1

Introduction

Sentiment analysis, also called *opinion mining*, is the field of study that analyzes people's opinions, sentiments, appraisals, attitudes, and emotions toward entities and their attributes expressed in written text. The entities can be products, services, organizations, individuals, events, issues, or topics. The field represents a large problem space. Many related names and slightly different tasks, for example, *sentiment analysis*, *opinion mining*, *opinion analysis*, *opinion extraction*, *sentiment mining*, *subjectivity analysis*, *affect analysis*, *emotion analysis*, and *review mining*, are now all under the umbrella of *sentiment analysis*. The term *sentiment analysis* perhaps first appeared in Nasukawa and Yi (2003), and the term *opinion mining* first appeared in Dave et al. (2003). However, research on *sentiment* and *opinion* began earlier (Wiebe, 2000; Das and Chen, 2001; Tong, 2001; Morinaga et al., 2002; Pang et al., 2002; Turney, 2002). Even earlier related work includes interpretation of metaphors; extraction of sentiment adjectives; affective computing; and analysis of subjectivity, viewpoints, and affects (Wiebe, 1990, 1994; Hearst, 1992; Hatzivassiloglou and McKeown, 1997; Picard, 1997; Wiebe et al., 1999). An early patent on text classification included sentiment, appropriateness, humor, and many other concepts as possible class labels (Elkan, 2001). Since existing research and applications of sentiment analysis have focused primarily on written text, it has been an active research field of natural language processing (NLP). However, the topic has also been widely studied in data mining, web mining, and information retrieval because many researchers in these fields deal with text data. My own first paper (Hu and Liu, 2004) on the topic was published in the proceedings of the data mining conference KDD (SIGKDD International Conference on Knowledge Discovery and Data Mining) in 2004. This paper defined the aspect-based sentiment analysis and summarization framework and some basic ideas and algorithms for solving the problem that are commonly used in research and industrial systems today.

Not surprisingly, there has been some confusion among practitioners and even researchers about the difference between *sentiment* and *opinion* and whether the field should be called *sentiment analysis* or *opinion mining*. Because the field originated from computer science rather than linguistics, little discussion has concerned the difference between the two words. In Merriam-Webster's dictionary, *sentiment* is defined as an attitude, thought, or judgment prompted by feeling, whereas *opinion* is defined as a view, judgment, or appraisal formed in the mind about a particular matter. The difference is quite subtle, and each contains some elements of the other. The definitions indicate that an opinion is more of a person's concrete view about something, whereas a sentiment is more of a feeling. For example, the sentence "*I am concerned about the current state of the economy*" expresses a sentiment, whereas the sentence "*I think the economy is not doing well*" expresses an opinion. In a conversation, if someone says the first sentence, we can respond by saying, "*I share your sentiment*," but for the second sentence, we would normally say, "*I agree/disagree with you*." However, the underlying meanings of the two sentences are related because the sentiment depicted in the first sentence is likely to be a feeling caused by the opinion in the second sentence. Conversely, we can also say that the first sentiment sentence implies a negative opinion about the economy, which is what the second sentence is saying. Although in most cases opinions imply positive or negative sentiments, some opinions do not, for example, "*I think he will go to Canada next year*."

Regarding the name of the field, *sentiment analysis* is used almost exclusively in industry, whereas both *opinion mining* and *sentiment analysis* are commonly employed in academia. In this book, I use the terms *sentiment analysis* and *opinion mining* interchangeably. Furthermore, I use the term *opinion* to mean the whole concept of sentiment, evaluation, appraisal, or attitude and associated information, such as the opinion target and the person who holds the opinion (see the formal definition in Section 2.1), and I use the term *sentiment* to mean the underlying positive or negative feeling implied by opinion. Sentiment analysis mainly focuses on opinions that express or imply positive or negative sentiments, also called *positive or negative opinions* in everyday language. This type of opinion is similar to the concept of *attitude* in social psychology. For example, Eagly and Chaiken (1998, p. 1) defined an attitude as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor." In discussing positive and negative sentiments, we must also consider expressions without any implied sentiment, which we call *neutral* expressions. Apart from sentiment and opinion, there are also the concepts of *affect*, *emotion*, and *mood*, which are psychological states of mind. We study natural language expressions of such states in detail in Section 2.3.

