Foundations and Trends® in
Information Retrieval
Vol. 2, Nos. 1–2 (2008) 1–135

DOI: 10.1561/1500000001

now
the essence of knowledge

Opinion Mining and Sentiment Analysis

Bo Pang[1] and Lillian Lee[2]

[1] *Yahoo! Research, 701 First Avenue, Sunnyvale, CA 94089, USA, bopang@yahoo-inc.com*
[2] *Computer Science Department, Cornell University, Ithaca, NY 14853, USA, llee@cs.cornell.edu*

Abstract

An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and challenges arise as people now can, and do, actively use information technologies to seek out and understand the opinions of others. The sudden eruption of activity in the area of opinion mining and sentiment analysis, which deals with the computational treatment of opinion, sentiment, and subjectivity in text, has thus occurred at least in part as a direct response to the surge of interest in new systems that deal directly with opinions as a first-class object.

This survey covers techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. Our focus is on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis. We include material

on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.

1

Introduction

> Romance should never begin with sentiment. It should begin with science and end with a settlement.
>
> — Oscar Wilde, *An Ideal Husband*

1.1 The Demand for Information on Opinions and Sentiment

"What other people think" has always been an important piece of information for most of us during the decision-making process. Long before awareness of the World Wide Web became widespread, many of us asked our friends to recommend an auto mechanic or to explain who they were planning to vote for in local elections, requested reference letters regarding job applicants from colleagues, or consulted *Consumer Reports* to decide what dishwasher to buy. But the Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics — that is, people we have never heard of. And conversely, more and more people are making their opinions available to strangers via the Internet.

Indeed, according to two surveys of more than 2000 American adults each [63, 127],

- 81% of Internet users (or 60% of Americans) have done online research on a product at least once;
- 20% (15% of all Americans) do so on a typical day;
- among readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73% and 87% report that reviews had a significant influence on their purchase;[1]
- consumers report being willing to pay from 20% to 99% more for a 5-star-rated item than a 4-star-rated item (the variance stems from what type of item or service is considered);
- 32% have provided a rating on a product, service, or person via an online ratings system, and 30% (including 18% of online senior citizens) have posted an online comment or review regarding a product or service.[2]

We hasten to point out that consumption of goods and services is not the only motivation behind people's seeking out or expressing opinions online. A need for political information is another important factor. For example, in a survey of over 2500 American adults, Rainie and Horrigan [248] studied the 31% of Americans — over 60 million people — that were 2006 *campaign internet users*, defined as those who gathered information about the 2006 elections online and exchanged views via email. Of these,

- 28% said that a major reason for these online activities was to get perspectives from within their community, and 34% said that a major reason was to get perspectives from outside their community;
- 27% had looked online for the endorsements or ratings of external organizations;

[1] Section 6.1 discusses quantitative analyses of actual economic impact, as opposed to consumer perception.

[2] Interestingly, Hitlin and Rainie [123] report that "Individuals who have rated something online are also more skeptical of the information that is available on the Web."

- 28% said that most of the sites they use share their point of view, but 29% said that most of the sites they use challenge their point of view, indicating that many people are not simply looking for validations of their pre-existing opinions; and
- 8% posted their own political commentary online.

The user hunger for and reliance upon online advice and recommendations that the data above reveals is merely one reason behind the surge of interest in new systems that deal directly with opinions as a first-class object. But, Horrigan [127] reports that while a majority of American internet users report positive experiences during online product research, at the same time, 58% also report that online information was missing, impossible to find, confusing, and/or overwhelming. Thus, there is a clear need to aid consumers of products and of information by building better information-access systems than are currently in existence.

The interest that individual users show in online opinions about products and services, and the potential influence such opinions wield, is something that vendors of these items are paying more and more attention to [124]. The following excerpt from a whitepaper is illustrative of the envisioned possibilities, or at the least the rhetoric surrounding the possibilities:

> With the explosion of Web 2.0 platforms such as blogs, discussion forums, peer-to-peer networks, and various other types of social media ... consumers have at their disposal a soapbox of unprecedented reach and power by which to share their brand experiences and opinions, positive or negative, regarding any product or service. As major companies are increasingly coming to realize, these consumer voices can wield enormous influence in shaping the opinions of other consumers — and, ultimately, their brand loyalties, their purchase decisions, and their own brand advocacy. ... Companies can respond to the consumer insights they generate through social media monitoring and analysis by modifying their

> marketing messages, brand positioning, product development, and other activities accordingly.
>
> — Zabin and Jefferies [327]

But industry analysts note that the leveraging of new media for the purpose of tracking product image requires new technologies; here is a representative snippet describing their concerns:

> Marketers have always needed to monitor media for information related to their brands — whether it's for public relations activities, fraud violations,[3] or competitive intelligence. But fragmenting media and changing consumer behavior have crippled traditional monitoring methods. Technorati estimates that 75,000 new blogs are created daily, along with 1.2 million new posts each day, many discussing consumer opinions on products and services. Tactics [of the traditional sort] such as clipping services, field agents, and ad hoc research simply can't keep pace.
>
> — Kim [154]

Thus, aside from individuals, an additional audience for systems capable of automatically analyzing consumer sentiment, as expressed in no small part in online venues, are companies anxious to understand how their products and services are perceived.

1.2 What Might be Involved? An Example Examination of the Construction of an Opinion/Review Search Engine

Creating systems that can process subjective information effectively requires overcoming a number of novel challenges. To illustrate some of these challenges, let us consider the concrete example of what building an *opinion- or review-search* application could involve. As we have discussed, such an application would fill an important and prevalent

[3] Presumably, the author means "*the detection or prevention of* fraud violations," as opposed to the *commission* thereof.

information need, whether one restricts attention to blog search [213] or considers the more general types of search that have been described above.

The development of a complete review- or opinion-search application might involve attacking each of the following problems.

(1) If the application is integrated into a general-purpose search engine, then one would need to determine whether the user is in fact looking for subjective material. This may or may not be a difficult problem in and of itself: perhaps queries of this type will tend to contain indicator terms like "review," "reviews," or "opinions," or perhaps the application would provide a "checkbox" to the user so that he or she could indicate directly that reviews are what is desired; but in general, query classification is a difficult problem — indeed, it was the subject of the 2005 KDD Cup challenge [185].
(2) Besides the still-open problem of determining which documents are topically relevant to an opinion-oriented query, an additional challenge we face in our new setting is simultaneously or subsequently determining which documents or portions of documents contain review-like or opinionated material. Sometimes this is relatively easy, as in texts fetched from review-aggregation sites in which review-oriented information is presented in relatively stereotyped format: examples include Epinions.com and Amazon.com. However, blogs also notoriously contain quite a bit of subjective content and thus are another obvious place to look (and are more relevant than shopping sites for queries that concern politics, people, or other non-products), but the desired material within blogs can vary quite widely in content, style, presentation, and even level of grammaticality.
(3) Once one has target documents in hand, one is still faced with the problem of identifying the overall sentiment expressed by these documents and/or the specific opinions regarding particular features or aspects of the items or topics in question, as necessary. Again, while some sites make this

kind of extraction easier — for instance, user reviews posted to Yahoo! Movies must specify grades for pre-defined sets of characteristics of films — more free-form text can be much harder for computers to analyze, and indeed can pose additional challenges; for example, if quotations are included in a newspaper article, care must be taken to attribute the views expressed in each quotation to the correct entity.

(4) Finally, the system needs to present the sentiment information it has garnered in some reasonable summary fashion. This can involve some or all of the following actions:

   (a) Aggregation of "votes" that may be registered on different scales (e.g., one reviewer uses a star system, but another uses letter grades).
   (b) Selective highlighting of some opinions.
   (c) Representation of points of disagreement and points of consensus.
   (d) Identification of communities of opinion holders.
   (e) Accounting for different levels of authority among opinion holders.

Note that it might be more appropriate to produce a visualization of sentiment data rather than a textual summary of it, whereas textual summaries are what is usually created in standard topic-based multi-document summarization.

1.3 Our Charge and Approach

Challenges (2), (3), and (4) in the above list are very active areas of research, and the bulk of this survey is devoted to reviewing work in these three sub-fields. However, due to space limitations and the focus of the journal series in which this survey appears, we do not and cannot aim to be completely comprehensive.

In particular, when we began to write this survey, we were directly charged to focus on information-access applications, as opposed to work of more purely linguistic interest. We stress that the importance of work in the latter vein is absolutely not in question.

Given our mandate, the reader will not be surprised that we describe the applications that sentiment-analysis systems can facilitate and review many kinds of approaches to a variety of opinion-oriented classification problems. We have also chosen to attempt to draw attention to single- and multi-document summarization of evaluative text, especially since interesting considerations regarding graphical visualization arise. Finally, we move beyond just the technical issues, devoting significant attention to the broader implications that the development of opinion-oriented information-access services have: we look at questions of privacy, manipulation, and whether or not reviews can have measurable economic impact.

1.4 Early History

Although the area of sentiment analysis and opinion mining has recently enjoyed a huge burst of research activity, there has been a steady undercurrent of interest for quite a while. One could count early projects on beliefs as forerunners of the area [48, 317]. Later work focused mostly on interpretation of metaphor, narrative, point of view, affect, evidentiality in text, and related areas [121, 133, 149, 262, 306, 310, 311, 312, 313].

The year 2001 or so seems to mark the beginning of widespread awareness of the research problems and opportunities that sentiment analysis and opinion mining raise [51, 66, 69, 79, 192, 215, 221, 235, 291, 296, 298, 305, 326], and subsequently there have been literally hundreds of papers published on the subject.

Factors behind this "land rush" include:

- the rise of machine learning methods in natural language processing and information retrieval;
- the availability of datasets for machine learning algorithms to be trained on, due to the blossoming of the World Wide Web and, specifically, the development of review-aggregation web-sites; and, of course
- realization of the fascinating intellectual challenges and commercial and intelligence applications that the area offers.

1.5 A Note on Terminology: Opinion Mining, Sentiment Analysis, Subjectivity, and All that

> 'The beginning of wisdom is the definition of terms,' wrote Socrates. The aphorism is highly applicable when it comes to the world of social media monitoring and analysis, where any semblance of universal agreement on terminology is altogether lacking.
>
> Today, vendors, practitioners, and the media alike call this still-nascent arena everything from 'brand monitoring,' 'buzz monitoring' and 'online anthropology,' to 'market influence analytics,' 'conversation mining' and 'online consumer intelligence'. ... In the end, the term 'social media monitoring and analysis' is itself a verbal crutch. It is placeholder [sic], to be used until something better (and shorter) takes hold in the English language to describe the topic of this report.
>
> — Zabin and Jefferies [327]

The above quotation highlights the problems that have arisen in trying to name a new area. The quotation is particularly apt in the context of this survey because the field of "social media monitoring and analysis" (or however one chooses to refer to it) is precisely one that the body of work we review is very relevant to. And indeed, there has been to date no uniform terminology established for the relatively young field we discuss in this survey. In this section, we simply mention some of the terms that are currently in vogue, and attempt to indicate what these terms tend to mean in research papers that the interested reader may encounter.

The body of work we review is that which deals with the computational treatment of (in alphabetical order) *opinion*, *sentiment*, and *subjectivity* in text. Such work has come to be known as *opinion mining*, *sentiment analysis*, and/or *subjectivity analysis*. The phrases *review mining* and *appraisal extraction* have been used, too, and there are some connections to *affective computing*, where the goals include enabling computers to recognize and express emotions [239]. This proliferation of terms reflects differences in the connotations that these terms carry,

both in their original general-discourse usages[4] and in the usages that have evolved in the technical literature of several communities.

In 1994, Wiebe [311], influenced by the writings of the literary theorist Banfield [26], centered the idea of *subjectivity* around that of *private states*, defined by Quirk et al. [245] as states that are not open to objective observation or verification. Opinions, evaluations, emotions, and speculations all fall into this category; but a canonical example of research typically described as a type of subjectivity analysis is the recognition of opinion-oriented language in order to distinguish it from objective language. While there has been some research self-identified as subjectivity analysis on the particular application area of determining the value judgments (e.g., "four stars" or "C+") expressed in the evaluative opinions that are found, this application has not tended to be a major focus of such work.

The term *opinion mining* appears in a paper by Dave et al. [69] that was published in the proceedings of the 2003 WWW conference; the publication venue may explain the popularity of the term within communities strongly associated with Web search or information retrieval. According to Dave et al. [69], the ideal opinion-mining tool would "process a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions

[4] To see that the distinctions in common usage can be subtle, consider how interrelated the following set of definitions given in *Merriam-Webster's Online Dictionary* are:

> Synonyms: opinion, view, belief, conviction, persuasion, sentiment mean a judgment one holds as true.
>
> - Opinion implies a conclusion thought out yet open to dispute ⟨each expert seemed to have a different opinion⟩.
> - View suggests a subjective opinion ⟨very assertive in stating his views⟩.
> - Belief implies often deliberate acceptance and intellectual assent ⟨a firm belief in her party's platform⟩.
> - Conviction applies to a firmly and seriously held belief ⟨the conviction that animal life is as sacred as human⟩.
> - Persuasion suggests a belief grounded on assurance (as by evidence) of its truth ⟨was of the persuasion that everything changes⟩.
> - Sentiment suggests a settled opinion reflective of one's feelings ⟨her feminist sentiments are well-known⟩.

about each of them (poor, mixed, good)." Much of the subsequent research self-identified as opinion mining fits this description in its emphasis on extracting and analyzing judgments on various aspects of given items. However, the term has recently also been interpreted more broadly to include many different types of analysis of evaluative text [190].

The history of the phrase *sentiment analysis* parallels that of "opinion mining" in certain respects. The term "sentiment" used in reference to the automatic analysis of evaluative text and tracking of the predictive judgments therein appears in 2001 papers by Das and Chen [66] and Tong [296], due to these authors' interest in analyzing market sentiment. It subsequently occurred within 2002 papers by Turney [298] and Pang et al. [235], which were published in the proceedings of the annual meeting of the Association for Computational Linguistics (ACL) and the annual conference on Empirical Methods in Natural Language Processing (EMNLP). Moreover, Nasukawa and Yi [221] entitled their 2003 paper, "Sentiment analysis: Capturing favorability using natural language processing", and a paper in the same year by Yi et al. [323] was named "Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques." These events together may explain the popularity of "sentiment analysis" among communities self-identified as focused on NLP. A sizeable number of papers mentioning "sentiment analysis" focus on the specific application of classifying reviews as to their polarity (either positive or negative), a fact that appears to have caused some authors to suggest that the phrase refers specifically to this narrowly defined task. However, nowadays many construe the term more broadly to mean the computational treatment of opinion, sentiment, and subjectivity in text.

Thus, when broad interpretations are applied, "sentiment analysis" and "opinion mining" denote the same field of study (which itself can be considered a sub-area of subjectivity analysis). We have attempted to use these terms more or less interchangeably in this survey. This is in no small part because we view the field as representing a unified body of work, and would thus like to encourage researchers in the area to share terminology regardless of the publication venues at which their papers might appear.

2

Applications

> Sentiment without action is the ruin of the soul.
> — Edward Abbey

We used one application of opinion mining and sentiment analysis as a motivating example in the Introduction, namely, web search targeted toward reviews. But other applications abound. In this section, we seek to enumerate some of the possibilities.

It is important to mention that because of all the possible applications, there are a good number of companies, large and small, that have opinion mining and sentiment analysis as part of their mission. However, we have elected not to mention these companies individually due to the fact that the industrial landscape tends to change quite rapidly, so that lists of companies risk falling out of date rather quickly.

2.1 Applications to Review-Related Websites

Clearly, the same capabilities that a review-oriented search engine would have could also serve very well as the basis for the creation and automated upkeep of review- and opinion-aggregation websites. That is, as an alternative to sites like Epinions that solicit feedback and reviews,

one could imagine sites that proactively gather such information. Topics need not be restricted to product reviews, but could include opinions about candidates running for office, political issues, and so forth.

There are also applications of the technologies we discuss to more traditional review-solicitation sites, as well. Summarizing user reviews is an important problem. One could also imagine that errors in user ratings could be fixed: there are cases where users have clearly accidentally selected a low rating when their review indicates a positive evaluation [47]. Moreover, as discussed later in this survey (see Section 5.2.4, for example), there is some evidence that user ratings can be biased or otherwise in need of correction, and automated classifiers could provide such updates.

2.2 Applications as a Sub-Component Technology

Sentiment-analysis and opinion-mining systems also have an important potential role as enabling technologies for other systems.

One possibility is as an augmentation to *recommendation systems* [292, 293], since it might behoove such a system not to recommend items that receive a lot of negative feedback.

Detection of "flames" (overly heated or antagonistic language) in email or other types of communication [276] is another possible use of subjectivity detection and classification.

In online systems that display ads as sidebars, it is helpful to detect webpages that contain sensitive content inappropriate for ads placement [137]; for more sophisticated systems, it could be useful to bring up product ads when relevant positive sentiments are detected, and perhaps more importantly, nix the ads when relevant negative statements are discovered.

It has also been argued that information extraction can be improved by discarding information found in subjective sentences [256].

Question answering is another area where sentiment analysis can prove useful [274, 284, 189]. For example, opinion-oriented questions may require different treatment. Alternatively, Lita et al. [189] suggest that for definitional questions, providing an answer that includes more information about how an entity is viewed may better inform the user.

Summarization may also benefit from accounting for multiple viewpoints [265].

Additionally, there are potentially relations to citation analysis, where, for example, one might wish to determine whether an author is citing a piece of work as supporting evidence or as research that he or she dismisses [238]. Similarly, one effort seeks to use semantic orientation to track literary reputation [287].

In general, the computational treatment of affect has been motivated in part by the desire to improve human–computer interaction [188, 192, 295].

2.3 Applications in Business and Government Intelligence

The field of opinion mining and sentiment analysis is well-suited to various types of intelligence applications. Indeed, business intelligence seems to be one of the main factors behind corporate interest in the field.

Consider, for instance, the following scenario (the text of which also appears in Lee [181]). A major computer manufacturer, disappointed with unexpectedly low sales, finds itself confronted with the question: "Why aren't consumers buying our laptop?" While concrete data such as the laptop's weight or the price of a competitor's model are obviously relevant, answering this question requires focusing more on people's personal views of such objective characteristics. Moreover, subjective judgments regarding intangible qualities — e.g., "the design is tacky" or "customer service was condescending" — or even misperceptions — e.g., "updated device drivers are not available" when such device drivers do in fact exist — must be taken into account as well.

Sentiment-analysis technologies for extracting opinions from unstructured human-authored documents would be excellent tools for handling many business-intelligence tasks related to the one just described. Continuing with our example scenario: it would be difficult to try to directly survey laptop purchasers who have not bought the company's product. Rather, we could employ a system that (a) finds reviews or other expressions of opinion on the Web — newsgroups, individual blogs, and aggregation sites such as Epinions are likely to

be productive sources — and then (b) creates condensed versions of individual reviews or a digest of overall consensus points. This would save an analyst from having to read potentially dozens or even hundreds of versions of the same complaints. Note that Internet sources can vary wildly in form, tenor, and even grammaticality; this fact underscores the need for robust techniques even when only one language (e.g., English) is considered.

Besides reputation management and public relations, one might perhaps hope that by tracking public viewpoints, one could perform trend prediction in sales or other relevant data [214]. (See our discussion of *Broader Implications* (Section 6) for more discussion of potential economic impact.)

Government intelligence is another application that has been considered. For example, it has been suggested that one could monitor sources for increases in hostile or negative communications [1].

2.4 Applications Across Different Domains

One exciting turn of events has been the confluence of interest in opinions and sentiment within computer science with interest in opinions and sentiment in other fields.

As is well known, opinions matter a great deal in politics. Some work has focused on understanding what voters are thinking [83, 110, 126, 178, 219], whereas other projects have as a long term goal the clarification of politicians' positions, such as what public figures support or oppose, to enhance the quality of information that voters have access to [27, 111, 294].

Sentiment analysis has specifically been proposed as a key enabling technology in eRulemaking, allowing the automatic analysis of the opinions that people submit about pending policy or government-regulation proposals [50, 175, 271].

On a related note, there has been investigation into opinion mining in weblogs devoted to legal matters, sometimes known as "blawgs" [64].

Interactions with sociology promise to be extremely fruitful. For instance, the issue of how ideas and innovations diffuse [258] involves the question of who is positively or negatively disposed toward whom,

and hence who would be more or less receptive to new information transmission from a given source. To take just one other example: *structural balance theory* is centrally concerned with the polarity of "ties" between people [54] and how this relates to group cohesion. These ideas have begun to be applied to online media analysis [58, 144].

3

General Challenges

3.1 Contrasts with Standard Fact-Based Textual Analysis

The increasing interest in opinion mining and sentiment analysis is partly due to its potential applications, which we have just discussed. Equally important are the new intellectual challenges that the field presents to the research community. So what makes the treatment of evaluative text different from "classic" text mining and fact-based analysis?

Take text categorization, for example. Traditionally, text categorization seeks to classify documents by topic. There can be many possible categories, the definitions of which might be user- and application-dependent; and for a given task, we might be dealing with as few as two classes (binary classification) or as many as thousands of classes (e.g., classifying documents with respect to a complex taxonomy). In contrast, with sentiment classification (see Section 4.1 for more details on precise definitions), we often have relatively few classes (e.g., "positive" or "3 stars") that generalize across many domains and users. In addition, while the different classes in topic-based categorization can be completely unrelated, the sentiment labels that are widely

considered in previous work typically represent opposing (if the task is binary classification) or ordinal/numerical categories (if classification is according to a multi-point scale). In fact, the regression-like nature of strength of feeling, degree of positivity, and so on seems rather unique to sentiment categorization (although one could argue that the same phenomenon exists with respect to topic-based relevance).

There are also many characteristics of answers to opinion-oriented questions that differ from those for fact-based questions [284]. As a result, opinion-oriented information extraction, as a way to approach opinion-oriented question answering, naturally differs from traditional information extraction (IE) [49]. Interestingly, in a manner that is similar to the situation for the classes in sentiment-based classification, the templates for opinion-oriented IE also often generalize well across different domains, since we are interested in roughly the same set of fields for each opinion expression (e.g., holder, type, strength) regardless of the topic. In contrast, traditional IE templates can differ greatly from one domain to another — the typical template for recording information relevant to a natural disaster is very different from a typical template for storing bibliographic information.

These distinctions might make our problems appear deceptively simpler than their counterparts in fact-based analysis, but this is far from the truth. In the next section, we sample a few examples to show what makes these problems difficult compared to traditional fact-based text analysis.

3.2 Factors that Make Opinion Mining Difficult

Let us begin with a *sentiment polarity* text-classification example. Suppose we wish to classify an opinionated text as either positive or negative, according to the overall sentiment expressed by the author within it. Is this a difficult task?

To answer this question, first consider the following example, consisting of only one sentence (by Mark Twain): "Jane Austen's books madden me so that I can't conceal my frenzy from the reader." Just as the topic of this text segment can be identified by the phrase "Jane Austen," the presence of words like "madden" and "frenzy" suggests

negative sentiment. So one might think this is an easy task, and hypothesize that the polarity of opinions can generally be identified by a set of keywords.

But, the results of an early study by Pang et al. [235] on movie reviews suggest that coming up with the right set of keywords might be less trivial than one might initially think. The purpose of Pang et al.'s pilot study was to better understand the difficulty of the document-level sentiment-polarity classification problem. Two human subjects were asked to pick keywords that they would consider to be good indicators of positive and negative sentiment. As shown in Figure 3.1, the use of the subjects' lists of keywords achieves about 60% accuracy when employed within a straightforward classification policy. In contrast, word lists of the same size but chosen based on examination of the corpus' statistics achieves almost 70% accuracy — even though some of the terms, such as "still," might not look that intuitive at first.

However, the fact that it may be non-trivial for humans to come up with the best set of keywords does not in itself imply that the problem is harder than topic-based categorization. While the feature "still" might not be likely for any human to propose from introspection, given training data, its correlation with the positive class can be discovered via a data-driven approach, and its utility (at least in

	Proposed word lists	Accuracy (%)	Ties (%)
Human 1	positive: *dazzling, brilliant, phenomenal, excellent, fantastic* negative: *suck, terrible, awful, unwatchable, hideous*	58	75
Human 2	positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting* negative: *bad, cliched, sucks, boring, stupid, slow*	64	39
Statistics-based	positive: *love, wonderful, best, great, superb, still, beautiful* negative: *bad, worst, stupid, waste, boring, ?, !*	69	16

Fig. 3.1 Sentiment classification using keyword lists created by human subjects ("Human 1" and "Human 2"), with corresponding results using keywords selected via examination of simple statistics of the test data ("Statistics-based"). Adapted from Figures 1 and 2 in Pang et al. [235].

the movie review domain) does make sense in retrospect. Indeed, applying machine learning techniques based on unigram models can achieve over 80% in accuracy [235], which is much better than the performance based on hand-picked keywords reported above. However, this level of accuracy is not quite on par with the performance one would expect in typical topic-based binary classification.

Why does this problem appear harder than the traditional task when the two classes we are considering here are so different from each other? Our discussion of algorithms for classification and extraction (Section 4) will provide a more in-depth answer to this question, but the following are a few examples (from among the many we know) showing that the upper bound on problem difficulty, from the viewpoint of machines, is very high. Note that not all of the issues these examples raise have been fully addressed in the existing body of work in this area.

Compared to topic, sentiment can often be expressed in a more subtle manner, making it difficult to be identified by any of a sentence or document's terms when considered in isolation. Consider the following examples:

- "If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut." (review by Luca Turin and Tania Sanchez of the Givenchy perfume Amarige, in *Perfumes: The Guide*, Viking 2008.) No ostensibly negative words occur.
- "She runs the gamut of emotions from A to B." (Dorothy Parker, speaking about Katharine Hepburn.) No ostensibly negative words occur.

In fact, the example that opens this section, which was taken from the following quote from Mark Twain, is also followed by a sentence with no ostensibly negative words:

> Jane Austen's books madden me so that I can't conceal my frenzy from the reader. Everytime I read 'Pride and Prejudice' I want to dig her up and beat her over the skull with her own shin-bone.

A related observation is that although the second sentence indicates an extremely strong opinion, it is difficult to associate the presence of this strong opinion with specific keywords or phrases in this sentence. Indeed, subjectivity detection can be a difficult task in itself. Consider the following quote from Charlotte Brontë, in a letter to George Lewes:

> You say I must familiarise my mind with the fact that "Miss Austen is not a poetess, has no 'sentiment'" (you scornfully enclose the word in inverted commas), "has no eloquence, none of the ravishing enthusiasm of poetry"; and then you add, I must "learn to acknowledge her as one of the greatest artists, of the greatest painters of human character, and one of the writers with the nicest sense of means to an end that ever lived."

Note the fine line between facts and opinions: while "Miss Austen is not a poetess" can be considered to be a fact, "none of the ravishing enthusiasm of poetry" should probably be considered as an opinion, even though the two phrases s (arguably) convey similar information.[1] Thus, not only can we not easily identify simple keywords for subjectivity, but we also find that like "the fact that" do not necessarily guarantee the objective truth of what follows them — and bigrams like "no sentiment" apparently do not guarantee the absence of opinions, either. We can also get a glimpse of how opinion-oriented information

[1] One can challenge our analysis of the "poetess" clause, as an anonymous reviewer indeed did — which disagreement perhaps supports our greater point about the difficulties that can sometimes present themselves.

Different researchers express different opinions about whether distinguishing between subjective and objective language is difficult for humans in the general case. For example, Kim and Hovy [159] note that in a pilot study sponsored by NIST, "human annotators often disagreed on whether a belief statement was or was not an opinion." However, other researchers have found inter-annotator agreement rates in various types of subjectivity-classification tasks to be satisfactory [45, 273, 274, 309]; a summary provided by one of the anonymous referees is that "[although] there is variation from study to study, on average, about 85% of annotations are not marked as uncertain by either annotator, and for these cases, inter-coder agreement is very high (kappa values over 80)." As in other settings, more careful definitions of the distinctions to be made tend to lead to better agreement rates.

In any event, the points we are exploring in the Brontë quote may be made more clear by replacing "Jane Austen is not a poetess" with something like "Jane Austen does not write poetry for a living, but is also no poet in the broader sense."

extraction can be difficult. For instance, it is non-trivial to recognize opinion holders. In the example quoted above, the opinion is not that of the author, but the opinion of "You," which refers to George Lewes in this particular letter. Also, observe that given the context ("you scornfully enclose the word in inverted commas," together with the reported endorsement of Austen as a great artist), it is clear that "has no sentiment" is not meant to be a show-stopping criticism of Austen from Lewes, and Brontë's disagreement with him on this subject is also subtly revealed.

