Evaluative Language Beyond Bags of Words: Linguistic Insights and Computational Applications

Farah Benamara*
IRIT-Université de Toulouse

Maite Taboada**
Simon Fraser University

Yannick Mathieu†
LLF-CNRS

The study of evaluation, affect, and subjectivity is a multidisciplinary enterprise, including sociology, psychology, economics, linguistics, and computer science. A number of excellent computational linguistics and linguistic surveys of the field exist. Most surveys, however, do not bring the two disciplines together to show how methods from linguistics can benefit computational sentiment analysis systems. In this survey, we show how incorporating linguistic insights, discourse information, and other contextual phenomena, in combination with the statistical exploitation of data, can result in an improvement over approaches that take advantage of only one of these perspectives. We first provide a comprehensive introduction to evaluative language from both a linguistic and computational perspective. We then argue that the standard computational definition of the concept of evaluative language neglects the dynamic nature of evaluation, in which the interpretation of a given evaluation depends on linguistic and extra-linguistic contextual factors. We thus propose a dynamic definition that incorporates update functions. The update functions allow for different contextual aspects to be incorporated into the calculation of sentiment for evaluative words or expressions, and can be applied at all levels of discourse. We explore each level and highlight which linguistic aspects contribute to accurate extraction of sentiment. We end the review by outlining what we believe the future directions of sentiment analysis are, and the role that discourse and contextual information need to play.

* 118 Route de Narbonne, 31062 Toulouse, France. E-mail: benamara@irit.fr.
** 8888 University Dr., Burnaby, BC, V5A 1S6, Canada. E-mail: mtaboada@sfu.ca.
† Université Paris 7, Paris Diderot, Bat. Olympe de Gouges 8 place Paul Ricœr, case 7031, 75205 Paris Cedex 13, France. E-mail: ymathieu@linguist.univ-paris-diderot.fr.

Submission received: 30 October 2015; revised version received: 19 February 2016; accepted for publication: 8 June 2016.

doi:10.1162/COLI_a_00278

1. Introduction

Evaluation aspects of language allow us to convey feelings, assessments of people, situations and objects, and to share and contrast those opinions with other speakers. An increased interest in subjectivity, evaluation, and opinion can be viewed as part of what has been termed the **affective turn** in philosophy, sociology, and political science (Clough and Halley 2007), and **affective computing** in artificial intelligence (Picard 1997). This interest has met with the rise of the social web, and the possibility of widely broadcasting emotions, evaluations, and opinions.

The study of evaluation, affect, and subjectivity is a multidisciplinary enterprise, including sociology (Voas 2014), psychology (Ortony et al. 1988; Davidson et al. 2003), economics (Rick and Loewenstein 2008), and computer science (Pang and Lee 2008; Scherer et al. 2010; Cambria and Hussain 2012; Liu 2012, 2015). In linguistics, studies have been framed within a wide range of theories, such as Appraisal theory (Martin and White 2005), stance (Biber and Finegan 1989), evaluation (Hunston and Thompson 2000b), and nonveridicality (Taboada and Trnavac 2013) (cf. Section 2.3). In computer science, most current research examines the expression and automatic extraction of opinion at three main levels of granularity (Liu 2012): the document, the sentence, and the aspect. The first level aims to categorize documents globally as being positive or negative, whereas the second one determines the subjective orientation and then the opinion orientation (positive or negative) of sequences of words in the sentence that are determined to be subjective. The aspect level focuses on extracting opinions according to the target domain features or aspects (cf. Section 3.2). Extraction methods used in each of the three levels rely on a variety of approaches going from bag-of-words representations and structured representations based on the use of grammar and dependency relations, to more sophisticated models that address the complexity of language, such as negation, speculation, and various context-dependent phenomena.

A number of excellent computational linguistics and linguistic surveys of the field exist. Hunston and Thompson (2000b) proposed in their book *Evaluation in Text* an overview of how evaluative expressions can be analyzed lexically, grammatically, and textually, and a recent edited collection (Thompson and Alba-Juez 2014) focuses on theoretical and empirical studies of evaluative text at different linguistic levels (phonological, lexical, or semantic), and in different text genres and contexts. Computational approaches to evaluative text (known as **sentiment analysis**) have been reviewed by Pang and Lee (2008), Liu (2012, 2015), and Feldman (2013), among others. This is clearly an important topic: A Google Scholar search for "sentiment analysis" yields about 31,000 publications, and the Pang and Lee survey alone has more than 4,700 citations.

In this survey, we focus on linguistic aspects of evaluative language, and show how the treatment of linguistic phenomena, in particular at the discourse level, can benefit computational sentiment analysis systems, and help such systems advance beyond representations that include only bags of words or bags of sentences. We also show how discourse and pragmatic information can help move beyond current sentence-level approaches that typically account for local contextual phenomena, and do so by relying on polarity lexicons and shallow or deep syntactic parsing.

More importantly, we argue that incorporating linguistic insights, discourse information, and other contextual phenomena, in combination with the statistical exploitation of data, can result in an improvement over approaches that take advantage of only one of those perspectives. Together with an affective turn, we believe computational linguistics is currently experiencing a **discourse turn**, a growing awareness of how multiple sources of information, and especially information from context and discourse,

can have a positive impact on a range of computational applications (Webber et al. 2012). We believe that future breakthroughs in natural language processing (NLP) applications will require developing synergies between machine learning techniques and bag-of-words representations on the one hand, and in-depth theoretical accounts, together with detailed analyses of linguistic phenomena, on the other hand. In particular, we investigate:

- The complex lexical semantics of evaluative expressions, including semantic categorization and domain dependency.
- Contextual effects deriving from negation, modality, and nonveridicality.
- Topicality, coherence relations, discourse structure, and other related discourse-level phenomena.
- Complex evaluative language phenomena such as implicit evaluation, figurative language (irony, sarcasm), and intent detection.
- Extra-linguistic information, such as social network structure and user profiles.

The article is organized around four main parts. The first (in Section 2) provides a comprehensive introduction to evaluative language, focusing on linguistic theories and how evaluative phenomena are described. Section 3 contains a computational definition of the problem statement, and a brief explanation of the standard approaches that have been applied in sentiment analysis. We argue that the standard definition of the concept of evaluative language is not sufficient to address discourse and contextual phenomena, and thus Section 4 introduces a new dynamic definition. Then, in Sections 5 and 6 we move from current approaches to future directions in sentiment analysis. Under current approaches, we discuss local and sentence-level phenomena, and in the new directions section we overview the main discourse and context-level phenomena that we believe need to be addressed to accurately capture sentiment. The last section summarizes what we believe are the future directions of sentiment analysis, emphasizing the importance of discourse and contextual information.