Sentences expressing opinions or sentiments are usually *subjective* sentences as opposed to *objective* sentences, which state facts, because opinions and sentiments are inherently subjective. However, objective sentences can imply positive or negative sentiments of their authors too, because they may describe desirable or undesirable facts. For example, based on our commonsense knowledge, we know that "*I bought the car yesterday and it broke today*" and "*after sleeping on the mattress for a month, a valley has formed in the middle*" describe two undesirable facts, and we can safely infer that the sentence authors feel negatively about the car and the mattress. Sentiment analysis also studies such objective sentences.

In a nutshell, sentiment analysis or opinion mining aims to identify positive and negative opinions or sentiments expressed or implied in text and also the targets of these opinions or sentiments (e.g., *the car* and *the mattress* in the preceding sentences). A more formal definition is given in Section 2.1.

Although sentiment analysis studies opinion text, there was almost no research on it from either the linguistics community or the NLP community before the year 2000. This is partly because almost no opinionated text was recorded in digital forms before then, although throughout history, spoken or written communication never had a shortage of opinion. With the explosive growth of the web and social media in the past fifteen years, we now have a constant flow of opinion data recorded in digital forms. Without these data, much of the existing research would not have been possible. It is thus no surprise that the inception and rapid growth of sentiment analysis coincide with the growth of social media on the web.

Over the years, social media systems on the web have provided excellent platforms to facilitate and enable audience participation, engagement, and community, which has resulted in our new participatory culture. From reviews and blogs to YouTube, Facebook, and Twitter, people have embraced these platforms enthusiastically because they enable their users to freely and conveniently voice their opinions and communicate their views on any subject across geographic and spatial boundaries. They also allow people to easily connect with others and to share their information. This participatory web and communications revolution has transformed our everyday lives and society as a whole. It has also popularized two major research areas, namely, *social network analysis* and *sentiment analysis*. Although social network analysis is not a new research area, as it started in the 1940s and 1950s when management science researchers began to study social actors (people in organizations) and their interactions and relationships, social media has certainly fueled its explosive growth in the past fifteen years. Sentiment analysis, conversely, is a new research area that essentially grew out of social media on the web.

Since the year 2002, research in sentiment analysis has been very active. Apart from the availability of a large volume of opinion data in social media, opinions and sentiments also have a very wide range of applications simply because opinions are central to almost all human activities. Whenever we need to make a decision, we often seek out others' opinions. This is true not only for individuals but also for organizations. It is thus no surprise that the industry and applications surrounding sentiment analysis have flourished since around 2006. On one hand, this application need provided a strong motivation for research. On the other, sentiment analysis also offers numerous challenging and fascinating research problems whose solutions have never before been attempted. In this book, I systematically define and discuss these problems and present the current state-of-the-art techniques for studying them.

Because a key function of social media is for people to express their views and opinions, sentiment analysis is right at the center of research and application of social media itself. It is now well recognized that, to extract and exploit information in social media, sentiment analysis is a necessary technology. One can even take a sentiment-centric view of social media content analysis because the most important information that one wants to extract from the social media content is what people talk about and what their opinions are. These are exactly the core tasks of sentiment analysis. Furthermore, we can claim that topics, events, and individuals discussed in social media are unlikely to be important if few people have expressed opinions about them. Human nature being what it is, everything that we consider important arouses our inner feelings or emotions, which are expressed with opinions and sentiments.