In general, sentiment and subjectivity are quite context-sensitive, and, at a coarser granularity, quite domain dependent (in spite of the fact that the general notion of positive and negative opinions is fairly consistent across different domains). Note that although domain dependency is in part a consequence of changes in vocabulary, even the exact same expression can indicate different sentiment in different domains. For example, "go read the book" most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews. (This example was furnished to us by Bob Bland.) We will discuss topic-sentiment interaction in more detail in Section 4.4.

It does not take a seasoned writer or a professional journalist to produce texts that are difficult for machines to analyze. The writings of Web users can be just as challenging, if not as subtle, in their own way — see Figure 3.2 for an example. In the case of Figure 3.2, it should be pointed out that might be more useful to learn to recognize the quality of a review (see Section 5.2 for more detailed discussions on that subject). Still, it is interesting to observe the importance of modeling discourse structure. While the overall topic of a document

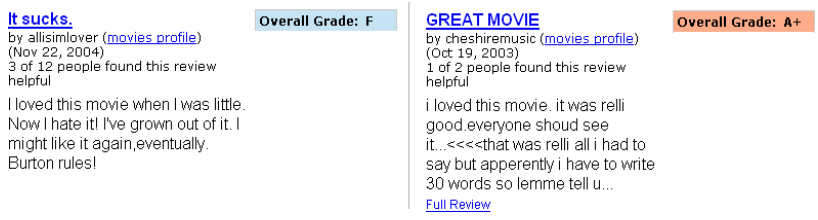

Fig. 3.2 Example of movie reviews produced by web users: a (slightly reformatted) screenshot of user reviews for *The Nightmare Before Christmas*.

should be what the majority of the content is focusing on regardless of the order in which potentially different subjects are presented, for opinions, the order in which different opinions are presented can result in a completely opposite overall sentiment polarity.

In fact, somewhat in contrast with topic-based text categorization, order effects can completely overwhelm frequency effects. Consider the following excerpt, again from a movie review:

> This film should be *brilliant*. It sounds like a *great* plot, the actors are *first grade*, and the supporting cast is *good* as well, and Stallone is attempting to deliver a *good* performance. However, it can't hold up.

As indicated by the (inserted) emphasis, words that are positive in orientation dominate this excerpt,[2] and yet the overall sentiment is negative because of the crucial last sentence; whereas in traditional text classification, if a document mentions "cars" relatively frequently, then the document is most likely at least somewhat related to cars.

Order dependence also manifests itself at more fine-grained levels of analysis: "A is better than B" conveys the exact opposite opinion from "B is better than A."[3] In general, modeling sequential information and discourse structure seems more crucial in sentiment analysis (further discussion appears in Section 4.7).

As noted earlier, not all of the issues we have just discussed have been fully addressed in the literature. This is perhaps part of the charm of this emerging area. In the following sections, we aim to give an overview of a selection of past heroic efforts to address some of these issues, and march through the positives and the negatives, charged with unbiased feeling, armed with hard facts.

> Fasten your seat belts. It's going to be a bumpy night!
>
> — Bette Davis, *All About Eve*,
> screenplay by Joseph Mankiewicz

[2] One could argue about whether in the context of movie reviews the word "Stallone" has a semantic orientation.

[3] Note that this is not unique to opinion expressions; "A killed B" and "B killed A" also convey different factual information.

4

Classification and Extraction

> "The Bucket List," which was written by Justin Zackham and directed by Rob Reiner, seems to have been created by applying algorithms to sentiment.
>
> — David Denby movie review,
> *The New Yorker*, January 7, 2007

A fundamental technology in many current opinion-mining and sentiment-analysis applications is *classification* — note that in this survey, we generally construe the term "classification" broadly, so that it encompasses regression and ranking. The reason that classification is so important is that many problems of interest can be formulated as applying classification/regression/ranking to given textual units; examples include making a decision for a particular phrase or document ("how positive is it?"), ordering a set of texts ("rank these reviews by how positive they are"), giving a single label to an entire document collection ("where on the scale between liberal and conservative do the writings of this author lie?"), and categorizing the relationship between two entities based on textual evidence ("does A approve of B's actions?"). This section is centered on approaches to these kinds of problems.

Part One (p. 24ff.) covers fundamental background. Specifically, Section 4.1 provides a discussion of key concepts involved in common formulations of classification problems in sentiment analysis and opinion mining. Features that have been explored for sentiment analysis tasks are discussed in Section 4.2.

Part Two (p. 37ff.) is devoted to an in-depth discussion of different types of approaches to classification, regression, and ranking problems. The beginning of Part Two should be consulted for a detailed outline, but it is appropriate here to indicate how we cover *extraction*, since it plays a key role in many sentiment-oriented applications and so some readers may be particularly interested in it.

First, extraction problems (e.g., retrieving opinions on various features of a laptop) are often solved by casting many sub-problems as classification problems (e.g., given a text span, determine whether it expresses any opinion at all). Therefore, rather than have a separate section devoted completely to the entirety of the extraction task, we have integrated discussion of extraction-oriented classification subproblems into the appropriate places in our discussion of different types of approaches to classification in general (Sections 4.3–4.8). Section 4.9 covers those remaining aspects of extraction that can be thought of as distinct from classification.

Second, extraction is often a means to the further goal of providing effective summaries of the extracted information to users. Details on how to combine information mined from multiple subjective text segments into a suitable summary can be found in Section 5.

Part One: Fundamentals

4.1 Problem Formulations and Key Concepts

Motivated by different real-world applications, researchers have considered a wide range of problems over a variety of different types of corpora. We now examine the key concepts involved in these problems. This discussion also serves as a loose grouping of the major problems, where each group consists of problems that are suitable for similar treatment as learning tasks.

4.1.1 Sentiment Polarity and Degrees of Positivity

One set of problems share the following general character: given an opinionated piece of text, wherein it is assumed that the overall opinion in it is about one single issue or item, classify the opinion as falling under one of two opposing sentiment polarities, or locate its position on the continuum between these two polarities. A large portion of work in sentiment-related classification/regression/ranking falls within this category. Eguchi and Lavrenko [84] point out that the polarity or positivity labels so assigned may be used simply for summarizing the content of opinionated text units on a topic, whether they be positive or negative, or for only retrieving items of a given sentiment orientation (say, positive).

The binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion is called *sentiment polarity classification* or *polarity classification*. Although this binary decision task has also been termed *sentiment classification* in the literature, as mentioned above, in this survey we will use "sentiment classification" to refer broadly to binary categorization, multi-class categorization, regression, and/or ranking.

Much work on sentiment polarity classification has been conducted in the context of reviews (e.g., "thumbs up" or "thumbs down" for movie reviews). While in this context "positive" and "negative" opinions are often evaluative (e.g., "like" vs. "dislike"), there are other problems where the interpretation of "positive" and "negative" is subtly different. One example is determining whether a political speech is in support of or opposition to the issue under debate [27, 294]; a related task is classifying predictive opinions in election forums into "likely to win" and "unlikely to win" [160]. Since these problems are all concerned with two opposing subjective classes, as machine learning tasks they are often amenable to similar techniques. Note that a number of other aspects of politically oriented text, such as whether liberal or conservative views are expressed, have been explored; since the labels used in those problems can usually be considered properties of a set of documents representing authors' attitudes over multiple issues rather than positive or negative sentiment with respect to a single issue, we

discuss them under a different heading further below ("viewpoints and perspectives," Section 4.1.4).

The input to a sentiment classifier is not necessarily always strictly opinionated. Classifying a news article into good or bad news has been considered a sentiment classification task in the literature [168]. But a piece of news can be good or bad news without being subjective (i.e., without being expressive of the private states of the author): for instance, "the stock price rose" is objective information that is generally considered to be good news in appropriate contexts. It is not our main intent to provide a clean-cut definition for what should be considered "sentiment polarity classification" problems,[1] but it is perhaps useful to point out that (a) in determining the sentiment polarity of opinionated texts where the authors do explicitly express their sentiment through statements like "this laptop is great," (arguably) objective information such as "long battery life"[2] is often used to help determine the overall sentiment; (b) the task of determining whether a piece of *objective* information is good or bad is still not quite the same as classifying it into one of several topic-based classes, and hence inherits the challenges involved in sentiment analysis; and (c) as we will discuss in more detail later, the distinction between subjective and objective information can be subtle. Is "long battery life" objective? Also consider the difference between "the battery lasts 2 hours" vs. "the battery only lasts 2 hours."

Related categories. An alternative way of summarizing reviews is to extract information on why the reviewers liked or disliked the product. Kim and Hovy [158] note that such "pro and con" expressions can differ from positive and negative opinion expressions, although the two concepts — opinion ("I think this laptop is terrific") and reason for opinion ("This laptop only costs $399") — are for the purposes of analyzing evaluative text strongly related. In addition to potentially forming the basis for the production of more informative sentiment-oriented summaries, identifying pro and con reasons can potentially be used to

[1] While it is of utter importance that the problem itself should be well-defined, it is of less, if any, importance to decide which tasks should be labeled as "polarity classification" problems.

[2] Whether this should be considered as an objective statement may be up for debate: one can imagine another reviewer retorting, "you call that *long* battery life?"

help decide the helpfulness of individual reviews: evaluative judgments that are supported by reasons are likely to be more trustworthy.

Another type of categorization related to degrees of positivity is considered by Niu et al. [226], who seek to determine the polarity of outcomes (improvement vs. death, say) described in medical texts.

Additional problems related to the determination of degree of positivity surround the analysis of comparative sentences [139]. The main idea is that sentences such as "The new model is more expensive than the old one" or "I prefer the new model to the old model" are important sources of information regarding the author's evaluations.

Rating inference (*ordinal regression*). The more general problem of rating inference, where one must determine the author's evaluation with respect to a multi-point scale (e.g., one to five "stars" for a review) can be viewed simply as a multi-class text categorization problem. Predicting degree of positivity provides more fine-grained rating information; at the same time, it is an interesting learning problem in itself.

But in contrast to many topic-based multi-class classification problems, sentiment-related multi-class classification can also be naturally formulated as a regression problem because ratings are ordinal. It can be argued to constitute a special type of (ordinal) regression problem because the semantics of each class may not simply directly correspond to a point on a scale. More specifically, each class may have its own distinct vocabulary. For instance, if we are classifying an author's evaluation into one of the positive, neutral, and negative classes, an overall neutral opinion could be a mixture of positive and negative language, or it could be identified with signature words such as "mediocre." This presents us with interesting opportunities to explore the relationships between classes.

Note the difference between rating inference and predicting strength of opinion (discussed in Section 4.1.2); for instance, it is possible to feel quite strongly (high on the "strength" scale) that something is mediocre (middling on the "evaluation" scale).

Also, note that the label "neutral" is sometimes used as a label for the objective class ("lack of opinion") in the literature. In this survey, we use neutral only in the aforementioned sense of a sentiment that lies between positive and negative.

Interestingly, Cabral and Hortaçsu [47] observe that neutral comments in feedback systems are not necessarily perceived by users as lying at the exact mid-point between positive and negative comments; rather, "the information contained in a neutral rating is perceived by users to be much closer to negative feedback than positive." On the other hand, they also note that in their data, "sellers were less likely to retaliate against neutral comments, as opposed to negatives: ... a buyer leaving a negative comment has a 40% chance of being hit back, while a buyer leaving a neutral comment only has a 10% chance of being retaliated upon by the seller."

Agreement. The opposing nature of polarity classes also gives rise to exploration of *agreement detection*, e.g., given a pair of texts, deciding whether they should receive the same or differing sentiment-related labels based on the relationship between the elements of the pair. This is often not defined as a standalone problem but considered as a subtask whose result is used to improve the labeling of the opinions held by the entities involved [272, 294]. A different type of agreement task has also been considered in the context of perspectives, where, for example, a label of "conservative" tends to indicate agreement with particular positions on a wide variety of issues.

4.1.2 Subjectivity Detection and Opinion Identification

Work in polarity classification often assumes the incoming documents to be opinionated. For many applications, though, we may need to decide whether a given document contains subjective information or not, or identify which portions of the document are subjective. Indeed, this problem was the focus of the 2006 Blog track at TREC [227]. At least one opinion-tracking system rates subjectivity and sentiment separately [108]. Mihalcea et al. [209] summarize the evidence of several projects on subsentential analysis [12, 90, 289, 319] as follows: "the problem of distinguishing subjective versus objective instances has often proved to be more difficult than subsequent polarity classification, so improvements in subjectivity classification promise to positively impact sentiment classification."

Early work by Hatzivassiloglou and Wiebe [120] examined the effects of adjective orientation and gradability on sentence subjectivity. The goal was to tell whether a given sentence is subjective or not judging from the adjectives appearing in that sentence. A number of projects address sentence-level or sub-sentence-level subjectivity detection in different domains [33, 156, 232, 255, 308, 315, 319, 326].Wiebe et al. [316] present a comprehensive survey of subjectivity recognition using different clues and features.

Wilson et al. [320] address the problem of determining clause-level opinion strength (e.g., "how mad are you?"). Note that the problem of determining opinion strength is different from rating inference. Classifying a piece of text as expressing a neutral opinion (giving it a midpoint score) for rating inference does not equal classifying that piece of text as objective (lack of opinion): one can have a strong opinion that something is "mediocre" or "so-so."

Recent work also considers relations between word sense disambiguation and subjectivity [307].

Subjectivity detection or ranking at the document level can be thought of as having its roots in studies in *genre* classification (see Section 4.1.5 for more detail). For instance, Yu and Hatzivassiloglou [326] achieve high accuracy (97%) with a Naive Bayes classifier on a particular corpus consisting of Wall Street Journal articles, where the task is to distinguish articles under *News and Business* (facts) from articles under *Editorial and Letter to the Editor* (opinions). (This task was suggested earlier by Wiebe et al. [315], and a similar corpus was explored in previous work [308, 316].) Work in this direction is not limited to the binary distinction between subjective and objective labels. Recent work includes the research by participants in the 2006 TREC Blog track [227] and others [69, 97, 222, 223, 234, 279, 316, 326].

4.1.3 Joint Topic–Sentiment Analysis

One simplifying assumption sometimes made by work on document-level sentiment classification is that each document under consideration is focused on the subject matter we are interested in. This is in part because one can often assume that the document set was created

by first collecting only on-topic documents (e.g., by first running a topic-based query through a standard search engine). However, it is possible that there are interactions between topic and opinion that make it desirable to consider the two simultaneously; for example, Rilof et al. [256] find that "topic-based text filtering and subjectivity filtering are complementary" in the context of experiments in information extraction.

Also, even a relevant opinion-bearing document may contain off-topic passages that the user may not be interested in, and so one may wish to discard such passages.

Another interesting case is when a document contains material on multiple subjects that may be of interest to the user. In such a setting, it is useful to identify the topics and separate the opinions associated with each of them. Two examples of the types of documents for which this kind of analysis is appropriate are (1) comparative studies of related products, and (2) texts that discuss various features, aspects, or attributes.[3]

4.1.4 Viewpoints and Perspectives

Much work on analyzing sentiment and opinions in politically oriented text focuses on general attitudes expressed through texts that are not necessarily targeted at a particular issue or narrow subject. For instance, Grefenstette et al. [112] experimented with determining the political orientation of websites essentially by classifying the concatenation of all the documents found on that site. We group this type of work under the heading of "viewpoints and perspectives," and include under this rubric work on classifying texts as liberal, conservative, libertarian, etc. [219], placing texts along an ideological scale [178, 202], or representing Israeli versus Palestinian viewpoints [186, 187].

Although binary or n-ary classification may be used, here, the classes typically correspond not to opinions on a single, narrowly defined topic, but to a collection of bundled attitudes and beliefs. This could potentially enable different approaches from polarity

[3] When the context is clear, we often use the term "feature" to refer to "feature, aspect, or attribute" in this survey.

classification. On the other hand, if we treat the set of documents as a meta-document, and the different issues being discussed as meta-features, then this problem still shares some common ground with polarity classification or its multi-class, regression, and ranking variants. Indeed, some of the approaches explored in the literature for these two problems individually could very well be adapted to work for either one of them.

The other point of departure from the polarity classification problem is that the labels being considered are more about attitudes that do not naturally correspond with degree of positivity. While assigning-simple labels remains a classification problem, if we move farther away and aim at serving more expressive and open-ended opinions to the user, we need to solve extraction problems. For instance, one may be interested in obtaining descriptions of opinions of a greater complexity than simple labels drawn from a very small set, i.e., one might be seeking something more like "achieving world peace is difficult" than like "mildly positive." In fact, much of the prior work on perspectives and viewpoints seeks to extract more perspective-related information (e.g., opinion holders). The motivation was to enable multi-perspective question answering, where the user could ask questions such as "what is Miss America's perspective on world peace?" rather than a fact-based question (e.g., "who is the new Miss America?"). Naturally, such work is often framed in the context of extraction problems, the particular characteristics of which are covered in Section 4.9.

4.1.5 Other Non-Factual Information in Text

Researchers have considered various affect types, such as the six "universal" emotions [86]: anger, disgust, fear, happiness, sadness, and surprise [192, 9, 285]. An interesting application is in human–computer interaction: if a system determines that a user is upset or annoyed, for instance, it could switch to a different mode of interaction [188].

Other related areas of research include computational approaches for humor recognition and generation [210]. Many interesting affectual aspects of text like "happiness" or "mood" are also being explored in the context of informal text resources such as weblogs [224]. Potential

applications include monitoring levels of hateful or violent rhetoric, perhaps in multilingual settings [1].

In addition to classification based on affect and emotion, another related area of research that addresses non-topic-based categorization is that of determining the *genre* of texts [97, 98, 150, 153, 182, 277]. Since subjective genres, such as "editorial," are often one of the possible categories, such work can be viewed as closely related to subjectivity detection. Indeed, this relation has been observed in work focused on learning subjective language [316].

There has also been research that concentrates on classifying documents according to their *source* or *source style*, with statistically detected stylistic variation [38] serving as an important cue. Authorship identification is perhaps the most salient example — Mosteller and Wallace's [216] classic Bayesian study of the authorship of the Federalist Papers is one well-known instance. Argamon-Engelson et al. [18] consider the related problem of identifying not the particular author of a text, but its publisher (e.g., the *New York Times* vs. *The Daily News*); the work of Kessler et al. [153] on determining a document's "brow" (e.g., high-brow vs. "popular," or low-brow) has similar goals. Several recent workshops have been dedicated to style analysis in text [15, 16, 17]. Determining stylistic characteristics can be useful in *multi-faceted* search [10].

Another problem that has been considered in intelligence and security settings is the detection of deceptive language [46, 117, 329].

4.2 Features

Converting a piece of text into a feature vector or other representation that makes its most salient and important features available is an important part of data-driven approaches to text processing. There is an extensive body of work that addresses feature selection for machine learning approaches in general, as well as for learning approaches tailored to the specific problems of classic text categorization and information extraction [101, 263]. A comprehensive discussion of such work is beyond the scope of this survey. In this section, we focus on findings in feature engineering that are specific to sentiment analysis.

4.2.1 Term Presence vs. Frequency

It is traditional in information retrieval to represent a piece of text as a feature vector wherein the entries correspond to individual terms. One influential finding in the sentiment-analysis area is as follows. Term frequencies have traditionally been important in standard IR, as the popularity of tf-idf weighting shows; but in contrast, Pang et al. [235] obtained better performance using *presence* rather than frequency. That is, binary-valued feature vectors in which the entries merely indicate whether a term occurs (value 1) or not (value 0) formed a more effective basis for review polarity classification than did real-valued feature vectors in which entry values increase with the occurrence frequency of the corresponding term. This finding may be indicative of an interesting difference between typical topic-based text categorization and polarity classification: While a topic is more likely to be emphasized by frequent occurrences of certain keywords, overall sentiment may not usually be highlighted through repeated use of the same terms. (We discussed this point previously in Section 3.2 on factors that make opinion mining difficult.)

On a related note, *hapax legomena*, or words that appear a single time in a given corpus, have been found to be high-precision indicators of subjectivity [316]. Yang et al. [322] look at rare terms that are not listed in a pre-existing dictionary, on the premise that novel versions of words, such as "bugfested," might correlate with emphasis and hence subjectivity in blogs.

4.2.2 Term-based Features Beyond Term Unigrams

Position information finds its way into features from time to time. The position of a token within a textual unit (e.g., in the middle vs. near the end of a document) can potentially have important effects on how much that token affects the overall sentiment or subjectivity status of the enclosing textual unit. Thus, position information is sometimes encoded into the feature vectors that are employed [158, 235].

Whether higher-order n-grams are useful features appears to be a matter of some debate. For example, Pang et al. [235] report that unigrams outperform bigrams when classifying movie reviews by sentiment

polarity, but Dave et al. [69] find that in some settings, bigrams and trigrams yield better product-review polarity classification.

Riloff et al. [254] explore the use of a subsumption hierarchy to formally define different types of lexical features and the relationships between them in order to identify useful complex features for opinion analysis. Airoldi et al. [5] apply a Markov Blanket Classifier to this problem together with a meta-heuristic search strategy called Tabu search to arrive at a dependency structure encoding a parsimonious vocabulary for the positive and negative polarity classes.

The "contrastive distance" between terms — an example of a high-contrast pair of words in terms of the implicit evaluation polarity they express is "delicious" and "dirty" — was used as an automatically computed feature by Snyder and Barzilay [272] as part of a rating-inference system.

4.2.3 Parts of Speech

Part-of-speech (POS) information is commonly exploited in sentiment analysis and opinion mining. One simple reason holds for general textual analysis, not just opinion mining: part-of-speech tagging can be considered to be a crude form of word sense disambiguation [318].

Adjectives have been employed as features by a number of researchers [217, 303]. One of the earliest proposals for the data-driven prediction of the semantic orientation of words was developed for adjectives [119]. Subsequent work on subjectivity detection revealed a high correlation between the presence of adjectives and sentence subjectivity [120]. This finding has often been taken as evidence that (certain) adjectives are good indicators of sentiment, and sometimes has been used to guide feature selection for sentiment classification, in that a number of approaches focus on the presence or polarity of adjectives when trying to decide the subjectivity or polarity status of textual units, especially in the unsupervised setting. Rather than focusing on isolated adjectives, Turney [298] proposed to detect document sentiment based on selected phrases, where the phrases are chosen via a number of pre-specified part-of-speech patterns, most including an adjective or an adverb.

The fact that adjectives are good predictors of a sentence being subjective does not, however, imply that other parts of speech do not contribute to expressions of opinion or sentiment. In fact, in a study by Pang et al. [235] on movie-review polarity classification, using only adjectives as features was found to perform much worse than using the same number of most frequent unigrams. The researchers point out that nouns (e.g., "gem") and verbs (e.g., "love") can be strong indicators for sentiment. Riloff et al. [257] specifically studied the extraction of subjective nouns (e.g., "concern," "hope") via bootstrapping. There have been several targeted comparisons of the effectiveness of adjectives, verbs, and adverbs, where further subcategorization often plays a role [34, 221, 316].

4.2.4 Syntax

There have also been attempts at incorporating syntactic relations within feature sets. Such deeper linguistic analysis seems particularly relevant with short pieces of text. For instance, Kudo and Matsumoto [173] report that for two sentence-level classification tasks, sentiment polarity classification and *modality identification* ("opinion," "assertion," or "description"), a subtree-based boosting algorithm using dependency-tree-based features outperformed the bag-of-words baseline (although there were no significant differences with respect to using n-gram-based features). Nonetheless, the use of higher-order n-grams and dependency or constituent-based features has also been considered for document-level classification; Dave et al. [69] on the one hand and Gamon [103], Matsumoto et al. [204], and Ng et al. [222] on the other hand come to opposite conclusions regarding the effectiveness of dependency information. Parsing the text can also serve as a basis for modeling *valence shifters* such as negation, intensifiers, and diminishers [152]. Collocations and more complex syntactic patterns have also been found to be useful for subjectivity detection [255, 316].

4.2.5 Negation

Handling negation can be an important concern in opinion- and sentiment-related analysis. While the bag-of-words representations

of "I like this book" and "I don't like this book" are considered to be very similar by most commonly-used similarity measures, the only differing token, the negation term, forces the two sentences into opposite classes. There does not really exist a parallel situation in classic IR where a single negation term can play such an instrumental role in classification (except in cases like "this document is about cars" vs. "this document is not about cars").

It is possible to deal with negations indirectly as a second-order feature of a text segment, that is, where an initial representation, such as a feature vector, essentially ignores negation, but that representation is then converted into a different representation that is negation-aware. Alternatively, as was done in previous work, negation can be encoded directly into the definitions of the initial features. For example, Das and Chen [66] propose attaching "NOT" to words occurring close to negation terms such as "no" or "don't," so that in the sentence "I don't like deadlines," the token "like" is converted into the new token "like-NOT."

However, not all appearances of explicit negation terms reverse the polarity of the enclosing sentence. For instance, it is incorrect to attach "NOT" to "best" in "No wonder this is considered one of the best." Na et al. [220] attempt to model negation more accurately. They look for specific part-of-speech tag patterns (where these patterns differ for different negation words), and tag the complete phrase as a negation phrase. For their dataset of electronics reviews, they observe about 3% improvement in accuracy resulting from their modeling of negations. Further improvement probably needs deeper (syntactic) analysis of the sentence [152].

Another difficulty with modeling negation is that negation can often be expressed in rather subtle ways. Sarcasm and irony can be quite difficult to detect, but even in the absence of such sophisticated rhetorical devices, we still see examples such as "[it] avoids all clichés and predictability found in Hollywood movies" (internet review by "Margie24") — the word "avoid" here is an arguably unexpected "polarity reverser." Wilson et al. [319] discuss other complex negation effects.

4.2.6 Topic-Oriented Features

Interactions between topic and sentiment play an important role in opinion mining. For example, in a hypothetical article on Wal-mart, the sentences "Wal-mart reports that profits rose" and "Target reports that profits rose" could indicate completely different types of news (good vs. bad) regarding the subject of the document, Wal-mart [116]. To some extent, topic information can be incorporated into features.

Mullen and Collier [217] examine the effectiveness of various features based on topic (e.g., they take into account whether a phrase follows a reference to the topic under discussion) under the experimental condition that topic references are manually tagged. Thus, for example, in a review of a particular work of art or music, references to the item receive a "THIS_WORK" tag.

For the analysis of predictive opinions (e.g., whether a message M with respect to party P predicts P to win), Kim and Hovy [160] propose to employ feature generalization. Specifically, for each sentence in M, each party name and candidate name is replaced by PARTY (i.e., P) or OTHER (not P). Patterns such as "PARTY will win," "go PARTY again," and "OTHER will win" are then extracted as n-gram features. This scheme outperforms using simple n-gram features by about 10% in accuracy when classifying which party a given message predicts to win.

Topic–sentiment interaction has also been modeled through parse tree features, especially in opinion extraction tasks. Relationships between candidate opinion phrases and the given subject in a dependency tree can be useful in such settings [244].

Part Two: Approaches

The approaches we will now discuss all share the common theme of mapping a given piece of text, such as a document, paragraph, or sentence, to a label drawn from a pre-specified finite set or to a real number.[4] As discussed in Section 4.1, opinion-oriented classification can range from sentiment-polarity categorization in reviews to determining

[4] However, unlike classification and regression, ranking does not *require* such a mapping for each individual document.

the strength of opinions in news articles to identifying perspectives in political debates to analyzing mood in blogs. Part of what is particularly interesting about these problems is the new challenges and opportunities that they present to us. In the remainder of this section, we examine different solutions proposed in the literature to these problems, loosely organized around different aspects of machine learning approaches. Although these aspects may seem to be general themes underlying most machine learning problems, we attempt to highlight what is unique for sentiment analysis and opinion mining tasks. For instance, some unsupervised learning approaches follow a sentiment-specific paradigm for how labels for words and phrases are obtained. Also, supervised and semi-supervised learning approaches for opinion mining and sentiment analysis differ from standard approaches to classification tasks in part due to the different features involved; but we also see a great variety of attempts at modeling various kinds of *relationships* between items, classes, or sub-document units. Some of these relationships are unique to our tasks; some become more imperative to model due to the subtleties of the problems we address.

The rest of this section is organized as follows. Section 4.3 covers the impact that the increased availability of labeled data has had, including the rise of supervised learning. Section 4.4 considers issues surrounding topic and domain dependencies. Section 4.5 describes unsupervised approaches. We next consider incorporating relationships between various types of entities (Section 4.6). This is followed by a section on incorporating discourse structure (4.7). Section 4.8 is concerned with the use of language models. Finally, Section 4.9 investigates certain issues in *extraction* that are somewhat particular to it, and thus are not otherwise discussed in the sections that precede it. One such issue is the identification of features and expressions of opinions in reviews. Another set of issues arise when opinion-holder identification needs to be applied.