2. Linguistic Approaches to Evaluative Text

Evaluation, as a cover term for many approaches, has long been the object of interest in linguistics. Unfortunately, however, no single theoretical framework has attempted to examine and account for the full range of evaluative devices available in language. The main exception is the Appraisal framework (Martin and White 2005), which does provide a very rich description of how evaluation is expressed and also implied in text. Before we devote a good part of this section to Appraisal, we briefly discuss other approaches in linguistics that have dealt with the resources deployed in the expression of evaluation. We are aware that the term *evaluation* is also used within the context of testing and benchmarking in computer science. To avoid confusion, we explicitly use the terms *evaluative language* or *evaluative expressions* to refer to the computational study of evaluative text. When we use the shorter form *evaluation,* it generally refers to expression of evaluation and opinion in language.

2.1 Stance

We could begin at different points, but one of the earliest attempts to describe evaluative language comes from Biber, Finegan, and colleagues. In a series of papers (Biber and Finegan 1988, 1989; Conrad and Biber 2000), they describe stance as the expression of the speaker's attitudes, feelings, and judgment, as well as their commitment towards the message. Stance, in this sense, encompasses evidentiality (commitment towards the message) and affect (positive or negative evaluation). The initial focus (Biber and Finegan 1988) was on adverbials (*personally, frankly, unquestionably, of course, apparently*). Later on, Biber and Finegan (1989) added adjectives, verbs, modal verbs, and hedges as markers of evidentiality and affect. In both papers, the classification of stance markers leads to an analysis of texts based on cluster analysis. The clusters represent different stance styles, such as Emphatic Expression of Affect, Expository Expression of Doubt, or Faceless. Within each of those styles one can find different genres/registers. For instance, Emphatic Expression of Affect includes texts such as personal letters, face-to-face conversations, and romance fiction. On the other hand, Faceless texts were found in the following genres: academic prose, press reportage, radio broadcasts, biographies, official documents and letters, among others.

This kind of genre classification, identifying how much the writer/speaker is involved, and how much evaluation is being displayed, is useful to gauge the "volume" of a text. If we know that we are dealing with a highly subjective genre, then finding high levels of subjective and evaluative words is not surprising. On the other hand, in what Biber and Finegan call Faceless texts, a high proportion of evaluative expressions is more significant, and is probably indicative of an overtly subjective text (i.e., one with a higher "volume," from the point of view of evaluative language).

The term *stance*, in Biber and Finegan's work, is quite close to our use of evaluative language or opinion. There is another meaning of stance, with regard to position in a debate, namely, for or against, or ideological position (Somasundaran and Wiebe 2009). We will not have much to say about recognizing positions, although it is an active and related area of research (Thomas et al. 2006; Hasan and Ng 2013). The SemEval 2016 competition included both sentiment analysis and stance detection tasks, where stance is defined as "automatically determining from text whether the author is in favor of, against, or neutral towards a proposition or target (Mohammad et al. in press). Targets mentioned in the SemEval task include legalization of abortion, atheism, climate change, or Hillary Clinton. The SemEval task organizers further specify that sentiment analysis and stance detection are different in that sentiment analysis attempts to capture polarity in a piece of text, whereas stance detection aims at extracting the author's standpoint towards a target that may not be explicitly mentioned in the text. Stance detection may be seen as a textual entailment task, because the stance may have to be inferred from what is present in the text, such as the fact that the author does not like Hillary Clinton if they express favorable views of another candidate.

2.2 Evidentiality

Biber and Finegan only briefly touch on evidentiality, but it is an area of study in its own right, examining the linguistic coding of attitudes towards knowledge, and in particular ways of expressing it cross-linguistically (Chafe and Nichols 1986). Evidentiality expresses three basic types of meaning (Chafe 1986; Boye and Harder 2009):

- Reliability of knowledge, with adverbs such as *maybe, probably, surely*.

- Mode of knowing: belief with constructions such as *I think, I guess*; induction with *must, seem, evidently*; and deduction with *should, could, presumably*.
- Source of knowledge indicated by verbs indicating the source of input (*see, hear, feel*).

Evidentiality has not received a great deal of attention in the sentiment analysis literature, even though it provides a set of resources to ascertain the reliability of an opinion, and is within the realm of speculation (Vincze et al. 2008; Saurí and Pustejovsky 2009). Some aspects of it, however, are captured in the Engagement system within the Appraisal framework, which we will discuss later.

2.3 Nonveridicality

Nonveridicality usually includes a host of phenomena that indicate that individual words and phrases may not be reliable for the purposes of sentiment analysis, also known as **irrealis**. Irrealis in general refers to expressions that indicate that the events mentioned in an utterance are not factual. Nonveridicality is wider, including all contexts that are not veridical—that is, which are not based on truth or existence (Giannakidou 1995; Zwarts 1995). The class of nonveridical operators typically includes negation, modal verbs, intensional verbs (*believe, think, want, suggest*), imperatives, questions, protasis of conditionals, habituals, and the subjunctive, (in languages that have an expression of subjunctive) (Trnavac and Taboada 2012).

Nonveridicality is different from evidentiality in that evidential markers may code nonveridical meanings, but also different shades within veridical propositions. A speaker may present something as fact (thus veridical), but at the same time distance themselves from the reliability of the statement through evidential markers (e.g., *Critics say it's a good movie*).

Nonveridicality is relevant in sentiment analysis because evaluative expressions in the scope of a nonveridical operator may not be reliable, that is, they may express the opposite polarity of the evaluative expression (when in the scope of negation), may see their polarity downtoned or otherwise hedged (in the presence of modal verbs or intensional verbs), or may interact with nonveridical operators in complex ways. The presence of conditionals as nonveridical operators has led to some work on the nature of coherence relations and their role in evaluative language (Asher et al. 2009; Trnavac and Taboada 2012). We discuss computational treatment of nonveridicality in Section 5.

2.4 Subjectivity

The term **subjectivity** has different senses and has been adopted in computational linguistics to encompass automatic extraction of both sentiment and polarity (Wiebe et al. 2004). Subjectivity in linguistics is a much more general and complex phenomenon, often relating to point of view. Researchers working with this definition are interested in deixis, locative expressions, and use of modal verbs, among other phenomena (Langacker 1990). The connections to sentiment analysis are obvious, because epistemic modals convey subjective meaning and are related to evidentiality.

White (2004) has also written about subjectivity as point of view, and some of his work is framed within the Appraisal framework (see Section 2.6). White discusses the opposition between objective and subjective statements in media discourse, and how

the latter can be conveyed not through overt subjective expressions, but through association, metaphor, or inference. Distilling such forms of opinion is certainly important in sentiment analysis, but also particularly difficult, as they rely on world knowledge that is inaccessible to most systems. We will discuss the role of some of these in Section 6.

Our discussion of subjectivity will then be circumscribed to the distinction between objective and subjective statements. Wiebe and colleagues have devoted considerable effort to finding indicators of subjectivity in sentences (Wiebe et al. 2004; Wiebe and Riloff 2005; Wilson et al. 2006). They propose a set of clues to subjectivity, some of them lexical and some syntactic. Among the lexical clues are psychological verbs and verbs of judgment (*dread, love, commend, reprove*); verbs and adjectives that usually involve an experiencer (*fuss, worry, please, upset, embarrass, dislike*); and adjectives that have been previously annotated for polarity. The syntactic clues are learned from manually annotated data (Riloff and Wiebe 2003; Wiebe et al. 2003).

