that the five pieces of information in the quintuple must correspond to one another. That is, the opinion s_{ijkl} must be given by opinion holder h_k about aspect a_{ij} of entity e_i at time t_l. Any mismatch is an error.
2. The five components are essential. Missing any of them is problematic in general. For example, if we do not have the time component, we will not be able to analyze opinions on an entity according to time, which is often very important in practice because an opinion two years ago and an opinion yesterday is not the same. Without opinion holder is also problematic. For example, in the sentence "*the mayor is loved by the people in the city, but he has been criticized by the state government,*" the two opinion holders, "*people in the city*" and "*state government,*" are clearly important for applications.
3. The definition covers most but not all possible facets of the semantic meaning of an opinion, which can be arbitrarily complex. For example, it does not cover the situation in "*The view finder and the lens are too close*," which expresses an opinion on the distance of two parts. It also does not cover the context of the opinion, e g., "*This car is too small for a tall person*," which does not say the car is too small for everyone. "Tall person" is the context here. Note also that in the original definition of entity, it is a hierarchy of parts, sub-parts, and so on. Every part can have its set of attributes. Due to the simplification, the quintuple representation can result in information loss. For example, "ink" is a part/component of a printer. In a printer review, one wrote "*The ink of this printer is expensive*." This does not say that the printer is expensive (which indicates the aspect *price*). If one does not care about any attribute of the ink, this sentence just gives a negative opinion to the ink, which is an aspect of the printer entity. However, if one also wants to study opinions about different aspects of the ink, e.g., price and quality, the ink needs to be treated as a separate entity. Then, the quintuple representation still applies, but the part-of relationship needs to be saved. Of course, conceptually we can also expand the representation of opinion target using a nested relation. Despite the limitations, the definition does cover the essential information of an opinion which is sufficient for most applications. As we mentioned above, too complex a definition can make the problem extremely difficult to solve.

4. This definition provides a framework to transform unstructured text to structured data. The quintuple above is basically a database schema, based on which the extracted opinions can be put into a database table. Then a rich set of qualitative, quantitative, and trend analyses of opinions can be performed using the whole suite of database management systems (DBMS) and OLAP tools.
5. The opinion defined here is just one type of opinion, called *regular opinion*. Another type is *comparative opinion* (Jindal and Liu, 2006b; Liu, 2006 and 2011), which needs a different definition. Section 2.3 will discuss different types of opinions. Chapter 8 defines and analyzes comparative opinions. For the rest of this section, we only focus on regular opinions. For simplicity, we just called them opinions.

2.1.2 Sentiment Analysis Tasks

With the definition, we can now present the objective and the key tasks of sentiment analysis (Liu, 2010; Liu, 2006 and 2011).

Objective of sentiment analysis: Given an opinion document d, discover all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) in d.

The key tasks are derived from the 5 components of the quintuple. The first component is the entity. That is, we need to extract entities. The task is similar to named entity recognition (NER) in information extraction (Hobbs and Riloff, 2010; Mooney and Bunescu, 2005; Sarawagi, 2008). Thus, the extraction itself is a problem. After extraction, we also need to categorize the extracted entities. In natural language text, people often write the same entity in different ways. For example, Motorola may be written as Mot, Moto, and Motorola. We need to recognize that they all refer to the same entity.

Definition (entity category and entity expression): An *entity category* represents a unique entity, while an *entity expression* is an actual word or phrase that appears in the text indicating an entity category.

Each entity category (or simply entity) should have a unique name in a particular application. The process of grouping entity expressions into entity categories is called *entity categorization*.

Now we look at aspects of entities. The problem is basically the same as for entities. For example, *picture*, *image*, and *photo* are the same aspect for cameras. We thus need to extract aspect expressions and categorize them.

Definition (aspect category and aspect expression): An *aspect category* of an entity represents a unique aspect of the entity, while an *aspect*

expression is an actual word or phrase that appears in the text indicating an aspect category.

Each aspect category (or simply aspect) should also have a unique name in a particular application. The process of grouping aspect expressions into aspect categories (aspects) is called *aspect categorization*.

Aspect expressions are usually nouns and noun phrases but can also be verbs, verb phrases, adjectives, and adverbs. The following definitions are useful (Hu and Liu, 2004).

Definition (**explicit aspect expression**): Aspect expressions that are nouns and noun phrases are called *explicit aspect expressions*.

For example, "picture quality" in "*The picture quality of this camera is great*" is an explicit aspect expression.

Definition (**implicit aspect expression**): Aspect expressions that are not nouns or noun phrases are called *implicit aspect expressions*.

For example, "expensive" is an implicit aspect expression in "*This camera is expensive*." It implies the aspect *price*. Many implicit aspect expressions are adjectives and adverbs that are used to describe or qualify some specific aspects, e.g., *expensive* (price), and *reliably* (reliability). They can also be verb and verb phrases, e.g., "*I can install the software easily*." "Install" indicates the aspect *installation*. Implicit aspect expressions are not just adjectives, adverbs, verbs and verb phrases; they can also be very complex, e.g., "*This camera will not easily fit in a coat pocket*." Here, "fit in a coat pocket" indicates the aspect *size* (and/or *shape*).