4.3 The Impact of Labeled Data

Work up to the early 1990s on sentiment-related tasks, such as determination of point of view and other types of complex recognition

problems, generally assumed the existence of sub-systems for sometimes rather sophisticated NLP tasks, ranging from parsing to the resolution of pragmatic ambiguities [121, 262, 310, 311, 313]. Given the state of the art of NLP at the time and, just as importantly, the lack of sufficient amounts of appropriate labeled data, the research described in these early papers necessarily considered only proposals for systems or prototype systems without large-scale empirical evaluation; typically, no learning component was involved (an interesting exception is Wiebe and Bruce [306], who proposed but did not evaluate the use of decomposable graphical models). Operational systems were focused on simpler classification tasks, relatively speaking (e.g., categorization according to affect), and relied instead on relatively shallow analysis based on manually constructed discriminant-word lexicons [133, 296], since with such a lexicon in hand, one can classify a text unit by considering which indicator terms or phrases from the lexicon appear in the given text.

The rise of the widespread availablity to researchers of organized collections of opinionated documents (two examples: financial-news discussion boards and review aggregation sites such as Epinions) and of other corpora of more general texts (e.g., newswire) and of other resources (e.g., WordNet) was a major contributor to a large shift in direction toward data-driven approaches. To begin with, the availability of the raw texts themselves made it possible to learn opinion-relevant lexicons in an unsupervised fashion, as is discussed in more detail in Section 4.5.1, rather than create them manually. But the increase in the amount of *labeled* sentiment-relevant data, in particular — where the labels are derived either through explicit researcher-initiated manual annotation efforts or by other means (see Section 7.1.1) — was a major contributing factor to activity in both supervised and unsupervised learning. In the unsupervised case, described in Section 4.5, it facilitated research by making it possible to evaluate proposed algorithms in a large-scale fashion. Unsupervised (and supervised) learning also benefitted from the improvements to sub-component systems for tagging, parsing, and so on that occurred due to the application of data-driven techniques in those areas. And, of course, the importance to supervised learning of having access to labeled data is paramount.

One very active line of work can be roughly glossed as the application of standard text-categorization algorithms, surveyed by Sebastiani [263], to opinion-oriented classification problems. For example, Pang et al. [235] compare Naive Bayes, Support Vector Machines, and maximum-entropy-based classification on the sentiment-polarity classification problem for movie reviews. More extensive comparisons of the performance of standard machine learning techniques with other types of features or feature selection schemes have been engaged in later work [5, 69, 103, 204, 217]; see Section 4.2 for more detail. We note that there has been some research that explicitly considers regression or ordinal-regression formulations of opinion-mining problems [109, 201, 233, 320]: example questions include, "how positive is this text?" and "how strongly held is this opinion?"

Another role that labeled data can play is in lexicon induction, although, as detailed in Section 4.5.1, the use of the unsupervised paradigm is more common. Morinaga et al. [215] and Bethard et al. [37] create an opinion-indicator lexicon by looking for terms that tend to be associated more highly with subjective-genre newswire, such as editorials, than with objective-genre newswire. Das and Chen [66, 67] start with a manually created lexicon specific to the finance domain (example terms: "bull," "bear"), but then assign discrimination weights to the items in the lexicon based on their cooccurrence with positively labeled vs. negatively labeled documents.

Other topics related to supervised learning are discussed in some of the more specific sections that follow.

4.4 Domain Adaptation and Topic-Sentiment Interaction

4.4.1 Domain Considerations

The accuracy of sentiment classification can be influenced by the domain of the items to which it is applied [21, 40, 88, 249, 298]. One reason is that the same phrase can indicate different sentiment in different domains: recall the Bob Bland example mentioned earlier, where "go read the book" most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews; or consider Turney's [298] observation that "unpredictable" is a positive

description for a movie plot but a negative description for a car's steering abilities. Difference in vocabularies across different domains also adds to the difficulty when applying classifiers trained on labeled data in one domain to test data in another.

Several studies show concrete performance differences from domain to domain. In an experiment auxiliary to their main work, Dave et al. [69] apply a classifier trained on a pre-assembled dataset of reviews of a certain type to product reviews of a different type. But they do not investigate the effect of training-test mis-match in detail. Engström [88] studies how the accuracy of sentiment classification can be influenced by topic. Read [249] finds standard machine learning techniques for opinion analysis to be both domain-dependent (with domains ranging from movie reviews to newswire articles) and temporally dependent (based on datasets spanning different ranges of time periods but written at least one year apart). Owsley et al. [229] also show the importance of building a domain-specific classifier.

Aue and Gamon [21] explore different approaches to customizing a sentiment classification system to a new target domain in the absence of large amounts of labeled data. The different types of data they consider range from lengthy movie reviews to short, phrase-level user feedback from web surveys. Due to significant differences in these domains along several dimensions, simply applying the classifier learned on data from one domain barely outperforms the baseline for another domain. In fact, with 100 or 200 labeled items in the target domain, an EM algorithm that utilizes in-domain unlabeled data and ignores out-of-domain data altogether outperforms the method based exclusively on (both in- and out-of-domain) labeled data.

Yang et al. [321] take the following simple approach to domain transfer: they find features that are good subjectivity indicators in both of two different domains (in their case, movie reviews versus product reviews), and consider these features to be good domain-independent features.

Blitzer et al. [40] explicitly address the domain transfer problem for sentiment polarity classification by extending the *structural correspondence learning algorithm* (*SCL*) [11], achieving an average of 46% improvement over a supervised baseline for sentiment polarity

classification of 5 different types of product reviews mined from Amazon.com. The success of SCL depends on the choice of *pivot* features in both domains, based on which the algorithm learns a projection matrix that maps features in the target domain into the feature space of the source domain. Unlike previous work that applied SCL for tagging, where frequent words in both domains happened to be good predictors for the target labels (part-of-speech tags), and were therefore good candidates for pivots, here the pivots are chosen from those with highest mutual information with the source label. The projection is able to capture correspondences (in terms of expressed sentiment polarity) between "predictable" for book reviews and "poorly designed" for kitchen appliance reviews. Furthermore, they also show that a measure of domain similarity can correlate well with the ease of adaptation from one domain to another, thereby enabling better scheduling of annotation efforts.

Cross-lingual adaptation. Much of the literature on sentiment analysis has focused on text written in English. As a result, most of the resources developed, such as lexica with sentiment labels, are in English. Adapting such resources to other languages is related to domain adaptation: the former aims at adapting from the source language to the target language in order to utilize existing resources in the source language; whereas the latter seeks to adapt from one domain to another in order to utilize the labeled data available in the source domain. Not surprisingly, we observe parallel techniques: instead of projecting unseen tokens from the new domain into the old one via co-occurrence information in the corpus [40], expressions in the new language can be aligned with expressions in the language with existing resources. For instance, one can determine cross-lingual projections through bilingual dictionaries [209], or parallel corpora [159, 209]. Alternatively, one can simply apply machine translation as a sentiment-analysis preprocessing step [32].

4.4.2 Topic (and sub-topic or feature) Considerations

Even when one is handling documents in the same domain, there is still an important and related source of variation: document topic. It is

true that sometimes the topic is pre-determined, such as in the case of free-form responses to survey questions. However, in many sentiment analysis applications, topic is another important consideration; for instance, one may be searching the blogosphere just for opinionated comments about Cornell University.

One approach to integrating sentiment and topic when one is looking for opinionated documents on a particular user-specified topic is to simply first perform one analysis pass, say for topic, and then analyze the results with respect to sentiment [134]. (See Sebastiani [263] for a survey of machine learning approaches to topic-based text categorization.) Such a two-pass approach was taken by a number of systems at the TREC Blog track in 2006, according to Ounis et al. [227], and others [234]. Alternatively, one may jointly model topic and sentiment simultaneously [84, 206], or treat one as a prior for the other [85].

But even in the case where one is working with documents known to be on-topic, not all the sentences within these documents need to be on-topic. Hurst and Nigam [134, 225] propose a two-pass process similar to that mentioned above, where each sentence in the document is first labeled as on-topic or off-topic, and sentiment analysis is conducted only for those that are found to be on-topic. Their work relies on a collocation assumption that if a sentence is found to be topical and to exhibit a sentiment polarity, then the polarity is expressed with respect to the topic in question. This assumption is also used by Nasukawa and Yi [221] and Gamon [103].

A related issue is that it is also possible for a document to contain multiple topics. For instance, a review can be a comparison of two products. Or, even when a single item is discussed in a document, one can consider features or aspects of the product to represent multiple (sub-) topics. If all but the main topic can be disregarded, then one possibility is as follows: simply consider the overall sentiment detected within the document — regardless of the fact that it may be formed from a mixture of opinions on different topics — to be associated with the primary topic, leaving the sentiment toward other topics undetermined (indeed, these other topics may never be identified). But it is more common to try to identify the topics and then determine the opinions regarding each of these topics separately. In some work, the important

topics are pre-defined, making this task easier [323]. In other work in extraction, this is not the case; the problem of the identification of product features is addressed in Section 4.9, and Section 4.6.3 discusses techniques that incorporate relationships between different features.

4.5 Unsupervised Approaches

4.5.1 Unsupervised Lexicon Induction

Quite a number of unsupervised learning approaches take the tack of first creating a *sentiment lexicon* in an unsupervised manner, and then determining the degree of positivity (or subjectivity) of a text unit via some function based on the positive and negative (or simply subjective) indicators, as determined by the lexicon, within it. Early examples of such an approach include Hatzivassiloglou and Wiebe [120], Turney [298], and Yu and Hatzivassiloglou [326]. Some interesting variants of this general technique are to use the polarity of the previous sentence as a tie-breaker when the scoring function does not indicate a definitive classification of a given sentence [130], or to incorporate information drawn from some labeled data as well [33].

A crucial component to applying this type of technique is, of course, the creation of the lexicon via the unsupervised labeling of words or phrases with their sentiment polarity (also referred to as *semantic orientation* in the literature) or subjectivity status [12, 45, 89, 90, 91, 92, 119, 130, 143, 146, 257, 286, 288, 289, 290, 299, 303, 304].

In early work, Hatzivassiloglou and McKeown [119] present an approach based on linguistic heuristics.[5] Their technique is built on the fact that in the case of polarity classification, the two classes of interest represent opposites, and we can utilize "opposition constraints" to help make labeling decisions. Specifically, constraints between pairs of adjectives are induced from a large corpus by looking at whether the two words are linked by conjunctions such as "but" (evidence for opposing orientations: "elegant but over-priced") or "and" (evidence for the same orientation: "clever and informative"). The task is then cast as a clustering or binary-partitioning problem where the inferred constraints are to be obeyed.

[5] For the purposes of the current discussion, we ignore the supervised aspects of their work.

Once the clustering has been completed, the labels of "positive orientation" and "negative orientation" need to be assigned; rather than use external information to make this decision, Hatzivassiloglou and McKeown [119] simply give the "positive orientation" label to the class whose members have the highest average frequency. But in other work, *seed words* for which the polarity is already known are assumed to be supplied, in which case labels can be determined by propagating the labels of the seed words to terms that co-occur with them in general text or in dictionary glosses, or to synonyms, words that co-occur with them in other WordNet-defined relations, or other related words (and, along the same lines, opposite labels can be given based on similar information) [12, 20, 89, 90, 130, 146, 148, 155, 288, 298, 299]. The joint use of mutual information and co-occurrence in a general corpus with a small set of seed words, a technique employed by a number of researchers, was suggested by Turney [298]; his idea was to essentially compare whether a phrase has a greater tendency to co-occur within certain context windows with the word "poor" or with the word "excellent," taking care to account for the frequencies with which "poor" and "excellent" occur, where the data on which such computations are to be made come from the results of particular types of Web search-engine queries.

Much of the work cited above focuses on identifying the *prior polarity* of terms or phrases, to use the terminology of Wilson et al. [319], or what we might by extension call terms' and phrases' *prior subjectivity status*, meaning the semantic orientation that these items might be said to generally bear when taken out of context. Such prior information is meant, of course, to serve toward further identifying *contextual* polarity or subjectivity [242, 319].

Lexicons for generation. It is worth noting that Higashinaka et al. [122] focus on a lexicon-induction task that facilitates natural language generation. They consider the problem of learning a dictionary that maps semantic representations to verbalizations, where the data comes from reviews. Although reviews are not explicitly marked up with respect to their semantics, they do contain explicit rating and aspect indicators. For example, from such data, they learn that one way to express the concept "atmosphere rating:5" is "nice and comfortable."

4.5.2 Other Unsupervised Approaches

Bootstrapping is another approach. The idea is to use the output of an available initial classifier to create labeled data, to which a supervised learning algorithm may be applied. Riloff and Wiebe [255] use this method in conjunction with an initial high-precision classifier to learn extraction patterns for subjective expressions. (An interesting, if simple, pattern discovered: the noun "fact," as in "The fact is . . . ," exhibits high correlation with subjectivity.) Kaji and Kitsuregawa [142] use a similar method to automatically construct a corpus of HTML documents with polarity labels. Similar work involving self-training is described in Wiebe and Riloff [314] and Riloff et al. [257].

Pang and Lee [234] experiment with a different type of unsupervised approach. The problem they consider is to rank search results for review-seeking queries so that documents that contain evaluative text are placed ahead of those that do not. They propose a simple "blank slate" method based on the rarity of words within the search results that are retrieved (as opposed to within a training corpus). The intuition is that words that appear frequently within the set of documents returned for a narrow topic (the *search set*) are more likely to describe objective information, since objective information should tend to be repeated within the search set; in contrast, it would seem that people's opinions and how they express them may differ. Counterintuitively, though, Pang and Lee find that when the vocabulary to be considered is restricted to the most frequent words in the search set (as a noise-reduction measure), the subjective documents tend to be those that contain a higher percentage of words that are *less* rare, perhaps due to the fact that most reviews cover the main features or aspects of the object being reviewed. (This echoes our previous observation that understanding the objective information in a document can be critical for understanding the opinions and sentiment it expresses.) The performance of this simple method is on par with that of a method based on a state-of-the-art subjectivity detection system, OpinionFinder [255, 314].

A comparison of supervised and unsupervised methods can be found in Chaovalit and Zhou [55].

4.6 Classification Based on Relationship Information

4.6.1 Relationships Between Sentences and Between Documents

One interesting characteristic of document-level sentiment analysis is the fact that a document can consist of sub-document units (paragraphs or sentences) with different, sometimes opposing labels, where the overall sentiment label for the document is a function of the set or sequence of labels at the sub-document level. As an alternative to treating a document as a bag of features, then, there have been various attempts to model the structure of a document via analysis of sub-document units, and to explicitly utilize the relationships between these units, in order to achieve a more accurate global labeling. Modeling the relationships between these sub-document units may lead to better sub-document labeling as well.

An opinionated piece of text can often consist of evaluative portions (those that contribute to the overall sentiment of the document, e.g., "this is a great movie") and non-evaluative portions (e.g., "the Powerpuff girls learned that with great power comes great responsibility"). The overlap between the vocabulary used for evaluative portions and non-evaluative portions makes it particularly important to model the context in which these text segments occur. Pang and Lee [232] propose a two-step procedure for polarity classification for movie reviews, wherein they first detect the objective portions of a document (e.g., plot descriptions) and then apply polarity classification to the remainder of the document after the removal of these presumably uninformative portions. Importantly, instead of making the subjective–objective decision for each sentence individually, they postulate that there might be a certain degree of continuity in subjectivity labels (an author usually does not switch too frequently between being subjective and being objective), and incorporate this intuition by assigning preferences for pairs of nearby sentences to receive similar labels. All the sentences in the document are then labeled as being either subjective or objective through a collective classification process, where this process employs a reformulation of the task as one of finding a *minimum s-t cut* in the appropriate graph [165]. Two key properties of this approach are (1) it

affords the finding of an *exact* solution to the underlying optimization problem via an algorithm that is efficient both in theory and in practice, and (2) it makes it easy to integrate a wide variety of knowledge sources both about individual preferences that items may have for one or the other class and about the pair-wise preferences that items may have for being placed in the same class regardless of which particular class that is. Follow-up work has used alternate techniques to determine edge weights within a minimum-cut framework for various types of sentiment-related binary classification problems at the document level [3, 27, 111, 294]. (The more general rating-inference problem can also, in special cases, be solved using a minimum-cut formulation [233].) Others have considered more sophisticated graph-based techniques [109].

4.6.2 Relationships Between Discourse Participants

An interesting setting for opinion mining is when the texts to be analyzed form part of a running discussion, such as in the case of individual turns in political debates, posts to online discussion boards, and comments on blog posts. One fascinating aspect of this kind of setting is the rich information source that references between such texts represent, since such information can be exploited for better collective labeling of the set of documents. Utilizing such relationships can be particularly helpful because many documents in the settings we have described can be quite terse (or complicated), and hence difficult to classify on their own, but we can easily categorize a difficult document if we find within it indications of agreement with a clearly, say, positive text.

Based on manual examination of 100 responses in newsgroups devoted to three distinct controversial topics (abortion, gun control and immigration), Agrawal et al. [4] observe that the relationship between two individuals in the "responded-to" network is more likely to be antagonistic — overall, 74% of the responses examined were found to be antagonistic, whereas only 7% were found to be reinforcing. By then assuming that "respond-to" links imply disagreement, they effectively classify users into opposite camps via graph partitioning, outperforming methods that depend solely on the textual information within a particular document.

Similarly, Mullen and Malouf [218] examine "quoting" behavior among users of the politics.com discussion site — a user can refer to another post by quoting part of it or by addressing the other user by name or user ID — who have been classified as either liberal or conservative. The researchers find that a significant fraction of the posts of interest to them contain quoted material, and that, in contrast to inter-blog linking patterns discussed in Adamic and Glance [2], where liberal and conservative blog sites were found to tend to link to sites of similar political orientations, and in accordance with the Agrawal et al. [4] findings cited above, politics.com posters tend to quote users at the opposite end of the political spectrum. To perform the final political-orientation classification, users are clustered so that those who tend to quote the same entities are placed in the same cluster. (Efron [83] similarly uses co-citation analysis for the same problem.)

Rather than assume that quoting always indicates agreement or disagreement regardless of the context, Thomas et al. [294] build an agreement detector for the task of analyzing transcripts of congressional floor-debates, where the classifier categorizes certain explicit references to other speakers as representing agreement (e.g., "I heartily support Mr Smith's views!") or disagreement. They then encode evidence of a high likelihood of agreement between two speakers as a relationship constraint between the utterances made by the speakers, and collectively classify the individual speeches as to whether they support or oppose the legislation under discussion, using a minimum-cut formulation of the classification problem, as described above. Follow-up work attempts to make more refined use of disagreement information [27].

4.6.3 Relationships Between Product Features

Popescu and Etzioni [244] treat the labeling of opinion words regarding product features as a collective labeling process. They propose an iterative algorithm wherein the polarity assignments for individual words are collectively adjusted through a relaxation-labeling process. Starting from "global" word labels computed over a large text collection that reflect the sentiment orientation for each particular word in general settings, Popescu and Etzioni gradually re-define the label from

one that is generic to one that is specific to a review corpus to one that is specific to a given product feature to, finally, one that is specific to the particular context in which the word occurs. They make sure to respect sentence-level local constraints that opinions connected by connectives such as "but" or "and" should receive opposite or the same polarities.

The idea of utilizing discourse information to help with the inference of relationships between product attributes can also be found in the work of Snyder and Barzilay [272], who utilize agreement information in a task where one must predict ratings for multiple aspects of the same item (e.g., food and ambiance for a restaurant). Their approach is to construct a linear classifier to predict whether all aspects of a product are given the same rating, and combine this prediction with that of individual-aspect classifiers so as to minimize a certain loss function (which they term the "grief"). Interestingly, Snyder and Barzilay [272] give an example where a collection of independent aspect-rating predictors cannot assign a correct set of aspect ratings, but augmentation with their agreement classification allows perfect rating assignment; in their specific example, the agreement classifier is able to use the presence of the phrase "but not" to predict a contrasting rating between two aspects. An important observation that Snyder and Barzilay [272] make about their formulation is that having the piece of information that all aspect ratings agree cuts down the space of possible rating tuples to a far greater degree than having the information that not all the aspect ratings are the same.

Note that the considerations discussed here relate to the topic-specific nature of opinions that we discussed in the context of domain adaptation in Section 4.4.

4.6.4 Relationships Between Classes

Regression formulations (where we include ordinal regression under this umbrella term) are quite well-suited to the rating reference problem of predicting the degree of positivity in opinionated documents such as product reviews, and to similar problems such as determining the strength with which an opinion is held. In a sense, regression implicitly models similarity relationships between classes that correspond to

points on a scale, such as the number of "stars" given by a reviewer. In contrast, standard multi-class categorization focuses on capturing the distinct features present in each class, and ignores the fact that "5 stars" is much more like "4 stars" than "2 stars." On a movie review dataset, Pang and Lee [233] observe that a one-vs-all multi-class categorization scheme can outperform regression for a three-class classification problem (positive, neutral, and negative), perhaps due to each class exhibiting a sufficiently distinct vocabulary, but for more fine-grained classification, regression emerges as the better of the two.

Furthermore, while regression-based models implicitly encode the intuition that similar items should receive similar labels, Pang and Lee [233] formulate rating inference as a *metric labeling problem*-[164], so that a natural notion of distance between classes ("2 stars" and "3 stars" are more similar to each other than "1 star" and "4 stars" are) is captured explicitly. More specifically, an optimal labeling is computed that balances the output of a classifier that considers items in isolation with the importance of assigning similar labels to similar items.

Koppel and Schler [167] consider a similar version of this problem, but where one of the classes, corresponding to "objective," does not lie on the positive-to-negative continuum. Goldberg and Zhu [109] present a graph-based algorithm that addresses the rating inference problem in the semi-supervised learning setting, where a closed-form solution to the underlying optimization problem is found through computation on a matrix induced by a graph representing inter-document similarity relationships, and the loss function encodes the desire for similar items to receive similar labels. Mao and Lebanon [201] (Mao and Lebanon [200] is a shorter version) propose to use isotonic conditional random fields to capture the ordinal labels of local (sentence-level) sentiments. Given words that are strongly associated with positive and negative sentiment, they formulate constraints on the parameters to reflect the intuition that adding a positive (negative) word should affect the local sentiment label positively (negatively).

Wilson et al. [320] treat *intensity* classification (e.g., classifying an opinion according to its strength) as an ordinal regression task.

McDonald et al. [205] leverage relationships between labels assigned at different classification stages, such as the word level or sentence level,

finding that a "fine-to-coarse" categorization procedure is an effective strategy.

4.7 Incorporating Discourse Structure

Compared to the case for traditional topic-based information access tasks, discourse structure (e.g., twists and turns in documents) tends to have more effect on overall sentiment labels. For instance, Pang et al. [235] observe that some form of discourse structure modeling can help to extract the correct label in the following example

> I hate the Spice Girls. . . . [3 things the author hates about them]. . . Why I saw this movie is a really, really, really long story, but I did, and one would think I'd despise every minute of it. But. . . Okay, I'm really ashamed of it, but I enjoyed it. I mean, I admit it's a really awful movie, . . . [they] act wacky as hell . . . the ninth floor of hell . . . a cheap [beep] movie . . . The plot is such a mess that it's terrible. But I loved it.

In spite of the predominant number of negative sentences, the overall sentiment toward the movie under discussion is positive, largely due to the order in which these sentences are presented. Needless to say, such information is lost in a bag-of-words representation.

Early work attempts to partially address this problem via incorporating location information in the feature set [235]. Specifically, the position at which a token appears can be appended to the token itself to form position-tagged features, so that the same unigram appearing in, say, the first quarter and the last quarter of the document are treated as two different features; but the performance of this simple scheme does not differ greatly from that which results from using unigrams alone.

On a related note, it has been observed that position matters in the context of summarizing sentiment in a document. In particular, in contrast to topic-based text summarization, where the beginnings of articles usually serve as strong baselines in terms of summarizing the objective information in them, the last n sentences of a review have

been shown to serve as a much better summary of the overall sentiment of the document than the first n sentences, and to be almost as good as using the n most (automatically-computed) subjective sentences, in terms of how accurately they represent the overall sentiment of the document [232].

Theories of lexical cohesion motivate the representation used by Devitt and Ahmad [73] for sentiment polarity classification of financial news.

Another way of capturing discourse structure information in documents is to model the global sentiment of a document as a trajectory of local sentiments. For example, Mao and Lebanon [200] propose using sentiment flow as a sequential model to represent an opinionated document. More specifically, each sentence in the document receives a *local sentiment* score from an isotonic-conditional-random-field-based sentence level predictor. The sentiment flow is defined as a function $h : [0, 1) \mapsto O$ (the ordinal set), where the interval $[(t-1)/n, t/n)$ is mapped to the label of the tth sentence in a document with n sentences. The flow is then smoothed out through convolution with a smoothing kernel. Finally, the distances between two flows (e.g., L_p distance between the two smoothed, continuous functions) should reflect, to some degree, the distances between global sentiments. On a small dataset, Mao and Lebanon observe that the sentiment flow representation (especially when objective sentences are excluded) outperforms a plain bag-of-words representation in predicting global sentiment with a nearest neighbor classifier.

4.8 Language Models

The rise of the use of language models in information retrieval has been an interesting recent development [65, 177, 179, 243]. They have been applied to various opinion-mining and sentiment-analysis tasks, and in fact the subjectivity-extraction work of Pang and Lee [232] is a demo application for the heavily language-modeling-oriented LingPipe system.[6]

[6] http://alias-i.com/lingpipe/demos/tutorial/sentiment/read-me.html.

One characteristic of language modeling approaches that differentiates them somewhat from other classification-oriented data-driven techniques we have discussed so far is that language models are often constructed using labeled data, but, given that they are mechanisms for assigning probabilities to text rather than labels drawn from a finite set, they cannot, strictly speaking, be defined as either supervised or unsupervised classifiers. On the other hand, there are various ways to convert their output to labels when necessary.

An example of work in the language-modeling vein is that of Eguchi and Lavrenko [84], who rank sentences by both sentiment relevancy and topic relevancy, based on previous work on *relevance language models* [179]. They propose a generative model that jointly models sentiment words, topic words, and sentiment polarity in a sentence as a triple. Lin and Hauptmann [186] consider the problem of examining whether two collections of texts represent different perspectives. In their study, employing Reuters data, two examples of different perspectives are the Palestinian viewpoint vs. the Israeli viewpoint in written text and Bush vs. Kerry in presidential debates. They base their notion of difference in perspective upon the Kullback–Leibler (KL) divergence between posterior distributions induced from document collection pairs, and discover that the KL divergence between different aspects is an order of magnitude smaller than that between different topics. This perhaps provides yet another reason that opinion-oriented classification has been found to be more difficult than topic-based classification.

Research employing *probabilistic latent semantic analysis* (*PLSA*) [125] or *latent Dirichlet allocation* (*LDA*) [39] can also be cast as language-modeling work [41, 195, 206]. The basic idea is to infer language models that correspond to unobserved "factors" in the data, with the hope that the factors that are learned represent topics or sentiment categories.

4.9 Special Considerations for Extraction

Opinion-oriented extraction. Many applications, such as summarization or question answering, require working with pieces of information that need to be pulled from one or more textual units. For example,

a *multi-perspective question–answering* (*MPQA*) system might need to respond to opinion-oriented questions such as "Was the most recent presidential election in Zimbabwe regarded as a fair election?" [51]; the answer may be encoded in a particular sentence of a particular document, or may need to be stitched together from pieces of evidence found in multiple documents. *Information extraction* (*IE*) is precisely the field of natural language processing devoted to this type of task [49]. Hence, it is not surprising that the application of information-extraction techniques to opinion mining and sentiment analysis has been proposed [51, 79]. In this survey, we use the term *opinion-oriented information extraction* (*opinion-oriented IE*) to refer to information extraction problems particular to sentiment analysis and opinion mining. (We sometimes shorten the phrase to *opinion extraction*, which should not be construed narrowly as focusing on the extraction of opinion expressions; for instance, determining product features is included under the umbrella of this term.)