2.5 Evaluation and Pattern Grammar

Under the label "evaluation" we will examine a fruitful area of research that has studied evaluation and opinion within a functional framework. The best example of such endeavor is the edited volume by Hunston and Thompson (2000b). Hunston and Thompson (2000a), in their Introduction, propose that there are two aspects to evaluative language: modality and something else, which is variously called evaluation, appraisal, or stance. Modality tends to express opinions about propositions, such as their likelihood (*It may rain*). It also tends to be more grammaticalized. Evaluation, on the other hand, expresses opinions about entities, and is mostly (although not exclusively) expressed through adjectives. Hunston and Thompson review some of the main approaches to the study of subjectivity, and note that there seem to be approaches that separate modality and evaluation as two distinct phenomena (Halliday 1985; Martin 2000; White 2003; Halliday and Matthiessen 2014). Other researchers combine the two expressions of opinion, often under one label, such as stance (Biber and Finegan 1989; Conrad and Biber 2000). In sentiment analysis, modality has been included as one of the phenomena that affect evaluative language. A comprehensive treatment of such phenomena, however, such as the one offered by nonveridicality (see Section 2.3) seems beneficial, as it provides a unified framework to deal with all valence shifters. Hunston and Thompson (2000a) espouse a combining approach, and propose that the cover term for both aspects should be "evaluation."

Hunston and Thompson (2000a, page 6) argue that evaluation has three major functions:

- To express the speaker's or writer's opinion.
- To construct and maintain relations between the speaker/writer and hearer/reader.
- To organize the discourse.

The first two functions are variously discussed in the sentiment literature, with emphasis on the first one, and some treatment of the second, in terms of how the writer presents, manipulates, or summarizes information. This is an area covered by the Engagement system in Appraisal. But it is the third aspect that has received the least attention in computational treatments of evaluative language, perhaps because it is the most difficult to process. Hunston and Thompson (2000a) present this function

as one of argumentation. A writer not only expresses opinion, and engages with the reader, but also presents arguments in a certain order and with a certain organization. Evaluation at the end of units such as paragraphs indicates that a point in the argument has been made, and that the writer assumes the reader accepts that point. In general, Hunston and Thompson argue that evaluation is expressed as much by text as it is by individual lexical items and by grammar. We would argue that, in addition to location in the text, other textual and discourse characteristics, in particular coherence relations (see Section 6.1), play a role in the interpretation of evaluation.

Pattern-based descriptions of language are of special relevance here, because they avoid a distinction between lexis and grammar, but rather treat them as part of the same object of description (Hunston and Francis 2000). Subjectivity spans over the two, sometimes being conveyed by a single word, sometimes by a phrase, and sometimes by an entire grammatical structure. Hunston and Francis (2000, page 37) define patterns of a word as "all the words and structures which are regularly associated with the word and which contribute to its meaning."

The most in-depth description of patterns and evaluation is Hunston (2011), where a case is clearly made that certain patterns contribute to evaluative meanings, with a distinction between patterns that perform the function of evaluation, namely, "performative" patterns, according to Hunston (2011, page 139), and patterns that report evaluation. Examples of performative patterns are *it* and *there* patterns, as in *It is amazing that. . . ; There is something admirable about. . .* An example of a pattern that reports evaluation is Verb + *that*, as in *Most people said that he was aloof*. Hunston also discusses phrases that accompany evaluation, such as as *(is) humanly possible; to the point of;* or *bordering on*.

Work on evaluative language within linguistics has seen steady publication in the last few years, including a follow-up volume, edited by Thompson and Alba-Juez (2014). This new collection places even more emphasis on the whole-text nature of evaluation: The title is *Evaluation in Context*, and the papers therein clearly demonstrate the pervasive nature of evaluation, in addition to its connection to emotion. In the introduction, Alba-Juez and Thompson argue that evaluation permeates all aspects of the language:

- The phonological level through intonation and pitch.
- The morphological with suffixes in morphologically rich languages, but also in English. Consider the use of the suffix *–let* in the term *deanlet*, coined by Ginsberg (2011) to describe a new class of academic administrators.
- The lexical level. This is self-evident, as most of our discussion, and the discussion in the sentiment analysis literature has focused on words.
- The syntactic level. Systems of modality and nonveridicality, but also word order and structural aspects, as discussed within pattern grammar.
- The semantic level. This is the thorniest aspect, and includes pragmatics as well. It refers to the non-compositional nature of some evaluative expressions and their context-dependence. *A long meal* may be positive if it is with friends, but is perhaps negative with business relations.

It is clear, then, that focusing on only one of those levels, the lexical level, will result in an incomplete picture of the evaluative spectrum that language has to offer.

That is why we believe that bags-of-words approaches need to be complemented with information from other levels.

2.6 Appraisal

Appraisal belongs in the systemic-functional tradition started by Halliday (Halliday and Matthiessen 2014), and has been developed by Jim Martin, Peter White, and colleagues (Martin 2000; Martin and White 2005; White 2012; Martin 2014). Martin (2000) characterizes appraisal as the set of resources used to negotiate emotions, judgments, and valuations, alongside resources for amplifying and engaging with those evaluations. He considers that appraisal resources form a system of their own within the language (in the sense of system within Systemic Functional Linguistics), and divides the Appraisal system into three distinct sub-systems (see Figure 1): Attitude, Graduation, and Engagement.

The central aspect of the theory is the Attitude system and its three subsystems, Affect, Judgment, and Appreciation. Affect is used to construe emotional responses about the speaker or somebody else's reactions (e.g., *happiness, sadness, fear*). Judgment conveys moral evaluations of character about somebody else than the speaker (e.g., *ethical, deceptive, brave*). Appreciation captures aesthetic qualities of objects and natural phenomena (*remarkable, elegant, innovative*).

Computational treatment of Appraisal (see Section 6) typically involves inscribed instances, that is, those that are explicitly present in the text via a word with positive or negative meaning. Instances that are not inscribed are considered to be invoked (also sometimes called evoked), in which "an evaluative response is projected by reference to events or states which are conventionally prized" (Hunston and Thompson 2000b, page 142). Thus, *a bright kid* or *a vicious kid* are inscribed. On the other hand, *a kid who reads a lot* or *a kid who tears the wings off butterflies* present invoked Appraisal. Because it is easier to identify automatically, most research in computational linguistics has focused on inscribed Appraisal and evaluation. We discuss implicit or invoked evaluation in Section 6, together with metaphor, irony, and sarcasm.