Apart from topics and opinions about topics, social media also allows us to study the participants themselves. We can produce a sentiment profile of each social media participant based on his or her topical interests and opinions about these interests expressed in the users' posts, because a person's topical interests and opinions reflect the nature and preferences of the person. Such information can be used in many applications, for example, recommending products and services and determining which political candidates to vote for. Additionally, social media participants can not only post messages but also interact with one another through discussions and debates, which involve sentiments such as *agreement* and *disagreement* (or *contention*). Discovery of such information is also of great importance. For example, contentious social and political issues and views of opposing positions can be exploited to frame political issues and to predict election results.

Owing to the importance of opinions in social media, imposters often game the system by posting fake or deceptive opinions to promote some target products, services, and ideological agendas. Detecting such fake or deceptive opinions is an important challenge, which again offers fertile ground for novel research and applications.

Although sentiment analysis originated from computer science, in recent years, it has spread to management sciences and social sciences because of its importance to business and society as a whole. Thus sentiment analysis research not only advances the field of NLP but also advances research in management science, political science, and economics, as these fields are all concerned with consumer and public opinions. It is thus not hard to imagine that sentiment analysis using social media can profoundly change the direction of research and practice in these fields. This book serves as an up-to-date and introductory text as well as a comprehensive survey of this important and fascinating subject.

1.1 Sentiment Analysis Applications

Opinions are very important to businesses and organizations because they always want to find consumer or public opinions about their products and services. Local and federal governments also want to know public opinions about their existing or proposed policies. Such opinions will enable relevant government decision makers to respond quickly to the fast-changing social, economic, and political climates. In international politics, every government wants to monitor the social media of other countries to find out what is happening in these countries and what people's views and sentiments are about current local and international issues and events. Such information is very useful to diplomacy, international relations, and economic decision making. Besides businesses, organizations, and government agencies, individual consumers also want to know the opinions of others about products, services, and political candidates before purchasing the products, using the services, and making election decisions.

In the past, when an individual needed opinions, he or she asked friends and family. When an organization or a business needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. When governments wanted to know what was happening in other countries, they monitored the traditional news media, for example, newspapers, radio, and TV, in these countries, and even sent spies to these countries to collect such information. Acquiring and analyzing public and consumer opinions have long been a huge business for marketing, public relations, and political campaign firms.

Nowadays, individuals, organizations, and government agencies are increasingly using the content in social media for decision making. If an individual wants to buy a consumer product, he or she is no longer limited to asking his or her 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, or focus groups to gather public or consumer opinions about the organization's products and services because an abundance of such information is publicly available. Governments can also easily obtain public opinions about their policies and measure the pulses of other nations simply by monitoring their social media.

In recent years, we have witnessed how opinionated posts on social media sites have helped reshape business and sway public sentiment, profoundly impacting our social and political lives. For instance, such posts have mobilized the masses for political change, such as during the Arab Spring in 2011. However, finding and monitoring opinion sites on the web and distilling the information contained in them remains a formidable task because of the diversity of sites. Each site typically contains a huge volume of opinion text that is not always easily deciphered from long blogs and forum posts. 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.

Opinionated documents not only exist on the web (often called external data); many organizations have internal data, for example, customer feedback collected from e-mails and call centers and results from surveys conducted by the organizations. It is critical to analyze both kinds of data to tease out the key product and service issues and to summarize customer opinions.

In recent years, sentiment analysis applications have spread to almost every possible domain, from consumer products, health care, tourism, hospitality, and financial services to social events and political elections. There are now hundreds of companies in this space, start-up companies and established large corporations, that have built or are in the process of building their own in-house capabilities, such as Google, Microsoft, Hewlett-Packard, Amazon, eBay, SAS, Oracle, Adobe, Bloomberg, and SAP. I myself have implemented a sentiment analysis system, called Opinion Parser, and worked on projects for clients in more than forty domains: automobile, mobile phone, earphone, printer, fridge, washing machine, stove, Blu-ray, laptop, home theater, television, e-book, GPS, LCD monitor, dieting, hair care product, coffee maker, mattress, paint, cruise, restaurant, hotel, cosmetics, fashion, drug, soft drink, beer and wine, movie, video editing software, financial software, search engine, health insurance, banking, investment, green technology, box-office revenue prediction for new movies, summer Olympic bidding, governor election, presidential election, and public mood during the 2008–9 financial crisis.