Past research in this area has been dominated by work on two types of texts:

- Opinion-oriented information extraction from *reviews* has, as noted above, attracted a great deal of interest in recent years. In fact, the term "opinion mining," when construed in its narrow sense, has often been used to describe work in this context. Reviews, while typically (but not always) devoted to a single item, such as a product, service, or event, generally comment on multiple aspects, facets, or features of that item, and all such commentary may be important. Extracting and analyzing opinions associated with each individual aspect can help provide more informative summarizations or enable more fine-grained opinion-oriented retrieval.
- Other work has focused on *newswire*. Unlike reviews, a news article is relatively likely to contain descriptions of opinions that do not belong to the article's author; an example is a quotation from a political figure. This property of journalistic text makes the identification of opinion holders (also known as opinion sources) and the correct association of opinion

> holders with opinions important tasks, whereas for reviews, all expressed opinions are typically those of the author, so opinion-holder identification is a less salient problem. Thus, when newswire articles are the focus, the emphasis has tended to be on identifying expressions of opinions, the agent expressing each opinion, and/or the type and strength of each opinion. Early work in this direction first carefully developed and evaluated a low-level opinion annotation scheme [45, 283, 309, 312], which facilitated the study of sub-tasks such as identifying opinion holders and analyzing opinions at the phrase level [37, 42, 43, 51, 60, 61, 157, 320].

It is important to understand the similarities and differences between opinion-oriented IE and standard fact-oriented IE. They share some sub-tasks in common, such as entity recognition; for example, as mentioned above, determination of opinion holders is an active line of research [37, 42, 61, 158]. What truly sets the problem apart from standard or classic IE is the specific types of entities and relations that are considered important. For instance, although identification of product features is in some sense a standard entity recognition problem, an opinion extraction system would be mostly interested in features for which associated opinions exist; similarly, an opinion holder is not just any named entity in a news article, but one that expresses opinions. Examples of the types of relations particularly pertinent to opinion mining are those centered around comparisons — consider, for example, the relations encoded by such sentences as "The new model is more expensive than the old one" or "I prefer product A over product B" ([139, 191], longer version of the latter available as Jindal and Liu [138]) — or between agents and reported beliefs, as described in Section 4.9.2. Note that the relations of interest can form a complex hierarchical structure, as in the case where an opinion is attributed to one party by another, so that it is unclear whether the first party truly holds the opinion in question [42].

It is also important to understand which aspects of opinion-oriented extraction are mentioned in this section as opposed to the previous sections. As discussed earlier, many sub-problems of opinion extraction are

in fact classification problems for relatively small textual units. Examples include both determining whether or not a text span is subjective and classifying a given text span already determined to be subjective by the strength of the opinion expressed. Thus, many key techniques involved in building an opinion extraction system are already discussed in the previous sections. In this section, we instead focus on the "missing pieces," describing approaches to problems that are somewhat special to extraction tasks in sentiment analysis. While these sub-tasks can be (and often are) cast as classification problems, they do not have natural counterparts outside of the extraction context. Specifically, Section 4.9.1 is devoted to the identification of features and expressions of opinions in reviews. Section 4.9.2 considers techniques that have been employed when opinion-holder identification is an issue.

Finally, we make the following organizational note. One may often want to present the output of opinion extraction in summarized form; conversely, some forms of sentiment summarization rely on the output of opinion extraction. Opinion-oriented summarization is discussed in Section 5.

4.9.1 Identifying Product Features and Opinions in Reviews

In the context of review mining [130, 166, 215, 244, 323, 324], two important extraction-related sub-tasks are

(1) The identification of product features, and
(2) the extraction of opinions associated with these features.

While the key features or aspects are known in some cases, many systems start from problem (1).

As noted above, identification of product features is in some sense a standard information extraction task with little to distinguish it from other non-sentiment-related problems. After all, the notion of the features that a given product has seems fairly objective. However, Hu and Liu [130] show that one can benefit from light sentiment analysis even for this sub-task, as described shortly.

Existing work on identifying product features discussed in reviews (task (1)) often relies on the simple linguistic heuristic that (explicit) features are usually expressed as nouns or noun phrases. This narrows down the candidate words or phrases to be considered, but obviously not all nouns or noun phrases are product features. Yi et al. [323] consider three increasingly strict heuristics to select from noun phrases based on part-of-speech-tag patterns. Hu and Liu [130] follow the intuition that frequent nouns or noun phrases are likely to be features. They identify frequent features through association mining, and then apply heuristic-guided pruning aimed at removing (a) multi-word candidates in which the words do not appear together in a certain order, and (b) single-word candidates for which subsuming super-strings have been collected (the idea is to concentrate on more specific concepts, so that, for example, "life" is discarded in favor of "battery life"). These techniques by themselves outperform a general-purpose term-extraction and -indexing system known as FASTR [135]. Furthermore — and here is the observation that is relevant to sentiment — the F-measure can be further improved (although precision drops slightly) via the following expansion procedure: adjectives appearing in the same sentence as frequent features are assumed to be opinion words, and nouns and noun phrases co-occurring with these opinion words in other sentences are taken to be infrequent features.

In contrast, Popescu and Etzioni [244] consider product features to be concepts forming certain relationships with the product (for example, for a scanner, its size is one of its properties, whereas its cover is one of its parts) and seek to identify the features connected with the product name through corresponding meronymy discriminators. Note that this approach, which does not involve sentiment analysis per se but simply focuses more on the task of identifying different types of features, achieved better performance than that yielded by the techniques of Hu and Liu [130].

There has also been work that focuses on extracting attribute-value pairs from textual product descriptions, but not necessarily in the context of opinion mining. Of work in this vein, Ghani et al. [105] directly compare against the method proposed by Hu and Liu [130].

To identify expressions of opinions associated with features (task (2)), a simple heuristic is to simply extract adjectives that appear in the same sentence as the features [130]. Deeper analyses can make use of parse information and manually or semi-automatically developed rules or sentiment-relevant lexicons [215, 244].

4.9.2 Problems Involving Opinion Holders

In the context of analysis of newswire and related genres, we need to identify text spans corresponding both to opinion holders and to expressions of the opinions held by them.

As is true with other segmentation tasks, identifying opinion holders can be viewed as a sequence labeling problem. Choi et al. [61] experiment with an approach that combines *Conditional Random Fields* (*CRFs*) [176] and extraction patterns. A CRF model is trained on a certain collection of lexical, syntactic, and semantic features. In particular, extraction patterns are learned to provide semantic tagging as part of the semantic features. (CRFs have also been used to detect opinion expressions [43].)

Alternatively, given that the status of an opinion holder depends by definition on the expression of an opinion, the identification of opinion holders can benefit from, or perhaps even require, accounting for opinion expressions either simultaneously or as a pre-processing step.

One example of simultaneous processing is the work of Bethard et al. [37], who specifically address the task of identifying both opinions and opinion sources. Their approach is based on semantic parsing where semantic constituents of sentences (e.g., "agent" or "proposition") are marked. By utilizing opinion words automatically learned by a bootstrapping approach, they further refine the semantic roles to identify *propositional opinions*, i.e., opinions that generally function as the sentential complement of a predicate. This enables them to concentrate on verbs and extract verb-specific information from semantic frames such as are defined in FrameNet [25] and PropBank [230].

As another example of the simultaneous approach, Choi et al. [60] employ an *integer linear programming* approach to handle the joint

extraction of entities and relations, drawing on the work of Roth and Yih [260] on using global inference based on constraints.

As an alternative to the simultaneous approach, a system can start by identifying opinion expressions, and then proceed to the analysis of the opinions, including the identification of opinion holders. Indeed, Kim and Hovy [159] define the problem of opinion holder identification as identifying opinion sources given an opinion expression in a sentence. In particular, structural features from a syntactic parse tree are selected to model the long-distance, structural relation between a holder and an expression. Kim and Hovy show that incorporating the patterns of paths between holder and expression outperforms a simple combination of local features (e.g., the type of the holder node) and other non-structural features (e.g., the distance between the candidate holder node and the expression node).

One final remark is that the task of determining which mentions of opinion holders are co-referent (*source coreference resolution*) differs in practice in interesting ways from typical noun phrase coreference resolution, due in part to the way in which opinion-oriented datasets may be annotated [282].

5

Summarization

So far, we have talked about *analyzing and extracting* opinion information from individual documents. The focus of this section is on *aggregating and representing* sentiment information drawn from an individual document or from a collection of documents. For example, a user might desire an at-a-glance presentation of the main points made in a single review; creating such single-document sentiment summaries is described in Section 5.1. Another application considered within this paradigm is the automatic determination of *market sentiment*, or the majority "leaning" of an entire body of investors, from the individual remarks of those investors [66, 67]; this is a type of multi-document opinion-oriented summarization, described in Section 5.2.

5.1 Single-Document Opinion-Oriented Summarization

There is clearly a tight connection between extraction of topic-based information from a single document [49] and topic-based summarization of that document, since the information that is pulled out can serve as a summary; see Radev et al. [247, Section 2.1] for a brief review.

Obviously, this connection between extraction and summarization holds in the case of sentiment-based summarization, as well.

One way in which this connection is made manifest in single-document opinion-oriented summarization is as follows: there are approaches that create textual sentiment summaries based on extraction of sentences or similar text units. For example, Beineke et al. [33] attempt to select a single passage[1] that reflects the opinion of the document's author(s), mirroring the practice of film advertisements that present "snippets" from reviews of the movie. Training and test data is acquired from the website Rotten Tomatoes (http://www.rottentomatoes.com), which provides a roughly sentence-length snippet for each review. However, Beineke et al. [33] note that low accuracy can result even for high-quality extraction methods because the Rotten Tomatoes data includes only a single snippet per review, whereas several sentences might be perfectly viable alternatives. In terms of creating longer summaries, Mao and Lebanon [200] suggest that by tracking the sentiment flow within a document — i.e., how sentiment orientation changes from one sentence to the next, as discussed in Section 4.7 — one can create sentiment summaries by choosing the sentences at local extrema of the flow (plus the first and last sentence). An interesting feature of this approach is that by incorporating a document's flow, the technique takes into account the entire document in a holistic way. Both approaches just mentioned seek to select the absolutely most important sentences to present. Alternatively, one could simply extract *all* subjective sentences, as was done by Pang and Lee [232] to create "subjectivity extracts." They suggested that these extracts could be used as summaries, although, as mentioned above, they focused on the use of these extracts as an aid to downstream polarity classification, rather than as summaries per se. Finally, we note that sentences are also used in multi-document sentiment summarization as well, as described in Section 5.2.

Other sentiment summarization methods can work directly off the output of opinion-oriented information-extraction systems. Indeed,

[1] Beineke et al. [33] use the term "sentiment summary" to refer to a single passage, but we prefer to not restrict that term's definition so tightly.

Cardie et al. [51], speaking about the more restricted type of extraction referred to by the technical term "information extraction," "propose to view ... summary representations as information extraction (IE) scenario templates ... [thus] we postulate that methods from information extraction ... will be adequate for the automatic creation of opinion-based summary representations." (A similar observation was made by Dini and Mazzini [79].) Note that these IE templates do not form coherent text on their own. However, they can be incorporated as is into *visual* summaries.

Indeed, one interesting aspect of the problem of extracting sentiment information from a single document (or from multiple documents, as discussed in Section 5.2) is that sometimes graph-based output seems much more appropriate or useful than text-based output. For example, graph-based summaries are very suitable when the information that is most important is the set of entities described and the opinions that some of these entities hold about each other [305]. Figure 5.1 shows an example of a human-generated summary in the form of a graph depicting various negative opinions expressed during the aftermath of Hurricane Katrina. Note the inclusion of text snippets on the arrows to support the inference of a negative opinion[2]; in general, providing some sense of the evidence from which opinions are inferred is likely to be helpful to the user.

While summarization technologies may not be able to achieve the level of sophistication of information presentation exhibited by Figure 5.1, current research is making progress toward that goal. In Figure 5.2, we see a proposed summary where opinion holders and the objects of their opinions are connected by edges, and various annotations derived from IE output are included, such as the strength of various attitudes.

Of course, graphical elements can also be used to represent a single bit, number or grade as a very succinct summary of a document's

[2] The exceptions are the edges from "news media"and the edges from "people who didn't evacuate." It is (perhaps intentionally) ambiguous whether the lack of supporting quotes is due merely to the lack of sufficiently "juicy" ones or is meant to indicate that it is utterly obvious that these entities blame many others. We also note that the hurricane itself is not represented.

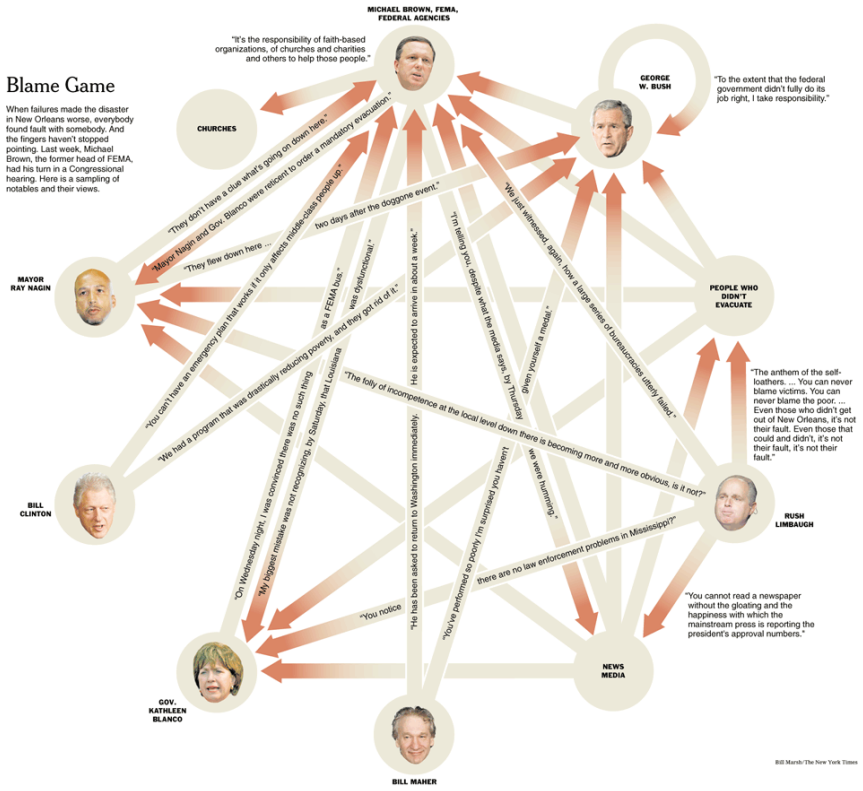

Fig. 5.1 Graphic by Bill Marsh for *The New York Times*, October 1, 2005, depicting negative opinions of various entities toward each other in the aftermath of Hurricane Katrina. Relation to opinion summarization pointed out by Eric Breck (Claire Cardie, personal communication).

sentiment. Variations of stars, letter grades, and thumbs up/thumbs down icons are common. More complex visualization schemes applied on a sentence-by-sentence basis have also been proposed [7].

5.2 Multi-Document Opinion-Oriented Summarization

> Language is itself the collective art of expression, a summary of thousands upon thousands of individual intuitions. The individual gets lost in the collective creation, but his personal expression has left some trace in a certain give and flexibility that are inherent in all

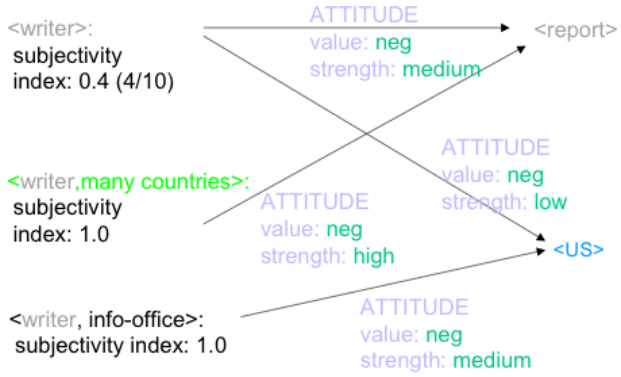

Fig. 5.2 Figure 2 (labeled 3) of Cardie et al. [51]: proposal for a summary representation derived from the output of an information-extraction system.

> collective works of the human spirit. — Edward Sapir, *Language and Literature*, 1921. Connection to sentiment analysis pointed out by Das and Chen [67].

5.2.1 Some Problem Considerations

> There never was in the world two opinions alike, no more than two hairs, or two grains; the most universal quality is diversity.
>
> — Michel de Montaigne, *Essays*

> Where an opinion is general, it is usually correct.
>
> — Jane Austen, *Mansfield Park*

We briefly discuss here some points to keep in mind in regards to multi-document sentiment summarization, although to a certain degree, work in sentiment summarization has not yet reached a level where these problems have come to the fore.

Determining which documents or portions of documents express the same opinion is not always an easy task; but, clearly it is one that needs to be addressed in the summarization setting, since readers of sentiment summaries surely are interested in the overall sentiment in the corpus — which means the system must determine shared sentiments within the document collection at hand.

This issue can still arise even when labels have been predetermined, if the items that have been pre-labeled come from different sub-collections. For instance, some documents may have polarity labels, whereas others may contain ratings on a 1-to-5 scale. And even when the ratings are drawn from the same set, calibration issues may arise. Consider the following from Rotten Tomatoes' frequently-asked-questions page (http://www.rottentomatoes.com/pages/faq#judge):

> *On the Blade 2 reviews page, you have a negative review from James Berardinelli (2.5/4 stars), and a positive review from Eric Lurio (2.5/5). Why is Berardinelli's review labeled Rotten and Lurio's review labeled Fresh?*
>
> You're seeing this discrepancy because star systems are not consistent between critics. For critics like Roger Ebert and James Berardinelli, 2.5 stars or lower out of 4 stars is always negative. For other critics, 2.5 stars can either be positive or negative. Even though Eric Lurio uses a 5 star system, his grading is very relaxed. So, 2 stars can be positive. Also, there's always the possibility of the webmaster or critic putting the wrong rating on a review.

As another example, in reconciling reviews of conference submissions, program-committee members must often take into account the fact that certain reviewers always tend to assign low scores to papers, while others have the opposite tendency. Indeed, we believe this calibration issue may be the reason why reviews of cars on Epinions come not only with a "number of stars" annotation, but also a "thumbs up/thumbs down" indicator, in order to clarify whether, regardless of the rating assigned, the review author actually intends to make a positive recommendation or not.

An additional observation to take note of is the fact that when two reviewers agree on a rating, they may have different reasons for doing so, and it may be important to indicate these reasons in the summary. A related point is that when a reviewer assigns a middling rating, it may be because he or she thinks that most aspects of the item under discussion are so-so, but it may also be because he or she sees

both strong positives and strong negatives. Or, reviewers may have the same opinions about individual item features, but weight these individual factors differently, leading to a different overall sentiment. Indeed, Rotten Tomatoes summarizes a set of reviews both with the Tomatometer — percentage of reviews judged to be positive — and an average rating on a 1-to-10 scale. The idea, again according to the FAQ (http://www.rottentomatoes.com/pages/faq#avgvstmeter), is as follows:

> The Average Rating measures the overall quality of a product based on an average of individual critic scores. The Tomatometer simply measures the percentage of critics who recommend a certain product.
>
> For example, while "Men in Black" scored 90% on the Tomatometer, the average rating is only 7.5/10. That means that while you're likely to enjoy MIB, it probably wasn't a contender for Best Picture at the Oscars.
>
> In contrast, "Toy Story 2" received a perfect 100% on the Tomatometer with an average rating of 9.6/10. That means, not only are you certain to enjoy it, you'll also be impressed with the direction, story, cinematography, and all the other things that make truly great films great.

The problem of deciding whether two sentences or text passages have the same semantic content is one that is faced not just by opinion-oriented multi-document summarizers, but by topic-based multi-document summarizers as well [247]; this has been one of the motivations behind work on paraphrase recognition [29, 30, 231] and textual entailment [28]. But, as pointed out in Ku et al. [170], while in traditional summarization redundant information is often discarded, in opinion summarization one wants to track and report the degree of "redundancy," since in the opinion-oriented setting the user is typically interested in the (relative) number of times a given sentiment is expressed in the corpus.

Carenini et al. [52] note that a challenge in sentiment summarization is that the pieces of information to be summarized — people's

opinions — are often conflicting, which is a bit different from the usual situation in topic-based summarization, where typically one does not assume that there are conflicting sets of facts in the document set (although there are exceptions [301, 302]).

5.2.2 Textual Summaries

In standard topic-based multi-document summarization, creating textual summaries has been a main focus of effort. Hence, despite the differences in topic- and opinion-based summarization mentioned above, several researchers have developed systems that create textual summaries of opinion-oriented information.

5.2.2.1 Leveraging Existing Topic-Based Technologies

One line of attack is to adapt existing topic-based multi-document summarization algorithms to the sentiment setting.

Sometimes the adaptation consists simply of modifying the input to these pre-existing algorithms. For instance, Seki et al. [264] propose that one apply standard multi-document summarization to a sub-collection of documents that are on the same topic and that are determined to belong to some relevant genre of text, such as "argumentative."

In other cases, pre-existing topic-based summarization techniques are modified. For example, Carenini et al. [52] generate natural-language summaries in the form of an "evaluative argument" using the classic natural-language generation pipeline of content selection, lexical selection and sentence planning, and sentence realization [251], assuming the existence of a pre-defined product-feature hierarchy. The system explicitly produces textual descriptions of aggregate information. The system is capable of relaying data about the average sentiment and signaling, if appropriate, that the distribution of responses is bi-modal (this allows one to report "split votes"). They compare this system against a modification of an existing sentence-extraction system, MEAD [246]. The former approach seems more well-suited for general overviews, whereas the latter seems better at providing more variety in expression and more detail; see Figure 5.3. Related to the

Summary created via a "true natural-language-generation" approach:

> Almost all users loved the Canon G3 possibly because some users thought the physical appearance was very good. Furthermore, several users found the manual features and the special features to be very good. Also, some users liked the convenience because some users thought the battery was excellent. Finally, some users found the editing/viewing interface to be good despite the fact that several customers really disliked the viewfinder. However, there were some negative evaluations. Some customers thought the lens was poor even though some customers found the optical zoom capability to be excellent. Most customers thought the quality of the images was very good.

Summary created by a modified sentence-extraction system:

> Bottom line, well made camera, easy to use, very flexible and powerful features to include the ability to use external flash and lense/filters choices. It has a beautiful design, lots of features, very easy to use, very configurable and customizable, and the battery duration is amazing! Great colors, pictures, and white balance. The camera is a dream to operate in automode, but also gives tremendous flexibility in aperture priority, shutter priority, and manual modes. I'd highly recommend this camera for anyone who is looking for excellent quality pictures and a combination of ease of use and the flexibility to get advanced with many options to adjust if you like.

Fig. 5.3 Sample automatically generated summaries. Adapted from Figure 2 of Carenini et al. [52].

latter approach, sentence extraction methods have also been used to create summaries for opinion-oriented queries or topics [265, 266].

While we are not aware of the following technique being used in standard topic-based summarization, we see no reason why it is not applicable to that setting, at least in principle. Ku et al. [170] (short version available as Ku et al. [169]) propose the following simple scheme to create a textual summary of a set of documents known in advance to be on the same topic. Sentences considered to be representative of the topic are collected, and the polarity of each such sentence is computed based on what sentiment-bearing words it contains, with negation taken into account. Then, to create a summary of the positive documents, the system simply returns the headline of the document with the most positive on-topic sentences, and similarly for the negative

documents. The authors show the following examples for the positive and the negative summary, respectively:

- Positive: "Chinese Scientists Suggest Proper Legislation for Clone Technology."
- Negative: "UK Government Stops Funding for Sheep Cloning Team."

The cleverness of this method is that headlines are, by construction, good summaries (at least of the article they are drawn from), so that fluency and informativeness, although perhaps not appropriateness, are guaranteed.

Another perhaps unconventional type of multi-document "summary" is the selection of a few documents of interest from the corpus for presentation to the user. In this vein, Kawai et al. [151] have developed a news portal site called "Fair News Reader" that attempts to determine the affect characteristics of articles the user has been reading so far (e.g., "happiness" or "fear") and then recommends articles that are on the same topic but have opposite affect characteristics. One could imagine extending this concept to a news portal that presented to the user opinions opposing his or her pre-conceived ones (Phoebe Sengers, personal communication). On a related note, Liu [190] mentions that one might desire a summarization system to present a "representative sample" of opinions, so that both positive and negative points of view are covered, rather than just the dominant sentiment. As of the time of this writing, Amazon presents the most helpful favorable review side-by-side with the most helpful critical review if one clicks on the "[x] customer reviews" link next to the stars indicator. Additionally, one could interpret the opinion-leader identification work of Song et al. [275] as suggesting that blog posts written by opinion leaders could serve as an alternative type of representative sample.

Summarizing online discussions and blogs is an area of related work [131, 300, 330]. The focus of such work is not on summarizing the opinions per se, although Zhou and Hovy [330] note that one may want to vary the emphasis on the opinions expressed versus the facts expressed.

5.2.2.2 Textual Summarization Without Topic-based Summarization Techniques

Other work in the area of textual multi-document sentiment summarization departs from topic-based work. The main reason seems to be that redundancy elimination is much less of a concern: users may wish to look at many individual opinions regardless of whether these individual opinions express the same overall sentiment, and these users may not particularly care whether the textual overview they peruse is coherent. Thus, in several cases, textual "summaries" are generated simply by listing some or all opinionated sentences. These are often grouped by feature (sub-topic) and/or polarity, perhaps with some ranking heuristic such as feature importance applied [129, 170, 324, 332].

5.2.3 Non-textual Summaries

In the previous section, we have discussed the creation of *textual* summaries of the opinion information expressed within a corpus. But in settings where the polarity or orientation of each individual document within a collection is summed up in a single bit (e.g., thumbs up/thumbs down), number (e.g., 3.5 stars), or grade (e.g., B+), an alternative way to obtain a succinct summary of the overall sentiment is to report *summary statistics*, such as the number of reviews that are "thumbs up" or the average number of stars or average grade. Many systems take this approach to summarization.

Summary statistics are often quite suited to graphical representations; we describe some noteworthy visual aspects of these summaries here (evaluation of the user-interface aspects has not been a focus of attention in the community to date).

5.2.3.1 "Bounded" Summary Statistics: Averages and Relative Frequencies

We use the term *bounded* to refer to summary statistics that lie within a predetermined range. Examples are the average number of stars (range: 0 to 5 stars, say) or the percentage of positive opinions (range: 0% to 100%).

"Thermometer"-type images are one means for displaying such statistics. One example is the "Tomatometer" on the Rotten Tomatoes website, which is simply a bar broken into two differently colored portions; the portion of the bar that is colored red indicates the fraction of positive reviews of a given movie. This representation extends straightforwardly to n-ary categorization schemes, such as positive/middling/negative, via the use of n colors. The thermometer-graphic concept also generalizes in other ways; for instance, the depiction of a number of stars can be considered to be a variant of this idea.

Instead of using size or extent to depict bounded summary statistics, as is done with thermometer representations, one can use color shading. This choice seems particularly appropriate in settings where the amount of display real-estate that can be devoted to any particular item under evaluation is highly limited or where size or location is reserved to represent some other information. For instance, Gamon et al. [104] use color to represent the general assessment of (automatically determined) product features. In Figure 5.4, we see that each of many features or topics, such as "handling" or "vw, service," is represented by a shaded box. The colors for any given box range from red to white to green, indicating gradations of the average sentiment toward that topic, moving from negative to neutral (or objective) to positive, respectively. Note that one can quickly glean from such a display what was liked and what was disliked about the product under discussion, despite the large number of topics under evaluation — people like driving this car but dislike the service. As shown in Figure 5.5, a similar interface (together with a usability study) is presented in Carenini et al. [53]. Some differences are that natural-language summarization is also employed, so that the summary is both "verbal" and visual; the features are grouped into a hierarchy, thus leveraging the ability of Treemaps [270] to display hierarchical data via nesting; and the interface also includes a way (not depicted in the figure) to see an "at-a-glance" summary of the polarities of the individual sentences commenting on a particular feature. A demo is available online at http://www.cs.ubc.ca/ carenini/storage/SEA/demo.html.