In addition to the very central Attitude system, Martin and White (2005) argue that two other systems play a crucial role in the expression of opinion. The Graduation system is responsible for a speaker's ability to intensify or weaken the strength of the

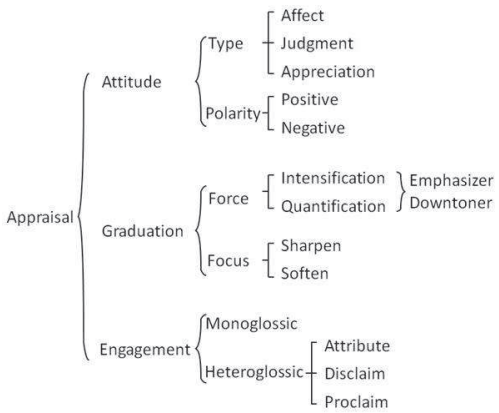

Figure 1
The Appraisal system.

opinions that they express, and has Force and Focus as subsystems. Force captures the intensification or downtoning of words that are inherently gradable, whereas Focus represents how speakers can sharpen or soften words that are usually non-gradable. Examples of intensification and downtoning are *somewhat interesting* and *a little bit sad*. In *a true friend* the meaning of *friend*, usually a non-gradable word, is sharpened. On the other hand, *a friend of sorts* implies a softening of the meaning.

The Engagement system is the set of linguistic options that allow the individual to convey the degree of their commitment to the opinion being presented. It makes a fundamental distinction between heteroglossic and monoglossic expressions, following proposals by Bakhtin (1981). In a heteroglossic expression, inter-subjective positioning is open, because utterances invoke, acknowledge, respond to, anticipate, revise, or challenge a range of convergent and divergent alternative utterances (Martin and White 2005; White 2012, 2003). The other option is monoglossia, where no alternative view or openness to accept one is present. Monoglossic utterances are presented as facts.

Appraisal is quite well understood (at least in English), with a wide range of studies dealing with different genres and other languages. The Appraisal framework provides a very rich description of different aspects of evaluation and subjective positioning. From the early stages of sentiment analysis and opinion mining, it was always obvious to researchers in those areas that Appraisal could be helpful in identifying types of opinions, and in pinpointing the contribution of intensifiers and valence shifters (Polanyi and Zaenen 2006) in general. We further discuss computational treatments of Appraisal in Section 5.1.4.

3. Evaluative Language in Computational Linguistics

After an introduction to evaluation and subjectivity in linguistics, we now present research in computational linguistics, starting with a basic problem definition and a common characterization of evaluative language. We then provide a short overview of standard approaches in sentiment analysis.

3.1 Problem Definition

In computational linguistics, evaluative language is used as an umbrella term that covers a variety of phenomena including opinion, sentiment, attitude, appraisal, affect, point of view, subjectivity, belief, desire, and speculation. Although computational linguists do not always seek to distinguish between these phenomena, most of them commonly agree to define evaluative language as being a subjective piece of language expressed by a holder (a person, a group, an institution) towards a topic or target (an object, a person, an action, an event). A key element in this definition is that an evaluative expression is always associated with a *polarized scale* regarding social or moral norms (*bad* vs. *good*, *love* vs. *hate*, *in favor of* vs. *against*, *prefer* vs. *dislike*, *better* vs. *worse*, etc.). Hence, the sentence in Example (1) is evaluative because it expresses a positive evaluation towards the food served in a restaurant, whereas Example (2) is not. Indeed, Example (1) can be paraphrased as *I loved the food*, or *Go to this restaurant; I recommend it*. On the other hand, Example (2) expresses an emotion, namely, the author's subjective feelings and affect, and is not as easily characterized on a polar scale (except if we accept that jealousy is conventionally defined as negative).

(1) This restaurant serves incredibly delicious food.

(2) I am jealous of the chef.

In the remainder of this survey, we focus on automatic detection of *polarized evaluation* in text, excluding other forms of evaluative language. For the sake of readability, the terms evaluation, evaluative language, and polarized evaluation are used interchangeably. We leave outside of the scope of this survey research on emotion detection and classification, which is surveyed by Khurshid (2013) and Mohammad (2016). Related, but also somewhat beyond our scope, is work on detecting negation and speculation, in particular in biomedical text (Saurí and Pustejovsky 2009; Councill et al. 2010; Morante and Sporleder 2012a; Cruz et al. 2016).

A frequently used definition of evaluative language as a structured model has been proposed by Liu (2012), drawing from Kim and Hovy (2004) and Hu and Liu (2004a). This model is a quintuple (e, a, s, h, t) where e is the entity that is the topic or target of the opinion (*restaurant* in Example (1)), a the specific aspect or feature of that entity (*food*), s the sentiment or evaluation towards a (*incredibly delicious*), h the opinion holder or source (the author in our example), and t the posting time of s. Liu (2012) further represents s by a triple (y, o, i) in order to capture the sentiment type or sentiment semantic category y (*delicious* denotes an appreciation), sentiment polarity or orientation o (*delicious* is positive), and sentiment valence i, also known as rate or strength, that indicates the degree of evaluation on a given scale (*incredibly delicious* is stronger than *delicious*). Sentiment type can be defined according to linguistic-based or psychology-based classifications. We discuss some of them in Section 5.1.

Liu (2012) notes that sentiment analysis based on this model corresponds to aspect-based or feature-based sentiment analysis in which systems relate each sentiment s to an aspect a (or more generally an entity e). Aspect-based models of sentiment have been most popular in the domain of consumer reviews, where movies, books, restaurants, hotels, or other consumer products are evaluated in a decompositional manner, with each aspect (e.g., ambiance, food, service, or price for a restaurant) evaluated separately. In Section 4.2, we will introduce a new definition, because we believe that contextual phenomena need to be accounted for in a definition of evaluative language.

3.2 Standard Approaches

The study of how to automatically extract evaluation from natural language data began in the 1990s (Hearst 1992; Wiebe 1994; Spertus 1997; Hatzivassiloglou and McKeown 1997; Bruce and Wiebe 1999). These initial efforts proceeded in a rather similar way by making an in-depth inspection of linguistic properties of evaluative language a prior step to any automatic treatment. The detection of expressions of evaluation relied not only on individual words taken in isolation but also on surrounding material or contextual information that were considered essential for a better understanding of evaluative language in text both at the expression and document level. For example, Hearst (1992) proposed a model inspired in cognitive linguistics, in which portions of a text are interpreted following a directionality criterion to determine if an author is in favor of, neutral, or opposed to some events in a document. The model involved a set of grammatical patterns relying on a syntactic parser. Spertus (1997) automatically identified hostile messages by leveraging insulting words and the syntactic context in which they are used (such as imperative statements, which tend to be more insulting). Wiebe (1994) proposed an algorithm to identify subjectivity in narratives following Banfield's (1982) theory, which characterized sentences of narration as objective (narrating events or describing the fictional world) or subjective (expressing the author's thoughts or perceptions). This algorithm relied on *the sentence proximity*

assumption where subjective vs. objective sentences were postulated to be more likely to appear together. Hatzivassiloglou and McKeown (1997) studied subjectivity at the expression level, focusing on the prior sentiment orientation of adjectives. Using a supervised learning method, they empirically demonstrated that adjectives connected with the conjunctions *and* and *or* usually share the same orientation, whereas the conjunction *but* tends to connect adjectives with opposite orientations. Turney and Littman (2002) extended this approach to infer semantic orientation of words (not only adjectives) in large corpora.