In addition to business interests, applications are also widespread in government agencies. Internally, agencies monitor social media to discover public sentiments and citizen concerns. Such monitoring is especially big in China, where social media has become the most popular channel for the general public to voice their opinions about government policies and to expose corruptions, sex scandals, and other wrongdoings of government officials. It is also the quickest and most popular way to report negative events in everyday lives. Weibo, which literally means "microblog" in Chinese and is similar to Twitter, is the most popular platform for such revelations. Several commercial social media monitoring tools are already available. The core technology in these tools is sentiment analysis. Externally, intelligence services discover issues and events being discussed in the social media of other countries and public sentiment about the issues and events by monitoring the main social media sites of these countries.

Besides real-life applications, many application-oriented research papers have also been published. For example, several researchers have used sentiment information to predict movie success and box-office revenue. Mishne and Glance (2006) showed that positive sentiment is a better predictor of movie success than simple buzz (keyword) count. Sadikov et al. (2009) made the same prediction using sentiment and other features. Liu et al. (2007) reported a sentiment model for predicting box-office revenue. The method consists of two steps. The first step builds a topic model based on probabilistic latent semantic analysis (PLSA) (Hofmann, 1999) using only sentiment words in a set of movie reviews. Sentiment words, also called opinion words, are words in a language that indicate desirable or undesirable states. For example, *good*, *great*, and *beautiful* are positive sentiment words, and *bad*, *awful*, and *dreadful* are negative sentiment words. The second step builds an autoregressive model employing both the revenues and sentiment topics in the past few days to predict future revenues. This same revenue prediction problem was also attempted in Asur and Huberman (2010) using both the tweet volume and the tweet sentiment. A linear regression–based approach using movie review text and movie meta-data was reported in Joshi et al. (2010). My own group also used tweet sentiment to predict movie revenues several years ago and found that they could be predicted fairly easily and accurately. We simply applied our

Opinion Parser system to identify and combine positive and negative opinions about each movie and user intentions to watch it. No additional model or algorithm was used.

Several researchers have also analyzed sentiments of public opinions in the context of electoral politics. For example, in O'Connor et al. (2010), a sentiment score was computed based simply on counting positive and negative sentiment words, which was shown to correlate well with presidential approval, political election polls, and consumer confidence surveys. In Bermingham and Smeaton (2011), tweet volume and positive and negative tweets were utilized as the independent variables and polling results as values for the dependent variable to train a linear regression model to predict election results. In Chung and Mustafaraj (2011) and Gayo-Avello et al. (2011), several limitations of current works on using Twitter data to predict political elections were discussed, one of them being poor sentiment analysis accuracy. The works in Diakopoulos and Shamma (2010) and Sang and Bos (2012) used manually annotated sentiments of tweets for election prediction. Tumasjan et al. (2010) even showed that simple party mentions on Twitter can be a good predictor of election results. In other related works, Yano and Smith (2010) reported a method for predicting comment volumes of political blogs, Chen et al. (2010) studied political standpoints, and Khoo et al. (2012) analyzed sentiment in political news articles about economic policies and political figures.