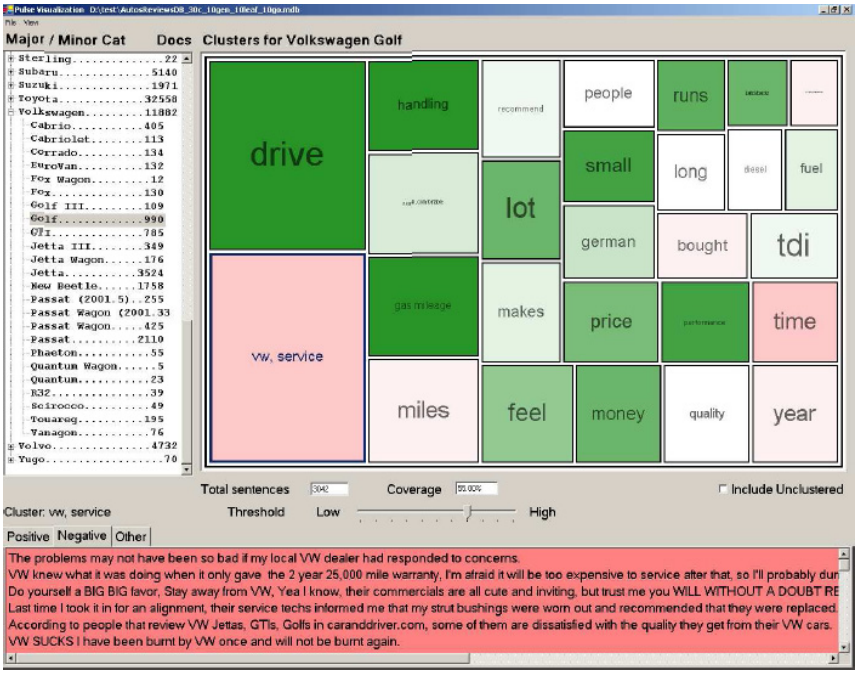

Fig. 5.4 Figure 2 of Gamon et al. [104], depicting (automatically determined) topics discussed in reviews of the Volkswagen Golf. The size of each topic box indicates the number of mentions of that topic. The shading of each topic box, ranging from red to white to green, indicates the average sentiment, ranging from negative to neutral/none to negative, respectively. At the bottom, the sentences most indicative of negative sentiment for the topic "vw, service" are displayed.

5.2.3.2 Unbounded Summary Statistics

As just described, thermometer graphics and color shading can be used to represent bounded statistics such as the mean or, in the case of n-color thermometers, relative distributions of ratings across different classes. But bounded statistics by themselves do not provide other important pieces of information, such as the actual number of opinions within each class. (We consider raw frequencies to be conceptually unbounded, although there are practical limits to how many opinions can be accounted for.) Intuitively, the observation that 50% of the reviews of a particular product are negative[3] is more of a

[3] We admit to being "glass-half-empty" people.

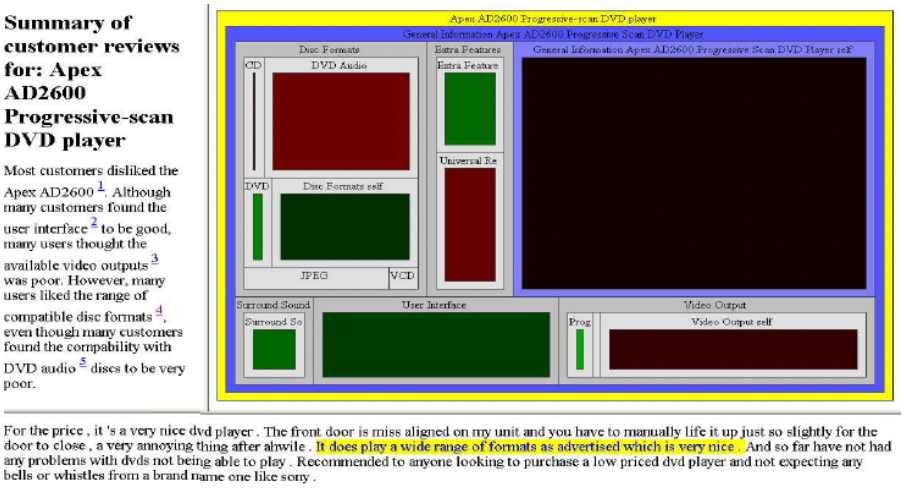

Fig. 5.5 Figure 4 of Carenini et al. [53], showing a summary of reviews of a particular product. An automatically generated text summary is on the left; a visual summary is on the right. The size of each each box in the visual summary indicates the number of mentions of the corresponding topic, which occupy a pre-defined hierarchy. The shading of each topic box, ranging from red to black to green, indicates the average sentiment, ranging from negative to neutral/none to negative, respectively. At the bottom is shown the source for the portion of the generated natural-language summary that has been labeled with the footnote "4."

big deal if that statistic is based on 10,000 reviews than if it based on only two.

Another problem specific to the mean as a summary statistic is that review-aggregation sites seem to often exhibit highly skewed rating distributions, with a particular bias toward highly positive reviews [74, 59, 128, 253, 132, 240].[4] Since there can often be a second mode, or bump, at the extreme low end of the rating scale, indicating polarization — for example, Hu et al. [132] remark that 54% of the items in a sample of Amazon book, DVD, and video products with more than 20 reviews fail both statistical normality and unimodality tests — reporting only the mean rating score may not provide enough information. To put it another way, divulging the average does not give the user

[4] On a related note, William Safire's *New York Times* May 1, 2005 article "Blurbosphere" quotes Charles McGrath, former editor of the *New York Times Book Review*, as asking, "has there ever been a book that wasn't acclaimed?"

enough information to distinguish between a set of middling reviews and a set of polarized reviews.

On the other hand, it is worth pointing out that just giving the number of positive and negative reviews, respectively, on the assumption that the user can always derive the percentages from these counts, may not suffice. Cabral and Hortaçsu [47] observe that once eBay switched to displaying the *percentage* of pieces of feedback on sellers that were negative, as opposed to simply the raw numbers, then negative reviews began to have a measurable economic impact (see Section 6).

Hence, not surprisingly, sentiment summaries tend to include data on the average rating, the distribution of ratings, and/or the number of ratings.

Visualization of unbounded summary statistics. Of the two systems described above that represent the average polarity of opinions via color, both represent the quantity of the opinions on a given topic via size. This means that the count data for positive and for negative opinions are not explicitly presented separately. In other systems, this is not the case; rather, frequencies for different classes are broken out and displayed.

For instance, as of the time of this writing, Amazon displays an average rating as a number of stars with the number of reviews next to it; mousing over the stars brings up a histogram of reviewer ratings annotated with counts for the 5-star reviews, 4-star reviews, etc. (Further mousing over the bars of the histogram brings up the percentage of reviews that each of those counts represent.)

As another example, a sample output of the Opinion Observer system [191] is depicted in Figure 5.6, where the portion of a bar projecting above the centered "horizon" line represents the number of positive opinions about a certain product feature, and the portion of the bar below the line represents the number of negative opinions. (The same idea can be used to represent percentages too, of course.) A nice feature of this visualization is that because of the use of a horizon line, two separate frequency datapoints — the positive and negative counts — can be represented by what is visually one object, namely, a solid bar, and one can easily simultaneously compare negatives against negatives

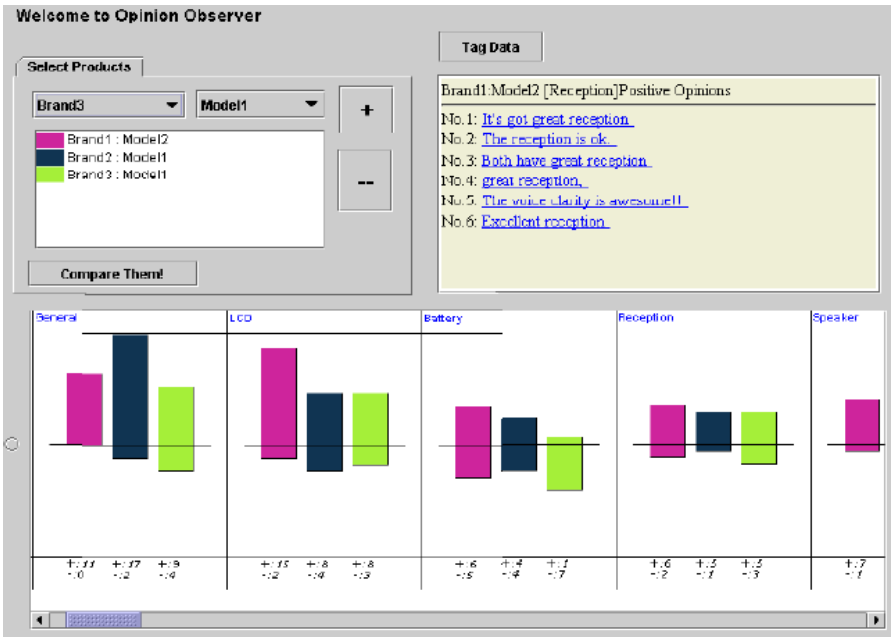

Fig. 5.6 Figure 2 of Liu et al. [191]. Three cellphones are represented, each by a different color. For each feature ("General," "LCD," etc.), three bars are shown, one for each of three cellphones. For a given feature and phone, the portions of the bar above and below the horizontal line represent the number of reviews that express positive or negative views about that camera's feature, respectively. (The system can also plot the percentages of positive or negative opinions, rather than the raw numbers.) The pane on the upper-right displays the positive sentences regarding one of the products.

and positives against positives. This simultaneous comparison is made much more difficult if the bars all have one end "planted" at the same location, as is the case for standard histograms such as the one depicted in Figure 5.7.

While the data for the features are presented sequentially in Figure 5.6 (first "General," then "LCD," and so forth), an alternative visualization technique called a *rose plot* is exemplified in Figure 5.8, which depicts a sample output of the system developed by Gregory et al. [113]. The median and quartiles across a document sub-collection of the percentage of positive and negative words per document, together with similar data for other possible affect-classification dimensions, are represented via a variant of box plots. (Adaptation to raw counts rather

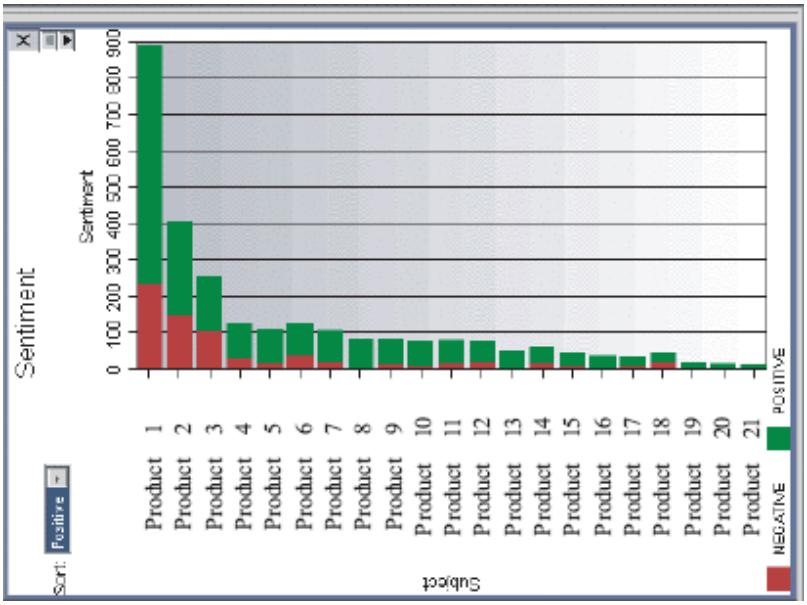

Fig. 5.7 A portion of Figure 4 of Yi and Niblack [324], rotated to save space and facilitate comparison with Figure 5.6. Notice that simultaneous comparison of the negative counts and the positive counts for two different products is not as easy as it is in Figure 5.6.

than percentages is straightforward.) Mapping this idea to product comparisons in the style of Opinion Observer, one could associate different features with different "compass directions," e.g., the feature "battery life" with "southwest," as long as the number of features being reported on is not too large. The reason that this representation might prove advantageous in some settings is that in some situations, a circular arrangement may be more compact than a sequential one, and it may be easier for a user to remember a feature as being "southwest" than as being "the fifth of eight." An additional functionality of the system that is not shown in the figure is the ability to depict how much an individual document's positive/negative percentage differs from the average for a given document group to which the document belongs. A similar circular layout is proposed in Subasic and Huettner [285] for visualizing various dimensions of affect within a single document.

Morinaga et al. [215] opt to represent degrees of association between products and opinion-indicative terms of a pre-specified polarity. First,

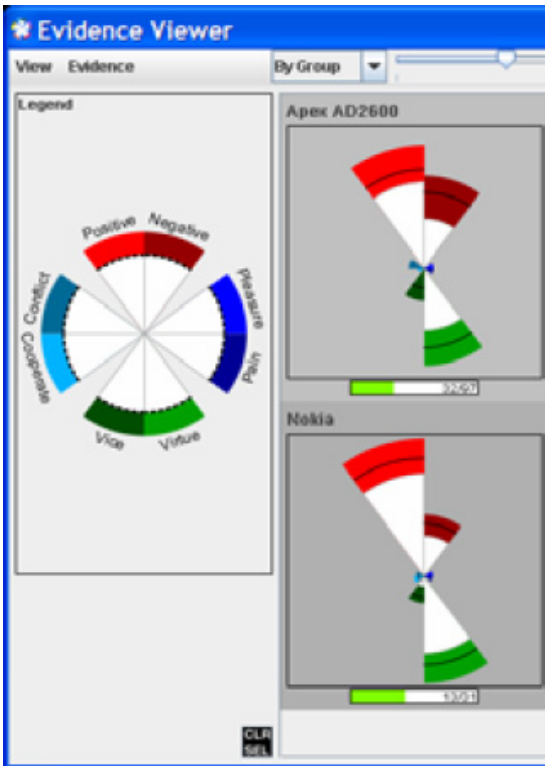

Fig. 5.8 Figure 7 of Gregory et al. [113]. On the right are two rose plots, one for each of two products; on the left is the plot's legend. In each rose plot, the upper two "petals" represent positivity and negativity, respectively (of the other six petals, the bottom two are vice and virtue, etc.). Similarly to box plots, the median value is indicated by a dark arc, and the quartiles by (colored) bands around the median arc. Darker shading for one of the two petals in a pair (e.g., "positive and negative") are meant to indicate the negative end of the spectrum for the affect dimension represented by the given petal pair. The histogram below each rose relates to the number of documents represented.

opinions are gathered using the authors' pre-existing system [291]. Coding-length and probabilistic criteria are used to determine which terms to focus on, and principal component analysis is then applied to produce a two-dimensional visualization, such that nearness corresponds to strength of association, as in the authors' previous work [184]. Thus, in Figure 5.9, we see that cellphone A is associated with what we recognize as positive terms, whereas cellphone C is associated with negative terms.

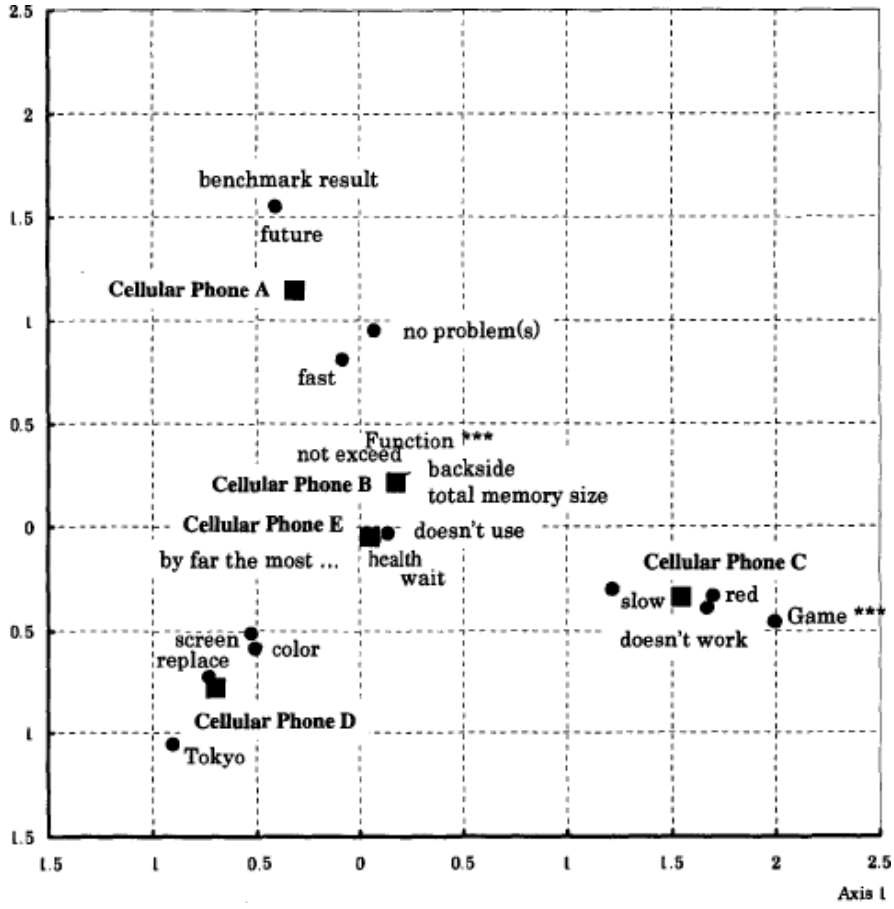

Fig. 5.9 Figure 5 of Morinaga et al. [215]: principal-components-analysis visualization of associations between products (squares) and automatically selected opinion-oriented terms (circles).

5.2.3.3 Temporal Variation and Sentiment Timelines

So far, the summaries we have considered do not explicitly incorporate any temporal dimension. However, time is often an important consideration.

First, users may wish to view individual reviews in reverse chronological order, i.e., newest first. Indeed, at the time of this writing, this is one of the two sorting options that Amazon presents.

Second, in many applications, analysts and other users are interested in tracking changes in sentiment about a product, political candidate, company, or issue over time. Clearly, one can create a sentiment timeline simply by plotting the value of a chosen summary statistic at different times; the chosen statistic can reflect the prevailing polarity [170, 296] or simply the number of mentions, in which case what is being measured is perhaps not so much public opinion, but rather public awareness [102, 197, 211, 212]. Such work is strongly related at a conceptual level to topic detection and tracking [8], a review of which is beyond the scope of this survey.

Mishne and de Rijke [212] also depict the derivative of the summary statistic considered as a function of time.

5.2.4 Review(er) Quality

> How do we identify what is good? And how do we censure what is bad? We will argue that developing a humane reputation system ecology can provide better answers to these two general questions — restraining the baser side of human nature, while liberating the human spirit to reach for ever higher goals. — "Manifesto for the reputation society." Masum and Zhang [203]

When creating summaries of reviews or opinionated text, an important type of information that deserves careful consideration is whether or not individual reviews are *helpful or useful*. For example, a system might want to downweight or even discard unhelpful reviews before creating summaries or computing aggregate statistics, as in Liu et al. [193]. Alternatively, the system could use all reviews, but provide helpfulness indicators for individual reviews as a summary of their expected utility. Indeed, non-summarization systems could use such information, too: for instance, a review-oriented search engine could rank its search results by helpfulness.

Some websites already gather helpfulness information from human readers. For example, Amazon.com annotates reviews with comments like "120 of 140 people found the following review helpful," meaning

that of the 140 people who pressed one of the "yes" or "no" buttons in response to the question "Was this review helpful to you?" — we deem these 140 people *utility evaluators* — 120 chose "yes." Similarly, the Internet Movie Database (IMDb, http://www.imdb.com) also annotates user comments with "x out of y people found the following comment useful." This similarity is perhaps not surprising due to the fact that Amazon owns IMDb, although from a research point of view, note that the two populations of users are probably at least somewhat disjoint, meaning that there might be interesting differences between the sources of data. Other sites soliciting utility evaluations include Yahoo! Movies and Yahoo! TV, which allow the user to sort reviews by helpfulness; CitySearch, which solicits utility evaluations from general users and gives more helpful reviews greater prominence; and Epinions, which only allows registered members to rate reviews and does not appear to have helpfulness as a sort criterion, at least for non-registered visitors.[5] (We learned about the solicitation of utility evaluations by IMDb from Zhuang et al. [332] and by Citysearch from Dellarocas [71].)

Despite the fact that many review-aggregation sites already provide helpfulness information gathered from human users, there are still at least two reasons why automatic helpfulness classification is a useful line of work to pursue.

Items that lack utility evaluations. Many reviews receive very few utility evaluations. For example, 38% of a sample of roughly 20,000 Amazon MP3-player reviews, and 31% of those aged at least three months, received three or fewer utility evaluations [161]. Similarly, Liu et al. [193] confirm one's prior intuitions that Amazon reviews that are youngest and reviews that are most lowly ranked (i.e., determined to be least helpful) by the site receive the fewest utility evaluations.

[5] We note that we were unable to find Amazon's definition of "helpful," and conclude that they do not supply one. In contrast, Yahoo! specifies the following: "Was [a review] informative, well written or amusing — above all was it was helpful to you in learning about the [film or show]? If so, then you should rate that review as helpful." It might be interesting to investigate whether these differing policies have implications. There have in fact been some comments that Amazon should clarify its question (http://www.amazon.com/Was-this-review-helpful-you/forum/Fx1JS1YLZ490S1O/Tx3QHE2JPEXQ1V7/1?_encoding=UTF8&asin =B000FL7CAU).

Perhaps some reviews receive no utility evaluations simply because they are so obviously bad that nobody bothers to rate them. But this does not imply that reviews without utility evaluations must necessarily be unhelpful; certainly we can not assume this of reviews too recently written to have been read by many people. One important role that automated helpfulness classifiers can play, then, is to provide utility ratings in the many cases when human evaluations are lacking.

Skew in utility evaluations. Another interesting potential application of automated helpfulness classification is to correct for biases in human-generated utility evaluations.

We first consider indirect evidence that such biases exist. It turns out that just as the distribution of ratings can often be heavily skewed toward the positive end, as discussed in Section 5.2.3.2, the distribution of utility evaluations can also be heavily skewed toward the helpful end, probably due at least in part to similar reasons as in the product-ratings case. In a crawl of approximately 4 million unique Amazon reviews for about 670,000 books (excluding alternate editions), the average percentage of "yes" responses among the utility evaluations is between 74% and 70%, depending on whether reviews with fewer than 10 utility evaluations are excluded (Gueorgi Kossinets and Cristian Danescu Niculescu-Mizil, personal communication). Similarly, half of a sample of about 23,000 Amazon digital-camera reviews had helpful/unhelpful vote ratios of over 9 to 1 [193]. As in the ratings distribution case, one's intuition is that the percentage of reviews that are truly helpful is not as high as these statistics would seem to indicate. Another type of indirect evidence of bias is that the number of utility evaluations received by a review appears to decrease exponentially in helpfulness rank as computed by Amazon [193]. (Certainly there has to be some sort of decrease, since Amazon's helpfulness ranking is based in part on the number of utility evaluations a review receives.) Liu et al. [193] conjecture that reviews that have many utility evaluations will have a disproportionate influence on readers (and utility evaluators) because they are viewed as more authoritative, but reviews could get many utility evaluations only because they are more prominently displayed, not because readers actually compared them against other reviews. (Liu et al. [193]

call this tendency for often-evaluated reviews to quickly accumulate even more utility evaluations as "winner circle" bias; in other literature on power-law effects, related phenomena are also referred to as "rich-get-richer.")

As for more direct evidence: Liu et al. [193] conduct an re-annotation study in which the Amazon reviewers' utility evaluations often did not match those of the human re-labelers. However, this latter evidence should be taken with a grain of salt. First, in some of the experiments in the study, "ground truth" helpfulness was measured by, among other things, the number of aspects of a product that are discussed by a review. Second, in all experiments, the test items appear to have consisted of only the text of a single review considered in isolation. It is not clear that the first point corresponds to the standard that all Amazon reviewers used, or should be required to use, and clearly, the second point describes an isolated-text setting that is not the one that real Amazon reviewers work in. To exemplify both these objections: a very short review written by a reputable critic (e.g., a "top reviewer") that points out something that *other* reviews missed can, indeed, be quite helpful, but would score poorly according to the specification of Liu et al. [193]. Indeed, the sample provided of a review that should be labeled "bad" starts,

> I want to point out that you should never buy a generic battery, like the person from San Diego who reviewed the S410 on May 15, 2004, was recommending. Yes you'd save money, but there have been many reports of generic batteries exploding when charged for too long.

We would view this comment, if true, to be quite helpful, despite the fact that it fails the specification. Another technical issue is that the re-labelers used a four-class categorization scheme, whereas essentially every possible percentage of positive utility evaluations could form a distinct class for the Amazon labels: it might have been better to treat reviews with helpfulness percentages of 60% and 61% as equivalent, rather than saying that Amazon reviewers rated the latter as better than the former.

Nonetheless, given the large predominance of "helpful" among utility evaluations despite the fact that anecdotal evidence we have gathered indicates that not all reviews deserve to be called "helpful," and given the suggestive results of the re-annotation experiment just described, it is likely that some of the human utility evaluations are not strongly related to the quality of the review at hand. Thus, we believe that correction of these utility evaluations by automatic means is a valid potential application.

A note regarding the effect of utility evaluations. It is important to mention one caveat before proceeding to describe research in this area. Park et al. [236] attempted to determine what the effect of review quality actually is on purchasing intention, running a study in which subjects engaged in hypothetical buying behavior. They found non-uniform effects: "low-involvement [i.e., motivated] consumers are affected by the quantity rather than the quality of reviews ... high-involvement consumers are affected by review quantity mainly when the review quality is high...The effect of review quality on high-involvement consumers is more pronounced with a sizable number of reviews, whereas the effect of review quantity is significant even when the review quality is low." (More on the economic impacts of sentiment analysis is described in Section 6.)

5.2.4.1 Methods for Automatically Determining Review Quality

In a way, one could consider the review-quality determination problem as a type of readability assessment and apply essay-scoring techniques [19, 99]. However, while some of the systems described below do try to take into account some readability-related features, they are tailored specifically to product reviews.

Kim et al. [161], Zhang and Varadarajan [328], and Ghose and Ipeirotis [106] attempt to automatically rank certain sets of reviews on the Amazon.com website according to their helpfulness or utility, using a regression formulation of the problem. The domains considered are a bit different: MP3 players and digital cameras in the first case; Canon electronics, engineering books, and PG-13 movies in the

second case; and AV players plus digital cameras in the third case. Liu et al. [193] convert the problem into one of low-quality review detection (i.e., binary classification), experimenting mostly with manually (re-)annotated reviews of digital cameras, although CNet editorial ratings were also considered on the assumption that these can be considered trustworthy. Rubin and Liddy [261] also sketch a proposal to consider whether reviews can be considered credible.

Kim et al. [161] study which of a multitude of length-based, lexical, POS-count, product-aspect-mention count, and metadata features are most effective when utilizing SVM regression. The best feature combination turned out to be review length plus tf-idf scores for lemmatized unigrams in the review plus the number of "stars" the reviewer assigned to the product. Somewhat disappointingly, the best pair of features among these was the length of the review and the number of stars. (Using "number of stars" as the only feature yielded similar results to using just the deviation of the number of stars given by the particular reviewer from the average number of stars granted by all reviewers for the item.) The effectiveness of using all unigrams appears to subsume that of using a select subset, such as sentiment-bearing words from the General Inquirer lexicon [281].

Zhang and Varadarajan [328] use a different feature set. They employ a finer classification of lexical types, and more sources for subjective terms, but do not include any meta-data information. Interestingly, they also consider the similarity between the review in question and the product specification, on the premise that a good review should discuss many aspects of the product; and they include the review's similarity to editorial reviews, on the premise that editorial reviews represent high-quality examples of opinion-oriented text. (David and Pinch [70] observe, however, that editorial reviews for books are paid for and are meant to induce sales of the book.) However, these latter two original features do not appear to enhance performance. The features that appear to contribute the most are the class of shallow syntactic features, which, the authors speculate, seem to characterize style; examples include counts of words, sentences, wh-words, comparatives and superlatives, proper nouns, etc. Review length seems to be very weakly correlated with utility score.

We thus see that Kim et al. [161] find that meta-data and very simple term statistics suffice, whereas Zhang and Varadarajan [328] observe that more sophisticated cues that appear correlated with linguistic aspects appear to be most important. Possibly, the difference is a result of the difference in domain choice: we speculate that book and movie reviews can involve more sophisticated language use than what is exhibited in reviews of electronics.

Declaring themselves influenced by prior work on creating subjectivity extracts [232], Ghose and Ipeirotis [106] take a different approach. They focus on the relationship between the subjectivity of a review and its helpfulness. The basis for measuring review subjectivity is as follows: using a classifier that outputs the probability of a sentence being subjective, one can compute for a given review the average subjectiveness-probability over all its sentences, or the standard deviation of the subjectivity scores of the sentences within the review. They found that both the standard deviation of the sentence subjectivity scores and a readability score (review length in characters divided by number of sentences) have a strongly statistically significant effect on utility evaluations, and that this is sometimes true of the average subjectiveness-probability as well. They then suggest on the basis of this and other evidence that it is extreme reviews that are considered to be most helpful, and develop a helpfulness predictor based on their analysis.