Since 2000, computational approaches to evaluative language, labeled either as sentiment analysis or opinion mining, have become one of the most popular applications of NLP in academic research institutions and industry. In the new century, however, and moving away from the early approaches where linguistics (and thus context) played a central role, sentiment analysis has become more a computational modelization problem than a linguistic one. Popular and commonly used approaches ("standard" approaches) focus on the automatic extraction of one or several elements of the quadruple (*e*, *a*, *s*, *h*), making sentiment analysis a field that involves roughly three main sub-tasks: (1) topic/aspect extraction, (2) holder identification, and (3) sentiment determination. These tasks are either performed independently from each other or simultaneously. When sub-tasks (1), (2), and (3) are treated independently, the dependencies between sentiment and topics are ignored. To account for these dependencies, two main approaches have been explored: sequential learning (Jakob and Gurevych 2010; Yang and Cardie 2012; Mitchell et al. 2013; Vo and Zhang 2015) and probabilistic joint sentiment-topic models, which are capable of detecting sentiment and topic at the same time in an unsupervised fashion (Hu and Liu 2004a; Zhuang et al. 2006; Lin and He 2009; Wang and Ester 2014; Nguyen and Shirai 2015), even though some approaches still require a seed of sentiment-bearing words and/or aspect seeds (Qiu et al. 2009; Hai et al. 2012).

In the following sections we overview the standard approaches to evaluative text on the three tasks just mentioned. The overview is necessarily brief, because our aim is not to provide an exhaustive survey of the field of sentiment analysis, but to focus on how more linguistically informed representations can contribute to the analysis and extraction of evaluation. For an excellent benchmark comparison of twenty-four different sentiment analysis systems, see Ribeiro et al. 2016.

3.2.1 Topic/Aspect and Holder Detection. Tasks (1) and (2) are important sub-tasks in sentiment analysis (Hu and Liu 2004a; Kim and Hovy 2005; Popescu and Etzioni 2005; Stoyanov and Cardie 2008; Wiegand and Klakow 2010). The holder can be the author, expressing their own evaluation (*The movie is great*), or the author stating or reporting someone else's evaluation (*My mother loves the movie; My mother said that the movie is great*) (Wiebe and Riloff 2005). The holder evaluates a topic or target that is the entity or a part or attribute of the entity that the sentiment is predicated upon (Liu 2015). A topic is thus a global entity *e* (e.g., a product, service, person, event, or issue) organized hierarchically into a set of attributes or aspects *a* (e.g., *engine, tires* are part of the entity *car*), as can be done in a thesaurus or domain ontology.

For holder recognition, it has been shown that semantic role labeling à la PropBank or FrameNet is beneficial (Bethard et al. 2004; Choi et al. 2006; Kim and Hovy 2006; Gangemi et al. 2014), although Ruppenhofer et al. (2008) argue that evaluation that is connected to its source indirectly via attribution poses challenges that go beyond the capabilities of automatic semantic role labeling, and that discourse structure has to be considered. Topic and aspect recognition, on the other hand, are seen as information

extraction tasks that generally exploit noun or noun phrases, dependency relations, some syntactic patterns at the sentence level, knowledge representation paradigms (like hierarchies or domain ontologies), and external sources (e.g., Wikipedia) to identify explicit aspects (Hu and Liu 2004a; Popescu and Etzioni 2005; Zhao and Li 2009; Wu et al. 2009).[1]

3.2.2 Sentiment Determination. Task (3), sentiment determination, is probably the most studied. It consists of identifying polarity or orientation. In its simplest form, it corresponds to the binary orientation (positive or negative) of a subjective span regardless of external context within or outside a sentence or document. Some researchers argue instead for a ternary classification, with a neutral category to indicate the absence of evaluation (Koppel and Schler 2006; Agarwal et al. 2011). The intensity of a subjective span is often combined with its prior binary polarity to form an **evaluation score** that tells us about the degree of the evaluation, that is, how positive or negative the word is. Several types of scales have been used in sentiment analysis research, going from continuous scales (Benamara et al. 2007) to discrete ones (Taboada et al. 2011). For example, if we have a three-point scale to encode strength, we can propose $score(good) = +1$ and $score(brilliant) = +3$. Generally, there is no consensus on how many points are needed on the scale, but the chosen length of the scale has to ensure a trade-off between a fine-grained categorization of subjective words and the reliability of this categorization with respect to human judgments.

Two main approaches have been proposed, corpus-based (mostly using machine learning techniques) and lexicon-based systems, which typically perform, in their most simplistic form, a lexical lookup and combine the meaning of individual words to derive the overall sentence or document sentiment (Hu and Liu 2004a; Kim and Hovy 2004; Taboada et al. 2011). For example, Kim and Hovy (2004) propose a three-step algorithm to compute sentence-level opinion orientation: First find the sentiment orientation of a word using a propagation algorithm that estimates its polarity on the basis of paradigmatic relations it may have with seed sentiments words, as encoded in WordNet. Then, compute the sentence sentiment orientation. In this second step, only words that appear near a holder and/or topic are considered. The sentiment scores of these words are then combined with various aggregation functions (average, geometric mean, etc.). Hu and Liu (2004a) propose a similar approach, taking into account in addition opposition words (like *but*, *however*). They also generate a feature-based summary of each review.

At the sentence level, the task is to determine the subjective orientation and then the opinion orientation of sequences of words in the sentence that are determined to be subjective or express an opinion (Riloff and Wiebe 2003; Yu and Vasileios 2003; Wiebe and Riloff 2005; Taboada et al. 2011), with the assumption that each sentence usually contains a single opinion. To better compute the contextual polarity of opinion expressions, some researchers have used subjectivity word sense disambiguation to identify whether a given word has a subjective or an objective sense (Akkaya et al. 2009). Other approaches identify valence shifters (negation, modality, and intensifiers) that strengthen, weaken, or reverse the prior polarity of a word or an expression (Polanyi and Zaenen 2006; Moilanen and Pulman 2007; Shaikh et al. 2007; Choi and Cardie 2008). The contextual polarity of individual expressions is then used for sentence as well as document classification (Kennedy and Inkpen 2006; Li et al. 2010).

1 For a survey on topic detection for aspect-based sentiment analysis, see Liu (2015), Chapter 6.

At the document level, the standard task is either a classification problem, categorizing documents globally as being positive, negative, or neutral towards a given topic (Pang et al. 2002; Turney 2002; Mullen and Nigel 2004; Blitzer et al. 2007), or regression, assigning a multi-scale rating to the document (Pang and Lee 2005; Goldberg and Zhu 2006; Snyder and Barzilay 2007; Lu et al. 2009; Lizhen et al. 2010; Leung et al. 2011; Moghaddam and Ester 2011; Ganu et al. 2013).