The third component in the opinion definition is the sentiment. This task classifies whether the sentiment on the aspect is positive, negative or neutral. The fourth component and fifth components are opinion holder and time respectively. They also need to be extracted and categorized as for entities and aspects. Note that an opinion holder (Bethard et al., 2004; Choi et al., 2005; Kim and Hovy, 2004) (also called opinion source in (Wiebe, Wilson and Cardie, 2005)) can be a person or organization who expressed an opinion. For product reviews and blogs, opinion holders are usually the authors of the postings. Opinion holders are more important for news articles as they often explicitly state the person or organization that holds an opinion. However, in some cases, identifying opinion holders can also be important in social media, e.g., identifying opinions from advertisers or people who quote advertisements of companies.

Based on the above discussions, we can define a model of entity and a model of opinion document (Liu, 2006 and 2011).

Model of entity: An entity e_i is represented by itself as a whole and a finite set of aspects $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$. e_i can be expressed with any one of a finite set of its entity expressions $\{ee_{i1}, ee_{i2}, \ldots, ee_{is}\}$. Each aspect $a_{ij} \in A_i$ of entity e_i can be expressed with any one of its finite set of aspect expressions $\{ae_{ij1}, ae_{ij2}, \ldots, ae_{ijm}\}$.

Model of opinion document: An opinion document d contains opinions on a set of entities $\{e_1, e_2, \ldots, e_r\}$ and a subset of their aspects from a set of opinion holders $\{h_1, h_2, \ldots, h_p\}$ at some particular time point.

Finally, to summarize, given a set of opinion documents D, sentiment analysis consists of the following 6 main tasks.

Task 1 (entity extraction and categorization): Extract all entity expressions in D, and categorize or group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster indicates a unique entity e_i.

Task 2 (aspect extraction and categorization): Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect a_{ij}.

Task 3 (opinion holder extraction and categorization): Extract opinion holders for opinions from text or structured data and categorize them. The task is analogous to the above two tasks.

Task 4 (time extraction and standardization): Extract the times when opinions are given and standardize different time formats. The task is also analogous to the above tasks.

Task 5 (aspect sentiment classification): Determine whether an opinion on an aspect a_{ij} is positive, negative or neutral, or assign a numeric sentiment rating to the aspect.

Task 6 (opinion quintuple generation): Produce all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) expressed in document d based on the results of the above tasks. This task is seemingly very simple but it is in fact very difficult in many cases as Example 4 below shows.

Sentiment analysis (or opinion mining) based on this framework is often called *aspect-based sentiment analysis* **(**or *opinion mining***),** or *feature-based sentiment analysis* (*or opinion mining*) as it was called in (Hu and Liu, 2004; Liu, Hu and Cheng, 2005).

We now use an example blog to illustrate the tasks (a sentence id is again associated with each sentence) and the analysis results.

Example 4: Posted by: bigJohn Date: Sept. 15, 2011

(1) *I bought a Samsung camera and my friends brought a Canon camera yesterday.* (2) *In the past week, we both used the cameras a lot.* (3) *The photos from my Samy are not that great, and the battery*

life is short too. (4) *My friend was very happy with his camera and loves its picture quality*. (5) *I want a camera that can take good photos*. (6) *I am going to return it tomorrow*.

Task 1 should extract the entity expressions, "Samsung," "Samy," and "Canon," and group "Samsung" and "Samy" together as they represent the same entity. Task 2 should extract aspect expressions "picture," "photo," and "battery life," and group "picture" and "photo" together as for cameras they are synonyms. Task 3 should find the holder of the opinions in sentence (3) to be bigJohn (the blog author) and the holder of the opinions in sentence (4) to be bigJohn's friend. Task 4 should also find the time when the blog was posted is Sept-15-2011. Task 5 should find that sentence (3) gives a negative opinion to the picture quality of the Samsung camera and also a negative opinion to its battery life. Sentence (4) gives a positive opinion to the Canon camera as a whole and also to its picture quality. Sentence (5) seemingly expresses a positive opinion, but it does not. To generate opinion quintuples for sentence (4) we need to know what "his camera" and "its" refer to. Task 6 should finally generate the following four opinion quintuples:

(Samsung, picture_quality, negative, bigJohn, Sept-15-2011)
(Samsung, battery_life, negative, bigJohn, Sept-15-2011)
(Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)
(Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

2.2 Opinion Summarization

Unlike factual information, opinions are essentially subjective. One opinion from a single opinion holder is usually not sufficient for action. In most applications, one needs to analyze opinions from a large number of people. This indicates that some form of summary of opinions is desired. Although an opinion summary can be in one of many forms, e.g., structured summary (see below) or short text summary, the key components of a summary should include opinions about different entities and their aspects and should also have a quantitative perspective. The quantitative perspective is especially important because 20% of the people being positive about a product is very different from 80% of the people being positive about the product. We will discuss this further in Chapter 7.

The opinion quintuple defined above actually provides a good source of information and also a framework for generating both *qualitative* and *quantitative* summaries. A common form of summary is based on aspects and is called *aspect-based opinion summary* (or *feature-based opinion summary*) (Hu and Liu, 2004; Liu, Hu and Cheng, 2005). In the past few