Liu et al. [193] considered features related to review and sentence length; brand, product and product-aspect mentions, with special consideration for appearances in review titles; sentence subjectivity and polarity; and "paragraph structure." This latter refers to paragraphs as delimited by automatically determined keywords. Interestingly, the technique of taking the 30 most frequent pairs of nouns or noun phrases that appear at the beginning of a paragraph as keywords yields separator pairs such as "pros"/"cons," "strength"/"weakness," and "the upsides"/"downsides." (Note that this differs from identifying pro or con reasons themselves [157], or identifying the polarity of sentences. Note also that other authors have claimed that different techniques are needed for situations in which pro/con delimiters are mandated by the format imposed by a review aggregation site

but a separate detailed textual description must also be included, as in Epinions, as opposed to settings where such delimiters need not be present or where all text is placed in the context of such delimiters [191].) Somewhat unconventionally with respect to other text-categorization work, the baseline was taken as $\text{SVM}^{\text{light}}$ run with three sentence-level statistics as features; that is, the performance of a classifier trained using bag-of-word features is not reported. Given this unconventional starting point, the addition of the features that do not reflect subjectivity or sentiment help. Including subjectivity and polarity on top of what has already been mentioned does not yield further improvement, and use of title-appearance for mentions did not seem to help.

Review- or opinion-spam detection — the identification of deliberately misleading reviews — is a line of work by Jindal and Liu ([141], short version available as Jindal and Liu [140]) in the same vein. One challenge these researchers faced was the difficulty in obtaining ground truth. Therefore, for experimental purposes they first re-framed the problem as one of trying to recognize duplicate reviews, since a priori it is hard to see why posting repeats of reviews is justified. (However, one potential problem with the assumption that repeated reviews constitute some sort of manipulation attempt, at least for the Amazon data that was considered, is that Amazon itself cross-posts reviews across different products — where "different" includes different instantiations (e.g., e-book vs. hardcover) or subsequent editions of the same item (Gueorgi Kossinets and Cristian Danescu Niculescu-Mizil, personal communication). Specifically, in a sample of over 1 million Amazon book reviews, about one-third were duplicates, but these were all due to Amazon's cross-posting. Human error (e.g., accidentally hitting the "submit" button twice) causes other cases of non-malicious duplicates.) A second round of experiments attempted to identify "reviews on brands only," ads, and "other irrelevant reviews containing no opinions" (e.g., questions, answers, and random texts). Some of the features used were similar to those employed in the studies described above; others included features on the review author and the utility evaluations themselves. The overall message was that this kind of spam is relatively easy to detect.

5.2.4.2 Reviewer-Identity Considerations

In the above, we have discussed determining the quality of individual reviews. An alternate approach is to look at the quality of the reviewers; doing so can be thought of as a way of classifying all the reviews authored by the same person at once.

Interestingly, one study has found that there is a real economic effect to be observed when factoring in reviewer credibility: Gu et al. [114] note that a weighted average of message-board postings in which poster credibility is factored in has "prediction power over future abnormal returns of the stock," but if postings are weighted uniformly, the predictive power disappears.

There has been work in a number of areas in the human-language-technologies community that incorporates the authority, trustworthiness, influentialness, or credibility of authors [94, 96, 141, 275]. PageRank [44, 241] and hubs and authorities (also known as HITS) [163] are very influential examples of work in link analysis on identifying items of great importance. Trust metrics also appear in other work, such as research into peer-to-peer and reputation networks and information credibility [71, 115, 147, 174, 252].

6

Broader Implications

> Sentiment is the mightiest force in civilization . . . —
> J. Ellen Foster, *What America Owes to Women*, 1893

As we have seen, sentiment-analysis technologies have many potential applications. In this section, we briefly discuss some of the larger implications that the existence of opinion-oriented information-access services has.

Privacy. One point that should be mentioned is that applications that gather data about people's preferences can trigger concerns about privacy violations. We suspect that in many people's minds, having one's public blog scanned by a coffee company for positive mentions of its product is one thing; having one's cell-phone conversations monitored by the ruling party of one's own country for negative mentions of government officials is quite another. It is not our intent to comment further here on privacy issues, these not being issues on which we are qualified to speak; rather, we simply want to be thorough by reminding the reader that these issues do exist and are important, and that these concerns apply to all data-mining technologies in general.

Manipulation. But even if we restrict attention to the apparently fairly harmless domain of business intelligence, certain questions regarding the potential for manipulation do arise. Companies already participate in managing online perceptions as part of the normal course of public-relations efforts:

> . . . companies can't control consumer-generated content. They can, however, pay close attention to it. In many cases, often to a large degree, they can even influence it. In fact, in a survey conducted by Aberdeen [of "more than 250 enterprises using social media monitoring and analysis solutions in a diverse set of enterprises"], more than twice as many companies with social media monitoring capabilities actively contribute to consumer conversations than remain passive observers (67% versus 33%). Over a third of all companies (39%) contribute to online conversations on a frequent basis, interacting with consumers in an effort to sway opinion, correct misinformation, solicit feedback, reward loyalty, test new ideas, or for any number of other reasons.
>
> — Zabin and Jefferies [327]

And it is also the case that some arguably mild forms of manipulation have been suggested. For instance, one set of authors, in studying the strategic implications for a company of offering online consumer reviews, notes that "if it is possible for the seller to decide the timing to offer consumer reviews at the individual product level, it may not always be optimal to offer consumer reviews at a very early stage of new product introduction, even if such reviews are available" ([57], quotation from the July 2004 working-paper version), and others have worked on a manufacturer-oriented system that ranks reviews "according to their expected effect on sales," noting that these might not be the ones that are considered to be most helpful to users [106].

But still, there are concerns that corporations might try to further "game the system" by taking advantage of knowledge of how ranking systems work in order to suppress negative publicity [124] or engage in other so-called "black-hat search engine optimization" and related

activities. Indeed, there has already been a term — "sock puppet" — coined to refer to ostensibly distinct online identities created to give the false impression of external support for a position or opinion; Stone and Richtel [280] list several rather attention-grabbing examples of well-known writers and CEOs engaging in sock-puppetry. On a related note, Das and Chen [67] recommend Leinweber and Madhavan [183] as an interesting review of the history of market manipulation through disinformation.

One reason these potentials for abuse are relevant to this survey is that, as pointed out earlier in the Introduction, sentiment-analysis technologies allow users to consult many people who are unknown to them; but this means precisely that it is harder for users to evaluate the trustworthiness of those people (or "people") they are consulting. Thus, opinion-mining systems might potentially make it easier for users to be mis-led by malicious entities, a problem that designers of such systems might wish to prevent. On the flip side, an information-access system that is (perhaps unfairly) perceived to be vulnerable to manipulation is one that is unlikely to be widely used; thus, again, builders of such systems might wish to take measures to make it difficult to "game the system."

In the remainder of this section, then, we discuss several aspects of the problem of possible manipulation of reputation. In particular, we look at evidence as to whether reviews have a demonstrable economic impact: if reviews do significantly affect customer purchases, then there is arguably an economic incentive for companies to engage in untoward measures to manipulate public perception; if reviews do not significantly affect customer purchases, then there is little reason, from an economic point of view, for entities to try to artificially change the output of sentiment-analysis systems — or, as Dewally [74] asserts, "the stock market does not appear to react to these recommendations. ... The fears raised by the media about the destabilizing power of such traders who participate in these discussions are thus groundless." If such claims are true, then it would seem that trying to manipulate perceptions conveyed by online review-access systems would offer little advantages to companies, and so they would not engage in it.

6.1 Economic Impact of Reviews

As mentioned earlier in the Introduction to this survey, many readers of online reviews say that these reviews significantly influence their purchasing decisions [63]. However, while these readers may have believed that they were "significantly influenced," perception and reality can differ. A key reason to understand the real economic impact of reviews is that the results of such an analysis have important implications for how much effort companies might or should want to expend on online reputation monitoring and management.

Given the rise of online commerce, it is not surprising that a body of work centered within the economics and marketing literature studies the question of whether the polarity (often referred to as "valence") and/or volume of reviews available online have a measurable, significant influence on actual consumer purchasing. Ever since the classic "market for lemons" paper [6] demonstrating some problems for makers of high-quality goods, economists have looked at the value of maintaining a good reputation as a means to overcome these problems [77, 162, 268, 269], among other strategies. (See the introduction to Dewally and Ederington [75], from which the above references have been taken, for a brief review.) One way to acquire a good reputation is, of course, by receiving many positive reviews of oneself as a merchant; another is for the products one offers to receive many positive reviews. For the purposes of our discussion, we regard experiments wherein the buying is hypothetical as being out of scope; instead, we focus on economic analyses of the behavior of people engaged in real shopping and spending real money.[1]

[1] Note that researchers in the economics community have a tradition of circulating and revising working papers, sometimes for years, before producing an archival version. In the references that follow, we have cited the archival version when journal-version publication data has been available to us, in order to enable the interested reader to access the final, peer-reviewed version of the work. But because of this policy, the reader who wishes to delve into this literature further should keep in mind the following two points. First, many citations within the literature are to preliminary working papers. This means that our citations may not precisely match those given in the papers themselves (e.g., there may be title mismatches). Second, work that was done earlier may be cited with a later publication date; therefore, the dates given in our citations should not be taken to indicate research precedence.

The general form that most studies take is to use some form of hedonic regression [259] to analyze the value and the significance of different item features to some function, such as a measure of utility to the customer, using previously recorded data. (Exceptions include Resnick et al. [253], who ran an empirical experiment creating "new" sellers on eBay, and Jin and Kato [136], who made actual purchases to validate seller claims.) Specific economic functions that have been examined include revenue (box-office take, sales rank on Amazon, etc.), revenue growth, stock trading volume, and measures that auction-sites like eBay make available, such as bid price or probability of a bid or sale being made. The type of product considered varies (although, understandably, those offered by eBay and Amazon have received more attention): examples include books, collectible coins, movies, craft beer, stocks, and used cars. It is important to note that some conclusions drawn from one domain often do not carry over to another; for instance, reviews seem to be influential for big-ticket items but less so for cheaper items. But there are also conflicting findings within the same domain. Moreover, different subsegments of the consumer population may react differently: for example, people who are more highly motivated to purchase may take ratings more seriously. Additionally, in some studies, positive ratings have an effect but negative ones do not, and in other studies the opposite effect is seen; the timing of such feedback and various characteristics of the merchant or of the feedback itself (e.g., volume) may also be a factor.

Nonetheless, to gloss over many details for the sake of brevity: if one allows any effect — including correlation even if said correlation is shown to be not predictive — that passes a statistical significance test at the 0.05 level to be classed as "significant," then many studies find that review polarity has a significant economic effect [13, 14, 23, 31, 35, 47, 59, 62, 68, 72, 75, 76, 81, 82, 128, 136, 145, 180, 195, 196, 198, 207, 208, 214, 237, 250, 253, 278, 297, 331]. But there are a few studies that conclude emphatically that review positivity or negativity has no significant economic effect [56, 74, 80, 87, 100, 194, 325]. Duan et al. [80] explicitly relate their findings to the issue of corporate manipulation: "From the managerial perspective, we show that consumers are rational in inferring movie quality from online user

reviews without being unduly influenced by the rating, thus presenting a challenge to businesses that try to influence sales through 'planting' online word-of-mouth."

With respect to effects that have been found, the literature survey contained in Resnick et al. [253] states that

> At the larger end of effect sizes for positive evaluations, the model in [Livingston [196]] finds that sellers with more than 675 positive comments earned a premium of $45.76, more than 10% of the mean selling price, as compared to new sellers with no feedback. ... At the larger end of effect sizes for negatives, [Lucking-Reiley et al. [198]], looking at collectible coins, finds that a move from 2 to 3 negatives cuts the price by 11%, about $19 from a mean price of $173.

But in general, the claims of statistically significant effects that have been made tend to be (a) qualified by a number of important caveats, and (b) quite small in absolute terms per item, although on the other hand again, small effects per item can add up when many items are involved. With regard to this discussion, the following excerpt from Houser and Wooders [128] is perhaps illuminating:

> ... on average, 3.46 percent of sales is attributable to the seller's positive reputation stock. Similarly, our estimates imply that the average cost to sellers stemming from neutral or negative reputation scores is $2.28, or 0.93 percent of the final sales price. If these percentages are applied to all of eBay's auctions ($1.6 billion in the fourth quarter of 2000), this would imply that sellers' positive reputations added more than $55 million to the value of sales, while non-positives reduced sales by about $15 million.

Ignoring for the moment the fact that, as mentioned above, other papers report differing or even opposite findings, we simply note that the choice of whether to focus on "0.93%," "$2.28," or "$55 million"

(and whether to view the latter amount as seeming particularly large or not) is one we prefer to leave to the reader.

Let us now mention some particular papers and findings of particular interest.

6.1.1 Surveys Summarizing Relevant Economic Literature

Resnick et al. [253] and Bajari and Hortaçsu [24] are good entry points into this body of literature. They provide very thorough overviews and discussion of the methodological issues underlying the studies mentioned above. Hankin [118] supplies several visual summaries that are modeled after the literature-comparison tables in Dellarocas [71], Resnick et al. [253], and Bajari and Hortaçsu [24]. A list of a number of papers on the general concept of sentiment in behavioral finance can be found at http://sentiment.behaviouralfinance.net/.

6.1.2 Economic-Impact Studies Employing Automated Text Analysis

In most of the studies cited above, the orientation of a review was derived from an explicit rating indication such as number of stars, but a few studies applied manual or automatic sentiment classification to review text [13, 14, 35, 47, 67, 68, 214, 237].

At least one related set of studies claims that "the text of the reviews contains information that influences the behavior of the consumers, and that the numeric ratings alone cannot capture the information in the text" [106] — see also Ghose et al. [107], who additionally attempt to assign a "dollar value" to various adjective-noun pairs, adverb-verb pairs, or similar lexical configurations. In a related vein, Pavlou and Dimoka [237] suggest that "the apparent success of feedback mechanisms to facilitate transactions among strangers does not mainly come from their crude numerical ratings, but rather from their rich feedback text comments." Also, Chevalier and Mayzlin [59] interpret their findings on the effect of review length as providing some evidence that people do read the reviews rather than simply relying on numerical ratings.

On the other hand, Cabral and Hortaçsu [47], in an interesting experiment, look at 41 odd cases of feedback on sellers posted on eBay: what was unusual was that the feedback text was clearly positive, but the numerical rating was negative (presumably due to user error). Analysis reveals that these reviews have a strongly significant ("both economically and statistically") detrimental effect on sales growth rate — indicating that customers seemed to ignore the text in favor of the incorrect summary information.

In some of these text-based studies, what was analyzed was not sentiment per se but the degree of polarization (disagreement) among a set of opinionated documents [13, 68] or, inspired in part by Pang and Lee [233], the average probability of a sentence being subjective within a given review [106]. Ghose and Ipeirotis [106] also take into account the standard deviation for sentence subjectivity within a review, in order to examine whether reviews containing a mix of subjective and objective sentences seem to have different effects from reviews that are mostly purely subjective or purely objective.

Some initially unexpected text effects are occasionally reported. For example, Archak et al. [14] found that "amazing camera," "excellent camera," and related phrases have a negative effect on demand. They hypothesize that consumers consider such phrases, especially if few details are subsequently furnished in the review, to indicate hyperbole and hence view the review itself as untrustworthy. Similarly, Archak et al. [14] and Ghose et al. [107] discover that apparently positive comments like "decent quality" or "good packaging" also had a negative effect, and hypothesize that the very fact that many reviews contain hyperbolic language mean that words like "decent" are interpreted as lukewarm.

These findings might seem pertinent to the distinction between the *prior* polarity and the *contextual* polarity of terms and phrases, borrowing the terminology of Wilson et al. [319]. Prior polarity refers to the sentiment a term evokes in isolation, as opposed to the sentiment the term evokes within a particular surrounding context; Polanyi and Zaenen [242] point out that identifying prior polarity alone may not suffice. With respect to this distinction, the status of the observations of Archak et al. [14] just mentioned is not entirely clear. The superlatives

("amazing") are clearly intended to convey positive sentiment regardless of whether the review authors actually managed to convince readers; that is, context is only needed to explain the economic effect of lowered sales, not the interpretation of the review itself. In the case of words like "decent," one could potentially make the case that the prior orientation of such words is in fact neutral rather than positive; but alternatively, one could argue instead that in a setting where many reviews are highly enthusiastic, the contextual orientation of "decent" is indeed different from its prior orientation.

6.1.3 Interactions with Word of Mouth (WOM)

One factor that some studies point out is that the number of reviews, positive or negative, may simply reflect "word of mouth," so that in some cases, what is really the underlying correlative (if any) of economic impact is not the amount of positive feedback per se but merely the amount of feedback in total. This explains why in some settings (but not all), negative feedback is seen to "increase" sales: the increased "buzz" brings more attention to the product (or perhaps simply indicates more attention is being paid to the product, in which case it would not be predictive per se).

6.2 Implications for Manipulation

Regarding the incentives for manipulation, it is difficult to draw a conclusion one way or the other from the studies we have just examined.

One cautious way to read the results summarized in the previous section is as follows. While there may be some economic benefit in some settings for a corporation to plant positive reviews or otherwise attempt to use untoward means to manufacture an artificially inflated reputation or suppress negative information, it seems that in general, a great deal of effort and resources would be required to do so for perhaps fairly marginal returns. More work is clearly required, though; as Bajari and Hortaçsu [24] conclude, "There is still plenty of work to be done to understand how market participants utilize the information contained in the feedback forum system." Surveying the state of the art in this subject is beyond the scope of this survey; a fairly concise

review of issues regarding online reputation systems may be found in Dellarocas [71].

We would like to conclude, though, by pointing out a result that indicates that even if illegitimate reviews do get through, opinion-mining systems can still be valuable to consumers. Awerbuch and Kleinberg [22] study the "competitive collaborative learning" setting in which some of the n users are assumed to be "Byzantine" (malicious, dishonest, coordinated, and able to eavesdrop on communications), and product or resource quality varies over time. The authors formulate the product selection problem as a type of "multi-armed bandit" problem. They show the striking result that even if *only a constant fraction* of users are honest and (unbeknownst to them) grouped into k market segments such that all members of a block share the same product preferences — with the implication that the recommendations of an honest user may be useless to honest users in different market segments — then there is still an algorithm by which, in time polynomial in $k \log(n)$, the average regret per honest user is arbitrarily small (assuming that the number of products or resources on offer is $O(n)$). Roughly speaking, the algorithm causes users to tend to raise the probability of getting recommendations from valuable sources. Thus, even in the face of rather stiff odds and formidable adversaries, honest users can — at least in theory — still get good advice from sentiment-analysis systems.

7

Publicly Available Resources

7.1 Datasets

7.1.1 Acquiring Labels for Data

One source of opinion, sentiment, and subjectivity labels is, of course, manual annotation [172, 309].

However, researchers in the field have also managed to find ways to avoid manual annotation by leveraging pre-existing resources. A common technique is to use labels that have been manually assigned, but not by the experimenters themselves; this explains why researchers in opinion mining and sentiment analysis have taken advantage of Rotten Tomatoes, Epinions, Amazon, and other sites where users furnish ratings along with their reviews. Some other noteworthy techniques are as follows:

- Sentiment summaries can be gathered by treating the review snippets that Rotten Tomatoes furnishes as one-sentence summaries [33].
- Subjective vs. non-subjective texts on the same topic can be gathered by selecting editorials versus non-editorial newswire

[308, 326] or by selecting movie reviews versus plot summaries [222, 232].

- If sentiment-oriented search engines already exist (one example used to be Opinmind), then one can issue topical queries to such search engines and harvest the results to get sentiment-bearing sentences more or less guaranteed to be on-topic [206]. (On the other hand, there is something circular about this approach, since it bootstraps off of someone else's solution to the opinion-mining problem.)
- One might be able to derive affect labels from emoticons [249].
- Text polarity may be inferred from correlations with stock-market behavior or other economic indicators [168, 107].
- Viewpoint labels can be derived from images of party logos that users display [160].
- Negative opinions can be gathered by assuming that when one newsgroup post cites another, it is typically done to indicate negative sentiment toward the cited post [4]. A more refined approach takes into account indications of "shouting," such as text rendered all in capital letters [110].

One point to mention with regards to sites where users rate the contributions of other users — such as the examples of Amazon and Epinions mentioned above — is a potential bias toward positive scores [59, 74, 128, 132, 240, 253], as we have mentioned above. In some cases, this comes about because of sociological effects. For example, Pinch and Athanasiades [240], in a study of a music-oriented site called ACIDplanet, found that various forces tend to cause users to give high ratings to each other's music. The users themselves refer to this phenomenon as "R=R" (review me and I will review you), among other, less polite, names, and the ACIDplanet administrators introduced a form of anonymous reviewing to avoid this issue in certain scenarios.

Thus, there is the question of whether one can trust the automatically determined labels that one is training one's classifiers upon. (After all, you often get what you pay for, as they say.) Indeed, Liu et al. [193] essentially re-labeled their review-quality Amazon data due to concerns

about bias, as discussed in Section 5.2.4. On the other hand, while this phenomenon implies that reviewers may not always be sincere, we hypothesize that this phenomenon does not greatly affect the quality of the authors' meta-data labels at reflecting the intended sentiment of the review itself. That is, we hypothesize that in many cases one can still trust the review's label, even if one does not trust the review.

7.1.2 An Annotated List of Datasets

The following list is in alphabetical order.

Blog06

[registration and fee required]
The University of Glasgow distributes this 25GB TREC test collection, consisting of blog posts over a range of topics. Access information is available at http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html. Included in the data set are "top blogs" that were provided by Nielsen BuzzMetrics and "supplemented by the University of Amsterdam" [227], and some spam blogs, also known as "splogs," that were planted in the corpus in order to simulate a more realistic setting. Assessments include relevance judgments and labels as to whether posts contain relevant opinions and what the polarity of the opinions was (positive, negative, or a mixture of both). Macdonald and Ounis [199] give more details on the creation of the corpus and the collection's features, and include some comparison with another collection of blog postings, the BlogPulse dataset (contact information can be found on the following agreement form: http://www.blogpulse.com/www2006-workshop/datashare-agreement.pdf, but it may be out of date).

Congressional floor-debate transcripts

URL: http://www.cs.cornell.edu/home/llee/data/convote.html
This dataset, first introduced in Thomas et al. [294], includes speeches as individual documents together with:

- Automatically derived labels for whether the speaker supported or opposed the legislation discussed in the debate the speech appears in, allowing for experiments with this kind of sentiment analysis.

- Indications of which "debate" each speech comes from, allowing for consideration of conversational structure.
- Indications of by-name references between speakers, allowing for experiments on *agreement classification* if one assigns gold-standard agreement labels from the support/oppose labels assigned to the pair of speakers in question.
- The edge weights and other information derived to create the graphs used in Thomas et al. [294], facilitating implementation of alternative graph-based methods upon the graphs constructed in that earlier work.

Cornell movie-review datasets

URL: http://www.cs.cornell.edu/people/pabo/movie-review-data/
These corpora, first introduced in Pang and Lee [232, 233], consist of the following datasets, which include automatically derived labels.

- Sentiment polarity datasets:
  - document-level: polarity dataset v2.0: 1000 positive and 1000 negative processed reviews. (An earlier version of this dataset (v1.0) was first introduced in Pang et al. [235].)
  - sentence-level: sentence polarity dataset v1.0: 5331 positive and 5331 negative processed sentences/ snippets.
- Sentiment-scale datasets: scale dataset v1.0: a collection of documents whose labels come from a rating scale.
- Subjectivity dataset v1.0: 5000 subjective and 5000 objective processed sentences.

We should point out that the existence of the polarity-based datasets does not indicate that the curators (i.e., us) believe that reviews with middling ratings are not important to consider in practice (indeed, the sentiment-scale corpora contain such documents). Rather, the rationale in creating the polarity dataset was as follows. At the

time the corpus creation was begun, the application of machine learning techniques to sentiment classification was very new, and, as discussed in Section 3, it was natural to assume that the problem could be very challenging to such techniques. Therefore, the polarity corpus was constructed to be as "easy" for text-categorization techniques as possible: the documents fell into one of two well-separated and size-balanced categories. The point was, then, to use this corpus as a lens to study the relative difficulty of sentiment polarity classification as compared to standard topic-based classification, where two-balanced-class problems with well-separated categories pose very little challenge.

A list of papers that use or report performance on the Cornell movie-review datasets can be found at http://www.cs.cornell.edu/people/pabo/movie-review-data/otherexperiments.html.

Customer review datasets

URL: http://www.cs.uic.edu/~liub/FBS/CustomerReviewData.zip
This dataset, introduced in Hu and Liu [129], consists of reviews of five electronics products downloaded from Amazon and Cnet. The sentences have been manually labeled as to whether an opinion is expressed, and if so, what feature from a pre-defined list is being evaluated. An addendum with nine products is also available (http://www.cs.uic.edu/~liub/FBS/Reviews-9-products.rar) and has been utilized in recent work [78]. The curator, Bing Liu, also distributes a comparative-sentence dataset that is available by request.

Economining

URL: http://economining.stern.nyu.edu/datasets.html
This site, hosted by the Stern School at New York University, consists of three sets of data:

- Transactions and price premiums.
- Feedback postings for merchants at Amazon.com.
- Automatically derived sentiment scores for frequent evaluation phrases at Amazon.com.

These formed the basis for the work reported in Ghose et al. [107], which focuses on interactions between sentiment, subjectivity, and economic indicators.

French sentences

URL: http://www.psor.ucl.ac.be/personal/yb/Resource.html
This dataset, introduced in Bestgen et al. [36], consists of 702 sentences from a Belgian–French newspaper, with labels assigned by ten judges as to unpleasant, neutral or pleasant content, using a seven-point scale.

MPQA Corpus

URL: http://www.cs.pitt.edu/mpqa/databaserelease/
The MPQA Opinion Corpus contains 535 news articles from a wide variety of news sources, manually annotated at the sentential and subsentential level for opinions and other private states (i.e., beliefs, emotions, sentiments, speculations, and so on). Wiebe et al. [309] describes the overall annotation scheme; Wilson et al. [319] describes the contextual polarity annotations and an agreement study.

Multiple-aspect restaurant reviews

URL: http://people.csail.mit.edu/bsnyder/naacl07
The corpus, introduced in Snyder and Barzilay [272], consists of 4,488 reviews, both in raw-text and in feature-vector form. Each review gives an explicit 1-to-5 rating for five different aspects — food, ambiance, service, value, and overall experience — along with the text of the review itself, all provided by the review author. A rating of five was the most common over all aspects, and Snyder and Barzilay [272] report that 30.5% of the 3,488 reviews in their randomly selected training set had a rating of five for all five aspects, although no other tuple of ratings was represented by more than 5% of the training set. The code used in Snyder and Barzilay [272] is also distributed at the aforementioned URL. The original source for the reviews was http://www.we8there.com/; data from the same website was also used by Higashinaka et al. [122].

Multi-Domain Sentiment Dataset

URL: http://www.cis.upenn.edu/~mdredze/datasets/sentiment/
This dataset, introduced in Blitzer et al. [40], consists of product

reviews from several different product types taken from Amazon.com, some with 1-to-5 star labels, some unlabeled.

NTCIR multilingual corpus

[registration required]
The corpus for the NTCIR 6 pilot task consists of news articles in Japanese, Chinese, and English and formed the basis of the Opinion Analysis Task at NTCIR6 [267]. The training data contains annotations regarding opinion holders, the opinions held by opinion holder, and sentiment polarity, as well as relevance information for a set of predetermined topics.

The corpus of the NTCIR Multilingual Opinion-Analysis Task (MOAT) is drawn from Japanese, Chinese, and English blogs.

Review-search results sets

URL: http://www.cs.cornell.edu/home/llee/data/search-subj.html
This corpus, used by Pang and Lee [234], consists of the top 20 results returned by the Yahoo! search engine in response to each of a set of 69 queries containing the word "review." The queries were drawn from the publicly available list of real MSN users' queries released for the 2005 KDD Cup competition [185]; the KDD data itself is available at http://www.acm.org/sigs/sigkdd/kdd2005/Labeled800Queries.zip. The search-engine results in the corpus are annotated as to whether they are subjective or not. Note that "sales pitches" were marked objective on the premise that they represent biased reviews that users might wish to avoid seeing.

7.2 Evaluation Campaigns

7.2.1 TREC Opinion-Related Competitions

The "TREC-BLOG" wiki, http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG/, is a useful source of information on the competitions sketched below.