The classification/regression approach to determining document sentiment typically uses bag-of-words representations (BOW), which model each text as a vector of the number of occurrences, or the frequency with which each word/construction appears. Bag-of-words features also include *n*-grams, parts of speech, and features that account for the presence/absence of subjective words (including emoticons), or valence shifting words (e.g., negation, intensifiers). Bags of words disregard grammar and cannot easily move beyond the scope of the analyzed corpus. The approach is, however, quite simple to implement and does keep token frequency. Bags of words are attractive because, through feature reduction techniques, they reduce a large feature space of possibly hundreds of thousands of dimensions to a manageable size that can help classifiers boost their performance (Abbasi et al. 2008; Liang et al. 2015).

To account for word order, BOWs may include syntactic dependency relations that rely on the assumption that specific syntactic phrases are likely to be used to express opinions such as adjective–noun, subject–verb, or verb–object relationships (Dave et al. 2003; Gamon 2004; Matsumoto et al. 2005; Ng et al. 2006; Joshi and Penstein-Rosé 2009; Xia and Zong 2010), thus becoming, more precisely, bags of phrases. These insights on the types of structures that convey sentiment are the ones formalized in pattern grammar (see Section 2.5). Experiments show a mitigated success. First of all, some dependency relations are not frequent enough for the classifier to generalize well. Second, dependencies provide only grammatical relations between words within a sentence, but inter-sentential relations that may influence the sentence's sentiment polarity are not considered.

Bag-of-words approaches are popular in particular for long documents and very large corpora, where classification accuracy can exceed 80%. The effectiveness of supervised learning, however, depends on the availability of labeled (sometimes unbalanced) data in one domain that usually involves high costs in terms of work and time, because it involves collecting and labeling data. In addition, some languages lack such resources. To overcome this problem, unsupervised (Turney and Littman 2002; Feng et al. 2013) or semi-supervised learning have been proposed. Semi-supervised learning methods use a large amount of unlabeled data together with labeled data to build better classifiers (Li et al. 2011; Täckström and McDonald 2011; Raksha et al. 2015; Yang et al. 2015). Popular approaches include self-training (He and Zhou 2011), co-training (Wan 2009), and structural learning methods that learn good functional structures using unlabeled data (Ren et al. 2014).

A second problem with BOW approaches is that they are difficult to generalize, since the interpretation of evaluative expressions is often domain or genre dependent. For example, Aue and Gamon (2005) reported a loss in classifier accuracy (up to 20% to 30%) when training on movie review data and testing on book and product reviews. Most of the classifiers built using movie review data suffer from bias toward that genre, and they would not be able to capture the particular characteristics of other types of text, such as formal reviews or blog posts. An alternative could involve using labeled data from a source domain to classify large amounts of unlabeled data (documents/sentences) in a new target domain. According to Jiang and Zhai (2007) and Xia et al. (2015), there are two methods to perform domain adaptation: **instance**

adaptation, which approximates the target-domain distribution by assigning different weights to the source domain labeled data, and **labeling adaptation**. The last one is the most widely used in sentiment analysis (Blitzer et al. 2007; Pan et al. 2010; Samdani and Yih 2011; Bollegala et al. 2013; Cambria and Hussain 2015). It aims at learning a new labeling function for the target domain to account for those words that are assigned different polarity labels. For example, *long* may be positive in the phone domain (*long battery life*) and negative in the restaurant domain (*long wait times*). This means that this method treats knowledge from every source domain as a valuable contribution to the task on the target domain. However, in addition to data sampled from different distributions (in the source and target domains), label adaptation approaches may suffer from *negative transfer* where instead of improving performance, the transfer from other domains degrades the performance on the target domain. One possible solution is active learning, which relies on a small amount of good labeled data in the target domain to quickly reduce the difference between the two domains (Rai et al. 2010; Li et al. 2013b).

Compared with machine learning, lexicon-based methods make use of the linguistic information contained in the text that makes them more robust across different domains (Kennedy and Inkpen 2006; Taboada et al. 2011). Furthermore, Brooke et al. (2009) showed that porting dictionaries to a new language or a new domain is not an onerous task, probably less onerous than labeling data in a new domain for a classifier. Combining lexicon-based learning and corpus-based learning could be a good solution to incorporate both domain-specific and domain-independent knowledge. This has been investigated in several studies relying either on a complete lexicon or fully labeled corpus (Andreevskaia and Bergler 2008; Qiu et al. 2009), or a partially labeled corpus as training examples (Yang et al. 2015).

4. Towards a Dynamic Model of Evaluative Language

The model presented in Section 3, where $evaluation = (e, a, s, h, t)$, is an operationalized representation that views sentiment analysis as an information extraction task. The model builds a structured representation from any unstructured evaluative text that captures the core elements relative to the evaluations expressed in the text. Liu (2012) points out that not all applications need all five elements of the quintuple. In some cases, it is hard to distinguish between entity and aspect, or there is no need to deal with aspect. In other cases, opinion holder or time can be ignored.

Although this model covers the essential elements for dealing with evaluative language in real scenarios, we argue that it is rather static because linguistic and extra-linguistic contextual factors that directly impact one or several elements of the quintuple are hidden in the structured representation. We show in Section 4.1 that each element is context-dependent at different linguistic levels and that a new dynamic model that explicitly incorporates context is necessary. We think that such a dynamic model needs to combine two complementary perspectives: a *conceptual* one, allowing for a theoretical characterization of the problem from a linguistic point of view, and a *computational* one, allowing computer algorithms to easily operationalize the extraction process. To this end, we propose extending Liu's (2012) model to account for the semantic and pragmatic contribution that an evaluative sentence or a clause makes to a discourse in terms of a relation between an input context prior to the sentence and an output context. This new characterization of evaluative language falls within the dynamic semantics paradigm (Kamp and Reyle 1993) and will offer the basis for investigating, detecting, and formalizing various discursive and pragmatic aspects of evaluation.

In this section, we first motivate why a new model is needed, discuss some of the quintuple elements, and then present a new dynamic definition of evaluative language.

4.1 Motivation

Holder (h) and topic + aspect (e,a). Aspects may be explicit or implicit (Hu and Liu 2004a). Explicit aspects appear as nouns or noun phrases, as in *The characters are great*, which explicitly refers to the movie characters. Implicit aspects are generally inferred from the text through common sense or pragmatic knowledge, as in the restaurant review in Example (3), where the pronoun *it* does not refer to the potential antecedent *new vegan restaurant*, but to an implicit aspect of a restaurant, its food. This is a form of bridging inference (Haviland and Clark 1974).

(3) We went to the new vegan restaurant yesterday. It was all too raw and chewy for me.

Despite substantial progress in the field, implicit topic/aspect identification as well as complex explicit aspects (which use, for example, verbal constructions) remain unsolved. See Section 6.2 for a survey of the existing work that tackles this hard problem.