Another popular application area is stock market prediction. Das and Chen (2007) identified opinions from message board posts by classifying each post into one of three sentiment classes: bullish (optimistic), bearish (pessimistic), or neutral (neither bullish nor bearish). The resulting sentiments across all stocks were then aggregated and used to predict the Morgan Stanley High-Tech Index. Instead of using bullish and bearish sentiments, Zhang et al. (2010) identified positive and negative public moods on Twitter and used them to predict the movement of stock market indices such as the Dow Jones, S&P 500, and NASDAQ. They showed that when emotions on Twitter fly high, that is, when people express a lot of hope, fear, or worry, the Dow goes down the next day. When people have less hope, fear, or worry, the Dow goes up. Along a similar line, Bollen et al. (2011) used Twitter moods to predict the movement of the Dow Jones Industrial Average (DJIA). In particular, the authors analyzed the text content of tweets to generate a six-dimensional daily time series of public mood: calm, alert, sure, vital, kind, and happy. The resulting mood time series were correlated with the DJIA to assess their ability to predict changes in the DJIA over time. Their results indicate that the accuracy of standard stock market prediction models can be significantly improved when certain mood dimensions are included, that is, calm and happiness, but not others. Instead of treating sentiments from all relevant Twitter authors equally, Bar-Haim et al. (2011) identified expert investors based on their past predictions of bullish and bearish stocks. Such expert investors are then used as one of the features in training stock price movement predictors. Feldman et al. (2011) reported a focused investigation of sentiment analysis of stock-related articles. Zhang and Skiena (2010) used blog and news sentiment to design trading strategies. Si et al. (2013) combined a topic-based sentiment time series and the index time series to predict the S&P 100 index's daily movements using vector autoregression. The topic-based sentiment analysis system first uses a nonparametric topic model to identify daily topics related to stocks and then computes people's sentiments about these topics.

In addition to research in the preceding three popular application areas, numerous papers have also been published on using sentiment analysis to help other types of applications. For example, in McGlohon et al. (2010), product reviews were used to rank products and merchants. In Hong and Skiena (2010), the relationships between the National Football League betting line and public opinions in blogs and on Twitter were studied. In Miller et al. (2011), sentiment flow in social networks was investigated. In Mohammad and Yang (2011), sentiments in males were used to find how genders differed on emotional axes. In Mohammad (2011), emotions in novels and fairy tales were tracked. In Sakunkoo and Sakunkoo (2009), social influences in online book reviews were studied, and in Groh and Hauffa (2011), sentiment analysis was used to characterize social relations. A deployed general-purpose sentiment analysis system and some case studies were reported in Castellanos et al. (2011).

1.2 Sentiment Analysis Research

Pervasive real-life applications provided strong motivations for research, but applications alone are not enough to generate strong research interests in academia. Researchers also need challenging technical problems. Sentiment analysis has provided plenty of such problems, most of which had not been attempted before, either in the NLP or linguistics communities. The novelty factor coupled with widespread applications and the availability of social media data attracted numerous researchers to the field. Since the year 2000, the field has grown rapidly to become one of the most active research areas in NLP, data mining, and web mining and is also widely studied in management sciences (Hu et al., 2006; Archak et al., 2007; Das and Chen, 2007; Dellarocas et al., 2007; Ghose et al., 2007; Park et al., 2007; Chen and Xie, 2008). Although sentiment analysis has been studied in different disciplines, their focuses are not the same. For example, in management science, the main focus is on the impact of consumer opinions on businesses and how to exploit such opinions to enhance business practices. However, for NLP and data mining, the objective is to design effective algorithms and models to extract opinions from natural language text and to summarize them suitably.

In terms of natural language understanding, sentiment analysis can be regarded as an important subarea of semantic analysis because its goal is to recognize topics that people talk about and their sentiments toward the topics. In the next few subsections, I briefly describe the key research topics covered in this book and also connect sentiment analysis with some general NLP tasks.

1.2.1 Different Levels of Analysis

Sentiment analysis research has been mainly carried out at three levels of granularity: document level, sentence level, and aspect level. We briefly introduce them here.

Document level. The task at the document level is to classify whether a whole opinion document expresses a positive or negative sentiment (Pang et al., 2002; Turney, 2002). It is thus known as *document-level sentiment classification*. For example, given a product review, the system determines whether the review expresses an overall positive or negative opinion about the product. This level of analysis implicitly assumes that each document expresses opinions on a single entity (e.g., a single product or service). Thus it is not applicable to documents that evaluate or compare multiple entities, for which more fine-grained analysis is needed. We study document-level sentiment analysis in Chapter 3.