TREC 2006 Blog Track. TREC 2006 involved a Blog track, with an opinion retrieval task designed precisely to focus on the opinionated character that many blogs have: participating systems had to retrieve

blog posts expressing an opinion about a specified topic. Fourteen groups participated; Ounis et al. [227] give an overview of the results. Some findings are as follows. With respect to performance on opinion detection, the participating systems seemed to fall into two groups. Opinion-detection ability and relevance-determination ability seemed to be strongly correlated. While the best systems were about equally good at detecting negative sentiment as positive sentiment, systems performing at the median seemed to be a bit more effective at locating documents with negative sentiment. Most participants followed a pipelined approach, where first topic relevance was tackled, and then opinion detection was applied upon the results. Perhaps the most surprising observation was that the organizers discovered that it was possible to achieve very good relative performance by omitting the second phase of the pipeline; but we take heart in the fact that the field is still relatively young and has room to grow and mature.

TREC 2007 Blog Track. The TREC 2007 Blog track retained the opinion retrieval task and instituted determining the sentiment status (positive, negative, or mixed) of the retrieved opinions as a subtask. The 2007 and 2006 Blog Track results are analyzed in Ounis et al. [228]. They found that lexicon-based approaches — either where the discriminativeness of terms was determined on labeled training data or where the terms were manually compiled — constituted the main effective approaches.

TREC 2008 Blog Track. In the TREC 2008 Blog track, the polarity-identification problem was re-posed as one of ranking of positive-polarity retrieved documents by degree of positivity, and, similarly, ranking of negative-polarity retrieved documents by degree of negativity. ("Mixed opinionated documents" were not to be included in these rankings.)

7.2.2 NTCIR Opinion-Related Competitions

The National Institute of Informatics (NII) runs annual meetings codenamed NTCIR (NII Test Collection for Information Retrieval Systems). Opinion analysis was featured at an NTCIR-5 workshop, and served as a pilot task at NTCIR-6 and a full-blown task at NTCIR-7.

NTCIR-6 opinion analysis pilot task. The dataset consists of newswire documents in Chinese, Japanese, and English; the organizers describe this as "what we believe to be the first multilingual opinion analysis data set over comparable data" [93]. The four constituent tasks, intentionally designed to be fairly simple so as to encourage participation from many groups, were as follows:

- Detection of opinionated sentences.
- Detection of opinion holders.
- (optional) Polarity labeling of opinionated sentences as positive, negative, or neutral.
- (optional) Detection of sentences relevant to a given topic.

Due to variation in annotator labelings, two evaluation standards were defined. In the strict evaluation, an answer is considered correct if all three annotators agreed on it. In the lenient evaluation, only a majority (i.e., two) of the annotators were required to agree with an answer for it to be considered correct.

Seki et al. [267] give an overview and the results of this evaluation exercise, noting that differences between languages make direct comparison difficult, especially since precision and recall were defined (slightly) differently across languages. A shortened version of this overview also exists [93].

NTCIR-7 Multilingual opinion analysis task (MOAT), 2008. Subsequent to the NTCIR-6 pilot task, a new dataset was selected, drawn from blogs in Japanese, traditional and simplified Chinese, and English; according to the organizers, "We plan to select and balance useful topics for opinion mining researchers, such as topics concerning product reviews, movie reviews, and so on." This exercise involves six subtasks:

- Detection of opinionated sentences and opinion fragments within opinionated sentences.
- Polarity labeling of opinion fragments as positive, negative or neutral.
- (optional) Strength labeling of opinion fragments as very weak, average, or very strong.

- (optional) Detection of opinion holders.
- (optional) Detection of opinion targets.
- (optional) Detection of sentences that are relevant to a given topic.

As in the previous competition, both strict and lenient evaluation standards are to be applied.

OpQA Corpus

[available by request]
Stoyanov et al. [283] describes the construction of this corpus, which is a collection of opinion questions and answers together with 98 documents selected from the MPQA dataset.

7.3 Lexical Resources

The following list is in alphabetical order.

General Inquirer

URL: http://www.wjh.harvard.edu/~inquirer/
This site provides entry-points to various resources associated with the General Inquirer [281]. Included are manually-classified terms labeled with various types of positive or negative semantic orientation, and words having to do with agreement or disagreement.

NTU Sentiment Dictionary

[registration required]
This sentiment dictionary listing the polarities of many Chinese words was developed by a combination of automated and manual means [171]. A registration form for acquiring it is available at http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp.

OpinionFinder's Subjectivity Lexicon

URL: http://www.cs.pitt.edu/mpqa/
The list of subjectivity clues that is part of OpinionFinder is available for download. These clues were compiled from several sources, representing several years of effort, and were used in Wilson et al. [319].

SentiWordnet

URL: http://sentiwordnet.isti.cnr.it/
SentiWordnet [91] is a lexical resource for opinion mining. Each synset of WordNet [95], a publicly available thesaurus-like resource, is assigned one of three sentiment scores — positive, negative, or objective — where these scores were automatically generated using a semi-supervised method described in Esuli and Sebastiani [90].

Taboada and Grieve's Turney adjective list

[available through the Yahoo! sentimentAI group]
Reported are the semantic-orientation values according to the method proposed by Turney [298] for 1700 adjectives.

7.4 Tutorials, Bibliographies, and Other References

Bing Liu has a chapter on opinion mining in his book on Web data mining [190]. Slides for that chapter are available at http://www.cs.uic.edu/~liub/teach/cs583-spring-07/opinion-mining.pdf.

Slides for Janyce Wiebe's tutorial, "Semantics, opinion, and sentiment in text," at the EUROLAN 2007 Summmer School are available at http://www.cs.pitt.edu/~wiebe/pubs/papers/EUROLAN07/eurolan07wiebe.ppt.

The following are online bibliographies that contain information in BibTeX format:

- http: / / www.cs.cornell.edu / home / llee / opinion-mining-sentiment-analysis-survey.html, the main website for this survey,
- http: / / liinwww.ira.uka.de/bibliography/Misc/Sentiment. html, maintained by Andrea Esuli,
- http: / / research.microsoft.com / ~ jtsun / OpinionMining PaperList. html, maintained by Jian-Tao Sun,
- http: / / www.cs.pitt.edu / ~ wiebe / pubs / papers / EURO LAN07/eurolan07bib.html with actual .bib file at http: / /

www.cs.pitt.edu / ~ wiebe / pubs / papers / EUROLAN07/ eurolan07.bib, maintained by Janyce Wiebe.

Esuli and Wiebe's sites have additional search capabilities.

Members of the Yahoo! group "sentimentAI" (http://tech.groups.yahoo.com/group/SentimentAI/) have access to the resources that have been contributed there (such as some links to corpora and papers) and are subscribed to the associated mailing list. Joining is free.

8

Concluding Remarks

> When asked how he knew a piece was finished, he responded, "When the dinner bell rings."
>
> — apocryphal anecdote about Alexander Calder

Our goal in this survey has been to cover techniques and approaches that promise to directly enable opinion-oriented information-seeking systems, and to convey to the reader a sense of our excitement about the intellectual richness and breadth of the area. We very much encourage the reader to take up the many open challenges that remain, and hope we have provided some resources that will prove helpful in this regard.

On the topic of resources: we have already indicated above that the bibliographic database used in this survey is publicly available. In fact, the URL mentioned above, http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html, is our personally maintained homepage for this survey. Any subsequent editions or versions of this survey that may be produced, or related news, will be announced there.[1]

[1] Indeed, we have vague aspirations to producing a "director's cut" one day. We certainly have accumulated some number of outtakes: we did not manage to find a way to work

Speaking of resources, we have drawn considerably on those of many others during the course of this work. We thus have a number of sincere *acknowledgments* to make.

This survey is based upon work supported in part by the National Science Foundation under grant no. IIS-0329064, a Cornell University Provost's Award for Distinguished Scholarship, a Yahoo! Research Alliance gift, and an Alfred P. Sloan Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views or official policies, either expressed or implied, of any sponsoring institutions, the US government, or any other entity.

We would like to wholeheartedly thank the anonymous referees, who provided outstanding feedback astonishingly quickly. Their insights contributed immensely to the final form of this survey on many levels. It is hard to describe our level of gratitude to them for their time and their wisdom, except to say this: we have, in various capacities, seen many examples of reviewing in the community, but this is the best we have ever encountered. We also thank Eric Breck for his careful reading of and commentary on portions of this survey. All remaining errors and faults are, of course, our own.

We are also very thankful to Fabrizio Sebastiani, for all of his editorial guidance and care. We owe him a great debt. We also greatly appreciate the help we received from Jamie Callan, who, along with Fabrizio, serves as Editor in Chief of the Foundations and Trends in Information Retrieval series, and James Finlay, of Now Publishers, the publisher of this series.

Finally, a number of unexpected health problems arose in our families during the writing of this survey. Despite this, it was our families who sustained us with their cheerful and unlimited support (on many levels), not the other way around. Thus — to end on a sentimental note — this work is dedicated to them.

---

some variant of "Once more, with feeling" into the title, or to find a place for the heading "Sentiment of a woman," or to formally prove a potential undecidability result for subjectivity detection (Jon Kleinberg, personal communication) based on reviews of *Brotherhood of the Wolf* ("it's the best darned French werewolf kung-fu movie I've ever seen").

References

[1] A. Abbasi, "Affect intensity analysis of dark web forums," in *Proceedings of Intelligence and Security Informatics* (*ISI*), pp. 282–288, 2007.

[2] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: Divided they blog," in *Proceedings of LinkKDD*, 2005.

[3] A. Agarwal and P. Bhattacharyya, "Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified," in *Proceedings of the International Conference on Natural Language Processing (ICON)*, 2005.

[4] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," in *Proceedings of WWW*, pp. 529–535, 2003.

[5] E. M. Airoldi, X. Bai, and R. Padman, "Markov blankets and meta-heuristic search: Sentiment extraction from unstructured text," *Lecture Notes in Computer Science*, vol. 3932 (Advances in Web Mining and Web Usage Analysis), pp. 167–187, 2006.

[6] G. A. Akerlof, "The market for "Lemons": Quality uncertainty and the market mechanism," *The Quarterly Journal of Economics*, vol. 84, pp. 488–500, 1970.

[7] S. M. Al Masum, H. Prendinger, and M. Ishizuka, "SenseNet: A linguistic tool to visualize numerical-valence based sentiment of textual data," in *Proceedings of the International Conference on Natural Language Processing (ICON)*, pp. 147–152, 2007. (Poster paper).

[8] J. Allan, "Introduction to topic detection and tracking," in *Topic Detection and Tracking: Event-based Information Organization*, (J. Allan, ed.), pp. 1–16, Norwell, MA, USA: Kluwer Academic Publishers, ISBN 0-7923-7664-1, 2002.

[9] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: Machine learning for text-based emotion prediction," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[10] A. Anagnostopoulos, A. Z. Broder, and D. Carmel, "Sampling search-engine results," *World Wide Web*, vol. 9, pp. 397–429, 2006.

[11] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, pp. 1817–1853, 2005.

[12] A. Andreevskaia and S. Bergler, "Mining WordNet for a fuzzy sentiment: Sentiment tag extraction from WordNet glosses," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

[13] W. Antweiler and M. Z. Frank, "Is all that talk just noise? The information content of internet stock message boards," *Journal of Finance*, vol. 59, pp. 1259–1294, 2004.

[14] N. Archak, A. Ghose, and P. Ipeirotis, "Show me the money! Deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2007.

[15] S. Argamon, ed., *Proceedings of the IJCAI Workshop on DOING IT WITH STYLE: Computational Approaches to Style Analysis and Synthesis*. 2003.

[16] S. Argamon, J. Karlgren, and J. G. Shanahan, eds., *Proceedings of the SIGIR Workshop on Stylistic Analysis of Text For Information Access*. ACM, 2005.

[17] S. Argamon, J. Karlgren, and O. Uzuner, eds., *Proceedings of the SIGIR Workshop on Stylistics for Text Retrieval in Practice*. ACM, 2006.

[18] S. Argamon-Engelson, M. Koppel, and G. Avneri, "Style-based text categorization: What newspaper am I reading?" in *Proceedings of the AAAI Workshop on Text Categorization*, pp. 1–4, 1998.

[19] Y. Attali and J. Burstein, "Automated essay scoring with e-rater v.2," *Journal of Technology, Learning, and Assessment*, vol. 26, February 2006.

[20] A. Aue and M. Gamon, "Automatic identification of sentiment vocabulary: Exploiting low association with known sentiment terms," in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in Natural Language Processing*, 2005.

[21] A. Aue and M. Gamon, "Customizing sentiment classifiers to new domains: A case study," in *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2005.

[22] B. Awerbuch and R. Kleinberg, "Competitive collaborative learning," in *Proceedings of the Conference on Learning Theory (COLT)*, pp. 233–248, 2005. (Journal version to appear in *Journal of Computer and System Sciences*, special issue on computational learning theory).

[23] P. Bajari and A. Hortaçsu, "The winner's curse, reserve prices, and endogenous entry: Empirical insights from eBay auctions," *RAND Journal of Economics*, vol. 34, pp. 329–355, 2003.

[24] P. Bajari and A. Hortaçsu, "Economic insights from internet auctions," *Journal of Economic Literature*, vol. 42, pp. 457–486, 2004.

[25] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The Berkeley Framenet Project," in *Proceedings of COLING/ACL*, 1998.

[26] A. Banfield, *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge and Kegan Paul, 1982.

[27] M. Bansal, C. Cardie, and L. Lee, "The power of negative thinking: Exploiting label disagreement in the min-cut classification framework," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2008. (Poster paper).

[28] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor, "The second PASCAL recognising textual entailment challenge," in *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

[29] R. Barzilay and L. Lee, "Learning to paraphrase: An unsupervised approach using multiple-sequence alignment," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pp. 16–23, 2003.

[30] R. Barzilay and K. McKeown, "Extracting paraphrases from a parallel corpus," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 50–57, 2001.

[31] S. Basuroy, S. Chatterjee, and S. A. Ravid, "How critical are critical reviews? The box office effects of film critics, star power and budgets," *Journal of Marketing*, vol. 67, pp. 103–117, 2003.

[32] M. Bautin, L. Vijayarenu, and S. Skiena, "International sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[33] P. Beineke, T. Hastie, C. Manning, and S. Vaithyanathan, "Exploring sentiment summarization," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, AAAI technical report SS-04-07, 2004.

[34] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. (Short paper).

[35] J. Berger, A. T. Sorensen, and S. J. Rasmussen, "Negative publicity: When is negative a positive?," Manuscript. PDF file's last modification date: October 16, 2007, *URL*: http://www.stanford.edu/~asorense/papers/Negative_Publicity.pdf, 2007.

[36] Y. Bestgen, C. Fairon, and L. Kerves, "Un baromètre affectif effectif: Corpus de référence et méthode pour déterminer la valence affective de phrases," in *Journées internationales d'analyse statistique des donnés textuelles (JADT)*, pp. 182–191, 2004.

[37] S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, and D. Jurafsky, "Automatic extraction of opinion propositions and their holders," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

[38] D. Biber, *Variation Across Speech and Writing*. Cambridge University Press, 1988.

[39] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[40] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

[41] S. R. K. Branavan, H. Chen, J. Eisenstein, and R. Barzilay, "Learning document-level semantic properties from free-text annotations," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2008.

[42] E. Breck and C. Cardie, "Playing the telephone game: Determining the hierarchical structure of perspective and speech expressions," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

[43] E. Breck, Y. Choi, and C. Cardie, "Identifying expressions of opinion in context," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Hyderabad, India, 2007.

[44] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th International World Wide Web Conference*, pp. 107–117, 1998.

[45] R. F. Bruce and J. M. Wiebe, "Recognizing subjectivity: A case study in manual tagging," *Natural Language Engineering*, vol. 5, 1999.

[46] J. K. Burgoon, J. P. Blair, T. Qin, and J. F. Nunamaker, Jr., "Detecting deception through linguistic analysis," in *Proceedings of Intelligence and Security Informatics (ISI),* number 2665 in *Lecture Notes in Computer Science*, p. 958, 2008.

[47] L. Cabral and A. Hortaçsu, "The dynamics of seller reputation: Theory and evidence from eBay," Working Paper, downloaded version revised in March, 2006, URL http://pages.stern.nyu.edu/~lcabral/workingpapers/Cabral Hortacsu_Mar06.pdf, 2006.

[48] J. Carbonell, *Subjective Understanding: Computer Models of Belief Systems.* PhD thesis, Yale, 1979.

[49] C. Cardie, "Empirical methods in information extraction," *AI Magazine*, vol. 18, pp. 65–79, 1997.

[50] C. Cardie, C. Farina, T. Bruce, and E. Wagner, "Using natural language processing to improve eRulemaking," in *Proceedings of Digital Government Research (dg.o)*, 2006.

[51] C. Cardie, J. Wiebe, T. Wilson, and D. Litman, "Combining low-level and summary representations of opinions for multi-perspective question answering," in *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, pp. 20–27, 2003.

[52] G. Carenini, R. Ng, and A. Pauls, "Multi-document summarization of evaluative text," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 305–312, 2006.

[53] G. Carenini, R. T. Ng, and A. Pauls, "Interactive multimedia summaries of evaluative text," in *Proceedings of Intelligent User Interfaces (IUI)*, pp. 124–131, ACM Press, 2006.

[54] D. Cartwright and F. Harary, "Structural balance: A generalization of Heider's theory," *Psychological Review*, vol. 63, pp. 277–293, 1956.

[55] P. Chaovalit and L. Zhou, "Movie review mining: A comparison between supervised and unsupervised classification approaches," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[56] P.-Y. S. Chen, S.-Y. Wu, and J. Yoon, "The impact of online recommendations and consumer feedback on sales," in *International Conference on Information Systems (ICIS)*, pp. 711–724, 2004.

[57] Y. Chen and J. Xie, "Online consumer review: Word-of-mouth as a new element of marketing communication mix," *Management Science*, vol. 54, pp. 477–491, 2008.

[58] P. Chesley, B. Vincent, L. Xu, and R. Srihari, "Using verbs and adjectives to automatically classify blog sentiment," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 27–29, 2006.

[59] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, vol. 43, pp. 345–354, August 2006.

[60] Y. Choi, E. Breck, and C. Cardie, "Joint extraction of entities and relations for opinion recognition," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[61] Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[62] E. K. Clemons, G. Gao, and L. M. Hitt, "When online reviews meet hyperdifferentiation: A study of the craft beer industry," *Journal of Management Information Systems*, vol. 23, pp. 149–171, 2006.

[63] comScore/the Kelsey group, "Online consumer-generated reviews have significant impact on offline purchase behavior," Press Release, http://www.comscore.com/press/release.asp?press=1928, November 2007.

[64] J. G. Conrad and F. Schilder, "Opinion mining in legal blogs," in *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*, pp. 231–236, New York, NY, USA: ACM, 2007.

[65] W. B. Croft and J. Lafferty, eds., *Language modeling for information retrieval*. Number 13 in the Information Retrieval Series. Kluwer/Springer, 2003.

[66] S. Das and M. Chen, "Yahoo! for Amazon: Extracting market sentiment from stock message boards," in *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)*, 2001.

[67] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," *Management Science*, vol. 53, pp. 1375–1388, 2007.

[68] S. R. Das, P. Tufano, and F. de Asis Martinez-Jerez, "eInformation: A clinical study of investor discussion and sentiment," *Financial Management*, vol. 34, pp. 103–137, 2005.

[69] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of WWW*, pp. 519–528, 2003.

[70] S. David and T. J. Pinch, "Six degrees of reputation: The use and abuse of online review and recommendation systems," *First Monday*, July 2006. (Special Issue on Commercial Applications of the Internet).

[71] C. Dellarocas, "The digitization of word-of-mouth: Promise and challenges of online reputation systems," *Management Science*, vol. 49, pp. 1407–1424, 2003. (Special issue on e-business and management science).

[72] C. Dellarocas, X. Zhang, and N. F. Awad, "Exploring the value of online product ratings in revenue forecasting: The case of motion pictures," *Journal of Interactive Marketing*, vol. 21, pp. 23–45, 2007.

[73] A. Devitt and K. Ahmad, "Sentiment analysis in financial news: A cohesion-based approach," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 984–991, 2007.

[74] M. Dewally, "Internet investment advice: Investing with a rock of salt," *Financial Analysts Journal*, vol. 59, pp. 65–77, July/August 2003.

[75] M. Dewally and L. Ederington, "Reputation, certification, warranties, and information as remedies for seller-buyer information asymmetries: Lessons from the online comic book market," *Journal of Business*, vol. 79, pp. 693–730, March 2006.

[76] S. Dewan and V. Hsu, "Adverse selection in electronic markets: Evidence from online stamp auctions," *Journal of Industrial Economics*, vol. 52, pp. 497–516, December 2004.

[77] D. W. Diamond, "Reputation acquisition in debt markets," *Journal of Political Economy*, vol. 97, pp. 828–862, 1989.

[78] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the Conference on Web Search and Web Data Mining (WSDM)*, 2008.

[79] L. Dini and G. Mazzini, "Opinion classification through information extraction," in *Proceedings of the Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields (Data Mining)*, pp. 299–310, 2002.

[80] W. Duan, B. Gu, and A. B. Whinston, "Do online reviews matter? — An empirical investigation of panel data," Social Science Research Network (SSRN) Working Paper Series, http://ssrn.com/paper=616262, version as of January, 2005.

[81] D. H. Eaton, "Valuing information: Evidence from guitar auctions on eBay," *Journal of Applied Economics and Policy*, vol. 24, pp. 1–19, 2005.

[82] D. H. Eaton, "The impact of reputation timing and source on auction outcomes," *The B. E. Journal of Economic Analysis and Policy*, vol. 7, 2007.

[83] M. Efron, "Cultural orientation: Classifying subjective documents by cociation [sic] analysis," in *Proceedings of the AAAI Fall Symposium on Style and Meaning in Language, Art, Music, and Design*, pp. 41–48, 2004.

[84] K. Eguchi and V. Lavrenko, "Sentiment retrieval using generative models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 345–354, 2006.

[85] K. Eguchi and C. Shah, "Opinion retrieval experiments using generative models: Experiments for the TREC 2006 blog track," in *Proceedings of TREC*, 2006.

[86] P. Ekman, *Emotion in the Human Face*. Cambridge University Press, Second ed., 1982.

[87] J. Eliashberg and S. M. Shugan, "Film critics: Influencers or predictors?," *Journal of Marketing*, vol. 61, pp. 68–78, April 1997.

[88] C. Engström, *Topic Dependence in Sentiment Classification*. Master's thesis, University of Cambridge, 2004.

[89] A. Esuli and F. Sebastiani, "Determining the semantic orientation of terms through gloss analysis," in *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, 2005.

[90] A. Esuli and F. Sebastiani, "Determining term subjectivity and term orientation for opinion mining," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

[91] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proceedings of Language Resources and Evaluation (LREC)*, 2006.

[92] A. Esuli and F. Sebastiani, "PageRanking WordNet synsets: An application to opinion mining," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

[93] D. K. Evans, L.-W. Ku, Y. Seki, H.-H. Chen, and N. Kando, "Opinion analysis across languages: An overview of and observations from the NTCIR6 opinion analysis pilot task," in *Proceedings of the Workshop on Cross-Language Information Processing,* vol. 4578 (Applications of Fuzzy Sets Theory) *of Lecture Notes in Computer Science*, pp. 456–463, 2007.

[94] A. Fader, D. R. Radev, M. H. Crespin, B. L. Monroe, K. M. Quinn, and M. Colaresi, "MavenRank: Identifying influential members of the US senate using lexical centrality," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

[95] C. Fellbaum, ed., *Wordnet: An Electronic Lexical Database*. MIT Press, 1998.

[96] D. Feng, E. Shaw, J. Kim, and E. Hovy, "Learning to detect conversation focus of threaded discussions," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pp. 208–215, 2006.

[97] A. Finn and N. Kushmerick, "Learning to classify documents according to genre," *Journal of the American Society for Information Science and Technology (JASIST)*, vol. 7, 2006. (Special issue on computational analysis of style).

[98] A. Finn, N. Kushmerick, and B. Smyth, "Genre classification and domain transfer for information filtering," in *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval,* number 2291 in *Lecture Notes in Computer Science*, pp. 353–362, Glasgow, 2002.

[99] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated essay scoring: Applications to education technology," in *Proceedings of ED-MEDIA*, pp. 939–944, 1999.

[100] C. Forman, A. Ghose, and B. Wiesenfeld, "Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets," *Information Systems Research*, vol. 19, 2008. (Special issue on the interplay between digital and social networks).

[101] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[102] T. Fukuhara, H. Nakagawa, and T. Nishida, "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[103] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

[104] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger, "Pulse: Mining customer opinions from free text," in *Proceedings of the International Symposium on Intelligent Data Analysis (IDA),* number 3646 in *Lecture Notes in Computer Science*, pp. 121–132, 2005.

[105] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *SIGKDD Explorations Newsletter*, vol. 8, pp. 41–48, 2006.

[106] A. Ghose and P. G. Ipeirotis, "Designing novel review ranking systems: Predicting usefulness and impact of reviews," in *Proceedings of the International Conference on Electronic Commerce (ICEC)*, 2007. (Invited paper).

[107] A. Ghose, P. G. Ipeirotis, and A. Sundararajan, "Opinion mining using econometrics: A case study on reputation systems," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2007.

[108] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[109] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*, 2006.

[110] A. B. Goldberg, X. Zhu, and S. Wright, "Dissimilarity in graph-based semi-supervised classification," in *Artificial Intelligence and Statistics (AISTATS)*, 2007.

[111] S. Greene, *Spin: Lexical Semantics, Transitivity, and the Identification of Implicit Sentiment*. PhD thesis, University of Maryland, 2007.

[112] G. Grefenstette, Y. Qu, J. G. Shanahan, and D. A. Evans, "Coupling niche browsers and affect analysis for an opinion mining application," in *Proceedings of Recherche d'Information Assistée par Ordinateur (RIAO)*, 2004.

[113] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner, "User-directed sentiment analysis: Visualizing the affective content of documents," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pp. 23–30, Sydney, Australia, July 2006.

[114] B. Gu, P. Konana, A. Liu, B. Rajagopalan, and J. Ghosh, "Predictive value of stock message board sentiments," McCombs Research Paper No. IROM-11-06, version dated November, 2006.

[115] R. V. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proceedings of WWW*, pp. 403–412, 2004.

[116] B. A. Hagedorn, M. Ciaramita, and J. Atserias, "World knowledge in broad-coverage information filtering," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, 2007. (Poster paper).

[117] J. T. Hancock, L. Curry, S. Goorha, and M. Woodworth, "Automated linguistic analysis of deceptive and truthful synchronous computer-mediated communication," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, p. 22c, 2005.

[118] L. Hankin, "The effects of user reviews on online purchasing behavior across multiple product categories," Master's final project report, UC Berkeley School of Information, http://www.ischool.berkeley.edu/files/lhankin_report.pdf, May 2007.

[119] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in *Proceedings of the Joint ACL/EACL Conference*, pp. 174–181, 1997.

[120] V. Hatzivassiloglou and J. Wiebe, "Effects of adjective orientation and gradability on sentence subjectivity," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2000.

[121] M. Hearst, "Direction-based text interpretation as an information access refinement," in *Text-Based Intelligent Systems*, (P. Jacobs, ed.), pp. 257–274, Lawrence Erlbaum Associates, 1992.

[122] R. Higashinaka, M. Walker, and R. Prasad, "Learning to generate naturalistic utterances using reviews in spoken dialogue systems," *ACM Transactions on Speech and Language Processing (TSLP)*, 2007.

[123] P. Hitlin and L. Rainie, "The use of online reputation and rating systems," Pew Internet & American Life Project Memo, October 2004.

[124] T. Hoffman, "Online reputation management is hot — but is it ethical?" Computerworld, February 2008.

[125] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of SIGIR*, pp. 50–57, 1999.

[126] D. Hopkins and G. King, "Extracting systematic social science meaning from text,". Manuscript available at http://gking.harvard.edu/files/words.pdf, 2007 version was the one most recently consulted, 2007.

[127] J. A. Horrigan, "Online shopping," Pew Internet & American Life Project Report, 2008.

[128] D. Houser and J. Wooders, "Reputation in auctions: Theory, and evidence from eBay," *Journal of Economics and Management Strategy*, vol. 15, pp. 252–369, 2006.

[129] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 168–177, 2004.

[130] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of AAAI*, pp. 755–760, 2004.

[131] M. Hu, A. Sun, and E.-P. Lim, "Comments-oriented blog summarization by sentence extraction," in *Proceedings of the ACM SIGIR Conference on*

Information and Knowledge Management (CIKM), pp. 901–904, 2007. (Poster paper).