In addition to the explicit vs. implicit nature of aspects, the way holders, topics, and aspects are extracted heavily depends on the corpus genre. For example, review-style corpora typically discuss one main topic and its related aspects and are the viewpoints of one holder (the review writer). Other corpus genres, however, do not meet these characteristics. Some are *author-oriented* like blogs where all the documents (posts and comments) can be attributed to the blogs' owners. A blogger, for example, may compare two different products in the same post. Others are both *multi-topic* and *multi-holder* documents like news articles, where each pair (topic, holder) has its own evaluation, which means that different quintuples are needed to account for opinions about different topics (old or new) and from different holders. In addition, social media corpora are composed of *follow-up* evaluations, where topics are dynamic over conversation threads (i.e., not necessarily known in advance). For example, posts on a forum or tweets are often responses to earlier posts, and the lack of context makes it difficult for machines to figure out whether the post is in agreement or disagreement. Finally, in a *multi-topic* setting, the author introduces and elaborates on a main topic, switches to other related topics, or reverts back to an older topic. This is known as **discourse popping**, where a topic switch is signaled by the fact that the new information does not attach to the prior clause, but rather to an earlier one that dominates it (Asher and Lascarides 2003).

Investigating evaluations toward a given topic and how related topics influence the holder's evaluation on this main topic is an interesting and challenging research problem (He et al. 2013). Multi-topic sentiment analysis is generally seen as a special case of multi-aspect sentiment analysis (Hu and Liu 2004b; Zhao et al. 2010). With the rise of social media, tracking follow-up evaluations and how they change towards a topic over time has become very popular (Wang et al. 2012; Farzindar and Inkpen 2015). See also Section 6.4 on the contribution of social network structure to sentiment analysis.

Sentiment (s). Polarized evaluative expressions may be explicit or implicit. The former are triggered by specific subjective words or symbols (adjectives, adverbs, verbs, nouns, interjections, emoticons, etc.), whereas the latter, also known as fact-implied opinions (Liu 2012), are triggered by situations that describe a desirable or an undesirable activity or state. These situations are understood to be evaluative on the basis of pragmatic, cultural, or common knowledge shared by authors and readers. For example, there

are three evaluative expressions in the movie review in Example (4): The first two are positive explicit expressions (underlined), whereas the last one (in italics) is an implicit positive opinion. Not moving from one's seat is not inherently positive or negative; it only becomes positive in the context of watching a movie.

(4) What a great animated movie. I was so scared the whole time that I *didn't even move from my seat*.

Compared with explicit evaluation, little work has been done on implicit evaluation, mainly because it is difficult to discriminate between explicit, implicit, and objective statements, even for humans (see Section 6.2). Obviously, the treatment of explicit evaluation is not always easy, especially when complex linguistic devices are used. Consider, for example, the ironic statement in Example (5). The overall evaluation is negative even though there are no explicit negative words. In this construction, the author conveys a message that fails to make sense against the context. The excerpt in Example (6) illustrates another interesting phenomenon particularly common in reviews, named **thwarted expectations**, where the author sets up a deliberate contrast to the preceding discourse (Pang et al. 2002; Turney 2002). This example contains four opinions: The first three are strongly negative whereas the last one (introduced by the conjunction *but* in the last sentence) is positive. A bag of words approach would classify this review as negative. An aspect-based sentiment analysis system would probably do better, by classifying the last sentence as being positive towards the TV series and the first three as negative towards particular aspects of the series. This is, in part, the counter-expectation strategy discussed within Appraisal as a way of flagging invoked Appraisal (Martin and White 2005). It is obvious from these examples that the *s* part of the definition of evaluation is context dependent.

(5) I love when my phone turns the volume down automatically.

(6) The characters are unpleasant. The scenario is totally absurd. The decoration seems to be made of cardboard. But, all these elements make the charm of this TV series.

Polarity (o) and Intensity (i). Sentiment s is further defined by Liu (2012) as a triple of $y =$ type of sentiment; $o =$ orientation or polarity; and $i =$ intensity of the opinion (cf. Section 3.1). With respect to polarity, the prior polarity of a word, that is, its polarity in the dictionary sense, may be different from *contextual polarity*, which is determined on the basis of a sentiment composition process that captures how opinion expressions interact with each other and with specific linguistic operators such as intensifiers, negation, or modality (Polanyi and Zaenen 2006; Moilanen and Pulman 2007; Wilson et al. 2009). For instance, in *This restaurant is not good enough*, the prior positive orientation of the word *good* has to be combined with the negation *not* and the modifier *enough*. Apart from local linguistic operators, prior polarity may also vary according to the context outside of the utterance, including domain factors (Aue and Gamon 2005; Blitzer et al. 2007; Bollegala et al. 2011). A given span may be subjective in one context and objective in another. Haas and Versley (2015) observe that seemingly neutral adjectives can become polar when combined with aspects of a movie (*elaborate continuation*, *expanded vision*), as can words that are intensified (*simply intrusive* was considered negative, but *intrusive* was neutral). Even if there is any ambiguity on the subjectivity status of a word, orientation can be highly context-dependent: *A horrible movie* may be positive if it is a thriller, but negative in a romantic comedy. Additionally, out of context, some subjective expressions can have both positive and negative orientations. This is particularly salient

for expressions of surprise and astonishment as in *This movie surprised me*. The way evaluative language is used depends also on cultural or social differences, which makes polarity vary according to a specific situation, group, or culture. For instance, Thompson and Alba-Juez (2014) point out that utterances such as *Yeah* and *right*, depending on the situation, the kind of speaker, and the way in which they are spoken, may be intended as positive signs of agreement or as very negative disagreement. Liu (2015) also points out that there may be a difference between author and reader standpoints: A *small restaurant* does not convey a universal negative feeling.

Another interesting property of evaluative language is its multi-dimensional nature. Apart from the traditional binary categorization (positive vs. negative) of evaluation towards a given target, researchers have suggested that sentiment analysis is a *quantification* task, one where the goal is not to classify an individual text, but to estimate the percentage of texts that are positive or negative (Esuli and Sebastiani 2015). Evaluative language may also take several other forms. For example, evaluation involving comparatives expresses an ordering towards targets based on some of their shared aspects, for example, *the picture quality of camera X is better than that of Y* (Jindal and Liu 2006a, 2006b). There are also evaluations that concern relative judgment towards actions or intentions, preferring them or not over others (e.g., *I prefer the first season over the second one*, or *Shall we see Game of Thrones next week?*). In this last case, reasoning about preferences determines an order over outcomes that predicts how a rational agent will act (Cadilhac et al. 2012; Chen et al. 2013). Other forms of evaluative language involve finding a consensus or conflict over participants by identifying agreement/disagreement on a given topic in a debate or discussion (Somasundaran and Wiebe 2009; Mukherjee and Bhattacharyya 2012). We address preferences and intentions in Section 6.5.