Sentence level. The next level is to determine whether each sentence expresses a positive, negative, or neutral opinion. Note that "neutral opinion" usually means "no opinion." This level of analysis is closely related to *subjectivity classification* (Wiebe et al., 1999), which distinguishes sentences that express factual information (called *objective sentences*) from sentences that express subjective views and opinions (called *subjective sentences*). However, subjectivity is not equivalent to sentiment or opinion because, as we discussed earlier, many objective sentences can imply sentiments or opinions, for example, "*We bought the car last month and the windshield wiper has fallen off*." Conversely, many subjective sentences may not express any opinion or sentiment, for example, "*I think he went home after lunch*." We study sentence-level sentiment analysis in Chapter 4.

Aspect level. Neither document-level nor sentence-level analyses discover what people like and dislike exactly. In other words, they do not tell what each opinion is about, that is, the target of opinion. For example, if we only know that the sentence "*I like the iPhone 5*" is positive, it is of limited use unless we know that the positive opinion is about the *iPhone 5*. One may say that if we can classify a sentence to be positive, everything in the sentence can take the positive opinion. However, that will not work either, because a sentence can have multiple opinions, for example, "*Apple is doing very well in this poor economy*." It does not make much sense to classify this sentence as positive or negative because it is positive about *Apple* but negative about *economy*. To obtain this level of fine-grained results, we need to go to the aspect level. This level of analysis was earlier called *feature level,* as in *feature-based opinion mining and summarization* (Hu and Liu, 2004; Liu, 2010), which is now called *aspect-based sentiment analysis*. Instead of looking at language units (documents, paragraphs, sentences, clauses, or phrases), aspect-level analysis directly looks at opinion and its target (called *opinion target*). Realizing the importance of opinion targets allows us to have a much better understanding of the sentiment analysis problem. Let us see another example sentence: "*Although the service is not great, I still love this restaurant*." This sentence clearly has a positive tone, but we cannot say that this sentence is entirely positive. We can only say that the sentence is positive about the *restaurant* (emphasized), but it is still negative about its *service* (not emphasized). If someone reading the opinion cares a lot about the service, he probably will not go to eat at the restaurant. In applications, opinion targets (e.g., *restaurant* and *service* in the preceding sentence) are often described by entities (e.g., *restaurant*) and/or their different aspects (e.g., *service* of the restaurant). Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects. On the basis of this level of analysis, a summary of opinions about entities and their aspects can be produced. We study aspect-level sentiment analysis in Chapters 5 and 6. Note that in some applications, the user may only be interested in opinions about entities. In that case, the system can just ignore its aspects. Aspect-level analysis is what is needed in applications, and almost all real-life sentiment analysis systems in industry are based on this level of analysis.

Besides different levels of analysis, there are two different types of opinions, that is, *regular opinions* and *comparative opinions* (Jindal and Liu, 2006b):

- A regular opinion expresses a sentiment about a particular entity or an aspect of the entity, for example, "*Coke tastes very good*" expresses a positive sentiment or opinion on the aspect *taste* of *Coke*. This is the most common type of opinion.
- A comparative opinion compares multiple entities based on some of their shared aspects, for example, "*Coke tastes better than Pepsi*" compares *Coke* and *Pepsi* based on their tastes (an aspect) and expresses a preference for *Coke* (see Chapter 8).

Along with these basic tasks, researchers have also studied opinion summarization and opinion search, which we study in Chapter 9.

1.2.2 Sentiment Lexicon and Its Issues

Not surprisingly, the most important indicators of sentiments are *sentiment words*, also called *opinion words*. 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, for example, *cost an arm and a leg*. Sentiment words and phrases are instrumental to sentiment analysis. 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 discuss these algorithms in Chapter 7.

Although sentiment words and phrases are important, they are far from sufficient for accurate sentiment analysis. The problem is much more complex. We highlight several issues in the following:

1. A positive or negative sentiment word may have opposite *orientations* or *polarities* in different application domains or sentence contexts. By