[132] N. Hu, P. A. Pavlou, and J. Zhang, "Can online reviews reveal a product's true quality?: Empirical findings and analytical modeling of online word-of-mouth communication," in *Proceedings of Electronic Commerce (EC)*, pp. 324–330, USA, New York, NY: ACM, 2006.

[133] A. Huettner and P. Subasic, "Fuzzy typing for document management," in *ACL 2000 Companion Volume: Tutorial Abstracts and Demonstration Notes*, pp. 26–27, 2000.

[134] M. Hurst and K. Nigam, "Retrieving topical sentiments from online document collections," in *Document Recognition and Retrieval XI*, pp. 27–34, 2004.

[135] C. Jacquemin, *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, 2001.

[136] G. Jin and A. Kato, "Price, quality and reputation: Evidence from an online field experiment," *The RAND Journal of Economics*, vol. 37, 2006.

[137] X. Jin, Y. Li, T. Mah, and J. Tong, "Sensitive webpage classification for content advertising," in *Proceedings of the International Workshop on Data Mining and Audience Intelligence for Advertising*, 2007.

[138] N. Jindal and B. Liu, "Identifying comparative sentences in text documents," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, 2006.

[139] N. Jindal and B. Liu, "Mining comparative sentences and relations," in *Proceedings of AAAI*, 2006.

[140] N. Jindal and B. Liu, "Review spam detection," in *Proceedings of WWW*, 2007. (Poster paper).

[141] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the Conference on Web Search and Web Data Mining (WSDM)*, pp. 219–230, 2008.

[142] N. Kaji and M. Kitsuregawa, "Automatic construction of polarity-tagged corpus from HTML documents," in *Proceedings of the COLING/ACL Main Conference Poster Sessions*, 2006.

[143] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of HTML documents," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1075–1083, 2007.

[144] A. Kale, A. Karandikar, P. Kolari, A. Java, T. Finin, and A. Joshi, "Modeling trust and influence in the blogosphere using link polarity," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007. (Short paper).

[145] K. Kalyanam and S. H. McIntyre, "The role of reputation in online auction markets," Santa Clara University Working Paper 02/03-10-WP, 2001, dated June 26.

[146] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke, "Using WordNet to measure semantic orientation of adjectives," in *Proceedings of LREC*, 2004.

[147] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust algorithm for reputation management in P2P networks," in *Proceedings of WWW*, pp. 640–651, New York, NY, USA: ACM, ISBN 1-58113-680-3, 2003.

[148] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Sydney, Australia), pp. 355–363, July 2006.

[149] M. Kantrowitz, "Method and apparatus for analyzing affect and emotion in text," U.S. Patent 6622140, Patent filed in November 2000, 2003.

[150] J. Karlgren and D. Cutting, "Recognizing text genres with simple metrics using discriminant analysis," in *Proceedings of COLING*, pp. 1071–1075, 1994.

[151] Y. Kawai, T. Kumamoto, and K. Tanaka, "Fair news reader: Recommending news articles with different sentiments based on user preference," in *Proceedings of Knowledge-Based Intelligent Information and Engineering Systems (KES),* number 4692 in *Lecture Notes in Computer Science*, pp. 612–622, 2007.

[152] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, pp. 110–125, 2006.

[153] B. Kessler, G. Nunberg, and H. Schütze, "Automatic detection of text genre," in *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pp. 32–38, 1997.

[154] P. Kim, "The forrester wave: Brand monitoring, Q3 2006," Forrester Wave (white paper), 2006.

[155] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

[156] S.-M. Kim and E. Hovy, "Automatic detection of opinion bearing words and sentences," in *Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2005.

[157] S.-M. Kim and E. Hovy, "Identifying opinion holders for question answering in opinion texts," in *Proceedings of the AAAI Workshop on Question Answering in Restricted Domains*, 2005.

[158] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pp. 483–490, 2006.

[159] S.-M. Kim and E. Hovy, "Identifying and analyzing judgment opinions," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, 2006.

[160] S.-M. Kim and E. Hovy, "Crystal: Analyzing predictive opinions on the web," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[161] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pp. 423–430, Sydney, Australia, July 2006.

[162] B. Klein and K. Leffler, "The role of market forces in assuring contractual performance," *Journal of Political Economy*, vol. 89, pp. 615–641, 1981.

[163] J. Kleinberg, "Authoritative sources in a hyperlinked environment," in *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 668–677, 1998. (Extended version in *Journal of the ACM*, 46:604–632, 1999).

[164] J. Kleinberg and E. Tardos, "Approximation algorithms for classification problems with pairwise relationships: Metric labeling and Markov random fields," *Journal of the ACM*, vol. 49, pp. 616–639, ISSN 0004-5411, 2002.

[165] J. Kleinberg and E. Tardos, *Algorithm Design*. Addison Wesley, 2006.

[166] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima, "Collecting evaluative expressions for opinion extraction," in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2004.

[167] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," in *Workshop on the Analysis of Informal and Formal Information Exchange During Negotiations (FINEXIN)*, 2005.

[168] M. Koppel and I. Shtrimberg, "Good news or bad news? Let the market decide," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pp. 86–88, 2004.

[169] L.-W. Ku, L.-Y. Li, T.-H. Wu, and H.-H. Chen, "Major topic detection and its application to opinion summarization," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pp. 627–628, 2005. (Poster paper).

[170] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 100–107, 2006.

[171] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Tagging heterogeneous evaluation corpora for opinionated tasks," in *Conference on Language Resources and Evaluation (LREC)*, 2006.

[172] L.-W. Ku, Y.-S. Lo, and H.-H. Chen, "Test collection selection and gold standard generation for a multiply-annotated opinion corpus," in *Proceedings of the ACL Demo and Poster Sessions*, pp. 89–92, 2007.

[173] T. Kudo and Y. Matsumoto, "A boosting algorithm for classification of semi-structured text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[174] S. Kurohashi, K. Inui, and Y. Kato, eds., *Workshop on Information Credibility on the Web*, 2007.

[175] N. Kwon, S. Shulman, and E. Hovy, "Multidimensional text analysis for eRulemaking," in *Proceedings of Digital Government Research (dg.o)*, 2006.

[176] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of ICML*, pp. 282–289, 2001.

[177] J. D. Lafferty and C. Zhai, "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of SIGIR*, pp. 111–119, 2001.

[178] M. Laver, K. Benoit, and J. Garry, "Extracting policy positions from political texts using words as data," *American Political Science Review*, vol. 97, pp. 311–331, 2003.

[179] V. Lavrenko and W. Bruce Croft, "Relevance-based language models," in *Proceedings of SIGIR*, pp. 120–127, 2001.

[180] C. G. Lawson and V. C. Slawson, "Reputation in an internet auction market," *Economic Inquiry*, vol. 40, pp. 533–650, 2002.

[181] L. Lee, ""I'm sorry Dave, I'm afraid I can't do that": Linguistics, statistics, and natural language processing circa 2001," in *Computer Science: Reflections on the Field, Reflections from the Field*, (Committee on the Fundamentals of Computer Science: Challenges and Opportunities, Computer Science and Telecommunications Board, National Research Council, ed.), pp. 111–118, The National Academies Press, 2004.

[182] Y.-B. Lee and S. H. Myaeng, "Text genre classification with genre-revealing and subject-revealing features," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, 2002.

[183] D. Leinweber and A. Madhavan, "Three hundred years of stock market manipulation," *Journal of Investing*, vol. 10, pp. 7–16, Summer 2001.

[184] H. Li and K. Yamanishi, "Mining from open answers in questionnaire data," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 443–449, 2001. (Journal version in *IEEE Intelligent Systems* vol. 17, no. 5, pp. 58–63, 2002).

[185] Y. Li, Z. Zheng, and H. Dai, "KDD CUP-2005 report: Facing a great challenge," *SIGKDD Explorations*, vol. 7, pp. 91–99, 2005.

[186] W.-H. Lin and A. Hauptmann, "Are these documents written from different perspectives? A test of different perspectives based on statistical distribution divergence," in *Proceedings of the International Conference on Computational Linguistics (COLING)/Proceedings of the Association for Computational Linguistics (ACL)*, pp. 1057–1064, Sydney, Australia: Association for Computational Linguistics, July 2006.

[187] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which side are you on? Identifying perspectives at the document and sentence levels," in *Proceedings of the Conference on Natural Language Learning (CoNLL)*, 2006.

[188] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," in *Interspeech*, pp. 1845–1848, 2005.

[189] L. V. Lita, A. H. Schlaikjer, W. Hong, and E. Nyberg, "Qualitative dimensions in question answering: Extending the definitional QA task," in *Proceedings of AAAI*, pp. 1616–1617, 2005. (Student abstract).

[190] B. Liu, "Web data mining; Exploring hyperlinks, contents, and usage data," *Opinion Mining*. Springer, 2006.

[191] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proceedings of WWW*, 2005.

[192] H. Liu, H. Lieberman, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proceedings of Intelligent User Interfaces (IUI)*, pp. 125–132, 2003.

[193] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 334–342, 2007. (Poster paper).

[194] Y. Liu, "Word-of-mouth for movies: Its dynamics and impact on box office revenue," *Journal of Marketing*, vol. 70, pp. 74–89, 2006.

[195] Y. Liu, J. Huang, A. An, and X. Yu, "ARSA: A sentiment-aware model for predicting sales performance using blogs," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, 2007.

[196] J. A. Livingston, "How valuable is a good reputation? A sample selection model of internet auctions," *The Review of Economics and Statistics*, vol. 87, pp. 453–465, August 2005.

[197] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *Proceedings of String Processing and Information Retrieval (SPIRE),* number 3772 in *Lecture Notes in Computer Science*, pp. 161–166, 2005.

[198] D. Lucking-Reiley, D. Bryan, N. Prasad, and D. Reeves, "Pennies from eBay: The determinants of price in online auctions," *Journal of Industrial Economics*, vol. 55, pp. 223–233, 2007.

[199] C. Macdonald and I. Ounis, "The TREC Blogs06 collection: Creating and analysing a blog test collection," Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.

[200] Y. Mao and G. Lebanon, "Sequential models for sentiment prediction," in *ICML Workshop on Learning in Structured Output Spaces*, 2006.

[201] Y. Mao and G. Lebanon, "Isotonic conditional random fields and local sentiment flow," in *Advances in Neural Information Processing Systems*, 2007.

[202] L. W. Martin and G. Vanberg, "A robust transformation procedure for interpreting political text," *Political Analysis*, vol. 16, pp. 93–100, 2008.

[203] H. Masum and Y.-C. Zhang, "Manifesto for the reputation society," *First Monday*, vol. 9, 2004.

[204] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," in *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2005.

[205] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar, "Structured models for fine-to-coarse sentiment analysis," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 432–439, Prague, Czech Republic: Association for Computational Linguistics, June 2007.

[206] Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proceedings of WWW*, pp. 171–180, New York, NY, USA: ACM Press, 2007. (ISBN 978-1-59593-654-7).

[207] M. I. Melnik and J. Alm, "Does a seller's eCommerce reputation matter? Evidence from eBay auctions," *Journal of Industrial Economics*, vol. 50, pp. 337–349, 2002.

[208] M. I. Melnik and J. Alm, "Seller reputation, information signals, and prices for heterogeneous coins on eBay," *Southern Economic Journal*, vol. 72, pp. 305–328, 2005.

[209] R. Mihalcea, C. Banea, and J. Wiebe, "Learning multilingual subjective language via cross-lingual projections," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 976–983, Prague, Czech Republic, June 2007.

[210] R. Mihalcea and C. Strapparava, "Learning to laugh (automatically): Computational models for humor recognition," *Journal of Computational Intelligence*, 2006.

[211] G. Mishne and M. de Rijke, "Capturing global mood levels using blog posts," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 145–152, 2006.

[212] G. Mishne and M. de Rijke, "Moodviews: Tools for blog mood analysis," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 153–154, 2006.

[213] G. Mishne and M. de Rijke, "A study of blog search," in *Proceedings of the European Conference on Information Retrieval Research (ECIR)*, 2006.

[214] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 155–158, 2006.

[215] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the Web," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 341–349, 2002. (Industry track).

[216] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, 1984.

[217] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 412–418, July 2004. (Poster paper).

[218] T. Mullen and R. Malouf, "Taking sides: User classification for informal online political discourse," *Internet Research*, vol. 18, pp. 177–190, 2008.

[219] T. Mullen and R. Malouf, "A preliminary investigation into sentiment analysis of informal political discourse," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 159–162, 2006.

[220] J.-C. Na, H. Sui, C. Khoo, S. Chan, and Y. Zhou, "Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews," in *Conference of the International Society for Knowledge Organization (ISKO)*, pp. 49–54, 2004.

[221] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the Conference on Knowledge Capture (K-CAP)*, 2003.

[222] V. Ng, S. Dasgupta, and S. M. N. Arifin, "Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews," in *Proceedings of the COLING/ACL Main Conference Poster Sessions*, pp. 611–618, Sydney, Australia: Association for Computational Linguistics, July 2006.

[223] X. Ni, G.-R. Xue, X. Ling, Y. Yu, and Q. Yang, "Exploring in the weblog space by detecting informative and affective articles," in *Proceedings of WWW*, 2007. (Industrial practice and experience track).

[224] N. Nicolov, F. Salvetti, M. Liberman, and J. H. Martin, eds., *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*. AAAI Press, 2006.

[225] K. Nigam and M. Hurst, "Towards a robust metric of polarity," in *Computing Attitude and Affect in Text: Theories and Applications,* number 20 in *The Information Retrieval Series*, (J. G. Shanahan, Y. Qu, and J. Wiebe, eds.), 2006.

[226] Y. Niu, X. Zhu, J. Li, and G. Hirst, "Analysis of polarity information in medical text," in *Proceedings of the American Medical Informatics Association 2005 Annual Symposium*, 2005.

[227] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff, "Overview of the TREC-2006 blog track," in *Proceedings of the 15th Text Retrieval Conference (TREC)*, 2006.

[228] I. Ounis, C. Macdonald, and I. Soboroff, "On the TREC blog track," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008.

[229] S. Owsley, S. Sood, and K. J. Hammond, "Domain specific affective classification of documents," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 181–183, 2006.

[230] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: A corpus annotated with semantic roles," *Computational Linguistics*, vol. 31, March 2005.

[231] B. Pang, K. Knight, and D. Marcu, "Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences," in *Proceedings of HLT/NAACL*, 2003.

[232] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 271–278, 2004.

[233] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 115–124, 2005.

[234] B. Pang and L. Lee, "Using very simple statistics for review search: An exploration," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2008. (Poster paper).

[235] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86, 2002.

[236] D.-H. Park, J. Lee, and I. Han, "The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement," *International Journal of Electronic Commerce*, vol. 11, pp. 125–148, (ISSN 1086-4415), 2007.

[237] P. A. Pavlou and A. Dimoka, "The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation," *Information Systems Research*, vol. 17, pp. 392–414, 2006.

[238] S. Piao, S. Ananiadou, Y. Tsuruoka, Y. Sasaki, and J. McNaught, "Mining opinion polarity relations of citations," in *International Workshop on Computational Semantics (IWCS)*, pp. 366–371, 2007. (Short paper).

[239] R. Picard, *Affective Computing*. MIT Press, 1997.

[240] T. Pinch and K. Athanasiades, "ACIDplanet: A study of users of an on-line music community," 2005. http://sts.nthu.edu.tw/sts_camp/files/ACIDplanet%20by%20Trevor%20Pinch.ppt, Presented at the 50th Society for Ethnomusicology (SEM) conference.

[241] G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics," *Information Processing and Management*, vol. 12, pp. 297–312, 1976.

[242] L. Polanyi and A. Zaenen, "Contextual lexical valence shifters," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, AAAI technical report SS-04-07, 2004.

[243] J. M. Ponte and W. Bruce Croft, "A language modeling approach to information retrieval," in *Proceedings of SIGIR*, pp. 275–281, 1998.

[244] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 2005.

[245] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik, *A comprehensive grammar of the English language*. Longman, 1985.

[246] D. Radev, T. Allison, S. Blair-Goldensohn, J. Blitzer, A. Çelebi, S. Dimitrov, E. Drabek, A. Hakim, W. Lam, D. Liu, J. Otterbacher, H. Qi, H. Saggion, S. Teufel, M. Topper, A. Winkel, and Z. Zhang, "MEAD — A platform for multidocument multilingual text summarization," in *Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 2004.

[247] D. R. Radev, E. Hovy, and K. McKeown, "Introduction to the special issue on summarization," *Computational Linguistics*, vol. 28, pp. 399–408, (ISSN 0891-2017), 2002.

[248] L. Rainie and J. Horrigan, "Election 2006 online," Pew Internet & American Life Project Report, January 2007.

[249] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL Student Research Workshop*, 2005.

[250] D. A. Reinstein and C. M. Snyder, "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics," *Journal of Industrial Economics*, vol. 53, pp. 27–51, 2005.

[251] E. Reiter and R. Dale, *Building Natural Language Generation Systems*. Cambridge, 2000.

[252] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the Association for Computing Machinery (CACM)*, vol. 43, pp. 45–48, (ISSN 0001-0782), 2000.

[253] P. Resnick, R. Zeckhauser, J. Swanson, and K. Lockwood, "The value of reputation on eBay: A controlled experiment," *Experimental Economics*, vol. 9, pp. 79–101, 2006.

[254] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature subsumption for opinion analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2006.

[255] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003.

[256] E. Riloff, J. Wiebe, and W. Phillips, "Exploiting subjectivity classification to improve information extraction," in *Proceedings of AAAI*, pp. 1106–1111, 2005.

[257] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping," in *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 25–32, 2003.

[258] E. Rogers, *Diffusion of Innovations*. Free Press, New York, 1962. (ISBN 0743222091. Fifth edition dated 2003).

[259] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," *The Journal of Political Economy*, vol. 82, pp. 34–55, Jan–Feb 1974.

[260] D. Roth and W. Yih, "Probabilistic reasoning for entity and relation recognition," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2004.

[261] V. L. Rubin and E. D. Liddy, "Assessing credibility of weblogs," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 187–190, 2006.

[262] W. Sack, "On the computation of point of view," in *Proceedings of AAAI*, p. 1488, 1994. (Student abstract).

[263] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.

[264] Y. Seki, K. Eguchi, and N. Kando, "Analysis of multi-document viewpoint summarization using multi-dimensional genres," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pp. 142–145, 2004.

[265] Y. Seki, K. Eguchi, N. Kando, and M. Aono, "Multi-document summarization with subjectivity analysis at DUC 2005," in *Proceedings of the Document Understanding Conference (DUC)*, 2005.

[266] Y. Seki, K. Eguchi, N. Kando, and M. Aono, "Opinion-focused summarization and its analysis at DUC 2006," in *Proceedings of the Document Understanding Conference (DUC)*, pp. 122–130, 2006.

[267] Y. Seki, D. Kirk Evans, L.-W. Ku, H.-H. Chen, N. Kando, and C.-Y. Lin, "Overview of opinion analysis pilot task at NTCIR-6," in *Proceedings of the*

Workshop Meeting of the National Institute of Informatics (NII) Test Collection for Information Retrieval Systems (NTCIR), pp. 265–278, 2007.

[268] C. Shapiro, "Consumer information, product quality, and seller reputation," *Bell Journal of Economics*, vol. 13, pp. 20–35, 1982.

[269] C. Shapiro, "Premiums for high quality products as returns to reputations," *Quarterly Journal of Economics*, vol. 98, pp. 659–680, 1983.

[270] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Transactions on Graphics*, vol. 11, pp. 92–99, 1992.

[271] S. Shulman, J. Callan, E. Hovy, and S. Zavestoski, "Language processing technologies for electronic rulemaking: A project highlight," in *Proceedings of Digital Government Research (dg.o)*, pp. 87–88, 2005.

[272] B. Snyder and R. Barzilay, "Multiple aspect ranking using the Good Grief algorithm," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, pp. 300–307, 2007.

[273] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, "Detecting arguing and sentiment in meetings," in *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2007.

[274] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov, "QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[275] X. Song, Y. Chi, K. Hino, and B. Tseng, "Identifying opinion leaders in the blogosphere," in *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, pp. 971–974, 2007.

[276] E. Spertus, "Smokey: Automatic recognition of hostile messages," in *Proceedings of Innovative Applications of Artificial Intelligence (IAAI)*, pp. 1058–1065, 1997.

[277] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2000.

[278] S. S. Standifird, "Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings," *Journal of Management*, vol. 27, pp. 279–295, 2001.

[279] A. Stepinski and V. Mittal, "A fact/opinion classifier for news articles," in *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, pp. 807–808, New York, NY, USA: ACM Press, 2007. (ISBN 978-1-59593-597-7).

[280] B. Stone and M. Richtel, "The hand that controls the sock puppet could get slapped," *The New York Times*, July 16 2007.

[281] P. J. Stone, *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press, 1966.

[282] V. Stoyanov and C. Cardie, "Partially supervised coreference resolution for opinion summarization through structured rule learning," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 336–344, Sydney, Australia: Association for Computational Linguistics, July 2006.

[283] V. Stoyanov, C. Cardie, D. Litman, and J. Wiebe, "Evaluating an opinion annotation scheme using a new multi-perspective question and answer corpus," in *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text*, AAAI Technical Report SS-04-07.

[284] V. Stoyanov, C. Cardie, and J. Wiebe, "Multi-perspective question answering using the OpQA corpus," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 923–930, Vancouver, British Columbia, Canada: Association for Computational Linguistics, October 2005.

[285] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," *IEEE Transactions on Fuzzy Systems*, vol. 9, pp. 483–496, 2001.

[286] M. Taboada, C. Anthony, and K. Voll, "Methods for creating semantic orientation dictionaries," in *Conference on Language Resources and Evaluation (LREC)*, pp. 427–432, 2006.

[287] M. Taboada, M. A. Gillies, and P. McFetridge, "Sentiment classification techniques for tracking literary reputation," in *LREC Workshop: Towards Computational Models of Literary Analysis*, pp. 36–43, 2006.

[288] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientation of words using spin model," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 133–140, 2005.

[289] H. Takamura, T. Inui, and M. Okumura, "Latent variable models for semantic orientations of phrases," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

[290] H. Takamura, T. Inui, and M. Okumura, "Extracting semantic orientations of phrases from dictionary," in *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*, 2007.

[291] K. Tateishi, Y. Ishiguro, and T. Fukushima, "Opinion information retrieval from the internet," *Information Processing Society of Japan (IPSJ) SIG Notes,* 2001, vol. 69, no. 7, pp. 75–82, 2001. (Also cited as "A reputation search engine that gathers people's opinions from the Internet", IPSJ Technical Report NL-14411. In Japanese).

[292] J. Tatemura, "Virtual reviewers for collaborative exploration of movie reviews," in *Proceedings of Intelligent User Interfaces (IUI)*, pp. 272–275, 2000.

[293] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter, "PHOAKS: A system for sharing recommendations," *Communications of the Association for Computing Machinery (CACM)*, vol. 40, pp. 59–62, 1997.

[294] M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 327–335, 2006.

[295] R. Tokuhisa and R. Terashima, "Relationship between utterances and 'enthusiasm' in non-task-oriented conversational dialogue," in *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pp. 161–167, Sydney, Australia: Association for Computational Linguistics, July 2006.

[296] R. M. Tong, "An operational system for detecting and tracking opinions in on-line discussion," in *Proceedings of the Workshop on Operational Text Classification (OTC)*, 2001.

[297] R. Tumarkin and R. F. Whitelaw, "News or noise? Internet postings and stock prices," *Financial Analysts Journal*, vol. 57, pp. 41–51, May/June 2001.

[298] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 417–424, 2002.

[299] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, pp. 315–346, 2003.

[300] S. Wan and K. McKeown, "Generating overview summaries of ongoing email thread discussions," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 549–555, Geneva, Switzerland, 2004.

[301] M. White, C. Cardie, and V. Ng, "Detecting discrepancies in numeric estimates using multidocument hypertext summaries," in *Proceedings of the Conference on Human Language Technology*, pp. 336–341, 2002.

[302] M. White, C. Cardie, V. Ng, K. Wagstaff, and D. McCullough, "Detecting discrepancies and improving intelligibility: Two preliminary evaluations of RIPTIDES," in *Proceedings of the Document Understanding Conference (DUC)*, 2001.

[303] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," in *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, pp. 625–631, ACM, 2005.

[304] J. Wiebe, "Learning subjective adjectives from corpora," in *Proceedings of AAAI*, 2000.

[305] J. Wiebe, E. Breck, C. Buckley, C. Cardie, P. Davis, B. Fraser, D. Litman, D. Pierce, E. Riloff, T. Wilson, D. Day, and M. Maybury, "Recognizing and organizing opinions expressed in the world press," in *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, 2003.

[306] J. Wiebe and R. Bruce, "Probabilistic classifiers for tracking point of view," in *Proceedings of the AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pp. 181–187, 1995.

[307] J. Wiebe and R. Mihalcea, "Word sense and subjectivity," in *Proceedings of the Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL)*, 2006.

[308] J. Wiebe and T. Wilson, "Learning to disambiguate potentially subjective expressions," in *Proceedings of the Conference on Natural Language Learning (CoNLL)*, pp. 112–118, 2002.

[309] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation (formerly Computers and the Humanities)*, vol. 39, pp. 164–210, 2005.

[310] J. M. Wiebe, "Identifying subjective characters in narrative," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 401–408, 1990.

[311] J. M. Wiebe, "Tracking point of view in narrative," *Computational Linguistics*, vol. 20, pp. 233–287, 1994.

[312] J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, "Development and use of a gold standard data set for subjectivity classifications," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 246–253, 1999.

[313] J. M. Wiebe and W. J. Rapaport, "A computational theory of perspective and reference in narrative," in *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 131–138, 1988.

[314] J. M. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing),* number 3406 in *Lecture Notes in Computer Science*, pp. 486–497, 2005.

[315] J. M. Wiebe, T. Wilson, and M. Bell, "Identifying collocations for recognizing opinions," in *Proceedings of the ACL/EACL Workshop on Collocation: Computational Extraction, Analysis, and Exploitation*, 2001.

[316] J. M. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, "Learning subjective language," *Computational Linguistics*, vol. 30, pp. 277–308, September 2004.

[317] Y. Wilks and J. Bien, "Beliefs, points of view and multiple environments," in *Proceedings of the international NATO symposium on artificial and human intelligence*, pp. 147–171, USA, New York, NY: Elsevier North-Holland, Inc., 1984.

[318] Y. Wilks and M. Stevenson, "The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation," *Journal of Natural Language Engineering*, vol. 4, pp. 135–144, 1998.

[319] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 347–354, 2005.

[320] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses," in *Proceedings of AAAI*, pp. 761–769, 2004. (Extended version in *Computational Intelligence*, vol. 22, no. 2, pp. 73–99, 2006).

[321] H. Yang, L. Si, and J. Callan, "Knowledge transfer and opinion detection in the TREC2006 blog track," in *Proceedings of TREC*, 2006.

[322] K. Yang, N. Yu, A. Valerio, and H. Zhang, "WIDIT in TREC-2006 blog track," in *Proceedings of TREC*, 2006.

[323] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2003.

[324] J. Yi and W. Niblack, "Sentiment mining in WebFountain," in *Proceedings of the International Conference on Data Engineering (ICDE)*, 2005.

[325] P.-L. Yin, "Information dispersion and auction prices," Social Science Research Network (SSRN) Working Paper Series, Version dated March 2005.

[326] H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,"

in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2003.

[327] J. Zabin and A. Jefferies, "Social media monitoring and analysis: Generating consumer insights from online conversation," Aberdeen Group Benchmark Report, January 2008.

[328] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, pp. 51–57, 2006.

[329] L. Zhou, J. K. Burgeon, and D. P. Twitchell, "A longitudinal analysis of language behavior of deception in e-mail," in *Proceedings of Intelligence and Security Informatics (ISI),* number 2665 in *Lecture Notes in Computer Science*, p. 959, 2008.

[330] L. Zhou and E. Hovy, "On the summarization of dynamically introduced information: Online discussions and blogs," in *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 237–242, 2006.

[331] F. Zhu and X. Zhang, "The influence of online consumer reviews on the demand for experience goods: The case of video games," in *International Conference on Information Systems (ICIS)*, 2006.

[332] L. Zhuang, F. Jing, X.-Y. Zhu, and L. Zhang, "Movie review mining and summarization," in *Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM)*, 2006.