4.2 A New Dynamic Model of Evaluative Language

We hope to have established by now that evaluative language is context-dependent at different discourse organization levels:

- The sentence: Interactions with linguistic operators like negation, modality, and intensifiers; or syntactic constraints such as altering the order of constituents in a clause or sentence.
- The document: Discourse connectives, discourse structure, rhetorical relations, topicality.
- Beyond the document:[2] Effects of various pragmatic phenomena such as common-sense knowledge, domain dependency, genre bias, cultural and social constraints, time constraints.

Inspired by Polanyi and Zaenen's (2006) first attempt to study contextual valence shifting phenomena, as well as recent linguistic studies on evaluation (Thompson and Alba-Juez 2014) that characterize it as a dynamic phenomenon, we propose a more flexible and abstract model of evaluative language that extends Liu's (2012) model to

2 We will use terms such as *document* and *text* throughout, but we believe that most of what we discuss here applies equally to spoken language, which has further special characteristics (prosody, pitch, gesture).

take into account context. The model is represented as a system of two main components $\langle \Omega, \Im \rangle$, where:

- $\Omega = (e, a, s, h)$ is a quadruple that corresponds to the *intrinsic properties* of evaluative language (target, aspect, sentiment, and holder). The sentiment element s is additionally composed of a quadruple (*span*, *category*, *pol*, *val*) to encode span (the evaluative expression), semantic category, polarity, and strength. Elements of Ω resemble those of Liu (2012), except that we restrict s to textual spans composed of explicit subjective tokens only (adjectives, verbs, nouns, adverbs, emoticons), excluding local operators at the sentence level. The aim is to explicitly separate the prior sentiment orientation of s (i.e., its value out of context) from its contextual interpretation that is considered to be part of the *extrinsic properties* of evaluation (see $\Im$ below). In addition, polarity (*pol*) is used as an umbrella term that covers any polarized scale in order to extend the traditional positive/negative categorization. Polarized scales may be positive–negative, for–against, positive–neutral–negative, or any number of stars in a star scale. If any of the four elements of Ω is not lexicalized, then their corresponding values are underspecified.[3]
- A set of functions $\Im = \{F_1, \ldots, F_n\}$ that capture the *extrinsic properties* of an evaluation by adjusting or adapting the prior values of Ω when they are interpreted in a given context, such that $\forall F_i \in \Im, F_i : \Omega \mapsto Update_l(\Omega)$. These functions act as complex update operators at different discourse organization levels l (sentence, document, beyond the document), and their interpretation reflects the special influence that they have on evaluative expressions. For example, at the sentence level, one can design a function to account for the role of negation or modality. At the document level, functions may use discourse relations or argumentative structure to update the prior polarity of an evaluation. At the extra-linguistic level, some functions can account for the conversation thread. We expressively avoid defining all the functions that may be included in $\Im$, so that users can specify them according to need and available NLP techniques. Not all functions are necessary for all applications, and some of them can be left out.

Let us illustrate this model with Example (4), reproduced again as Example (7). We explain how the model behaves when applied at the sentence level,[4] but it can also be easily applied at a finer or more-coarse grained level.

(7) What a great animated movie. I was so scared the whole time that I didn't even move from my seat.

In Example (7), the immediate interpretation of *great* in the first sentence leads to a positive evaluation that remains stable after updating the intrinsic property of polarity, i.e., $\Omega_1 = Update(\Omega_1) =$ (movie, _ , (*great*, _ ,+,1), author). In the second sentence, the sen-

3 Note that we do not include a time dimension in our definition for simplicity purposes. In cases where time is important, an element t may be added.

4 In this example, we deliberately leave the sentiment semantic category underspecified. For strength, we use a discrete scale.

timent span *scared* is generally given a negative evaluation out of context, Ω_2 = (movie, _ , (*scared*, _ ,−,1), author). At the sentence level, the evaluation in *scared* is influenced by the adverb *so*, which modifies the prior intensity of the evaluation: $Update_{sentence}(\Omega_2)$ = (movie, _ , (*so scared*, _ ,−,2), author). Then, if one takes into account the discursive context given by the first sentence, the prior polarity of *scared* has to be updated: $Update_{discourse}(\Omega_2)$ = (movie, _ , (*so scared*, _ ,+,2), author). Finally, the pragmatic level tells us about the presence of an additional positive evaluation, as already explained in the previous section: $Update_{pragmatic}(\Omega_2)$ = (movie, _ , {(*so scared*, _ ,+,2), (*didn't move...seat*, _ ,+,2)}, author).

In practical sentiment analysis systems, the complete instantiation of the model $\langle \Omega, \Im \rangle$ depends on their context-sensitiveness. As far as we know, no existing system deals simultaneously with all levels of context. Most systems are either static and context insensitive, or they exploit one or two levels at most. Some deal with local factors at the sentence level, for example, the presence of negation in the proximity of polarized words using dependency parsing; others deal with discourse; whereas others deal with domain and genre factors, or figurative language. Overall, two main directions have been explored to deal with context: (a) extend the bag-of-words model by incorporating more contextual information; and (b) combine theories and knowledge from linguistics with the statistical exploitation of data. In the next two sections, we explore these two directions, focusing on how sentiment analysis systems can be made much more effective and accurate by explicitly considering context in its wider sense. We investigate several dimensions of context, and for each dimension, we outline which linguistic aspects contribute to accurate extraction of sentiment.

5. Current Approaches: From Lexical Semantics to the Sentence

The sentiment analysis problem can be described as the process of moving from the bottom–up, starting at the word level and ending with context. In this section, we describe how sentiment is expressed and extracted at the word, phrase, and sentence levels, and in the next section we address discourse and contextual phenomena.

5.1 Lexical Semantics of Evaluative Expressions

Finding subjective words (and their associated prior polarity) is an active research topic where both corpus-based and lexicon-based approaches have been deployed. There are several manually or automatically created lexical resources. Most of them share four main characteristics: They are domain-and language-specific, of limited coverage, and they group evaluative expressions along a binary positive/negative axis.

Domain adaptation has been extensively studied in the literature (cf. Section 3.2 for a discussion). Language adaptation often consists of transferring knowledge from a resource-rich language such as English to a language with fewer resources, using parallel corpora or standard machine translation techniques (Mihalcea et al. 2007; Abbasi et al. 2008; Balahur et al. 2012; Balahur and Perea-Ortega 2015; Gao et al. 2015). Other approaches make few assumptions about available resources by using a holistic statistical model that discovers connections across languages (Boyd-Graber and Resnik 2010). Under the assumption that similar terms have similar emotional or subjective orientation (Hatzivassiloglou and McKeown 1997), lexicon expansion techniques grow an initial set of subjective seed words by diverse semantic similarity metrics. These techniques may also exploit word relationships such as synonyms, antonyms, and hypernyms

within general linguistic resources such as WordNet, or syntactic properties such as dependency relations.[5] Compared with lexicon expansion and domain and language adaptation, few studies propose to enhance the binary categorization of evaluative expressions. In the remainder of this section, we focus on these studies, as we believe they can be beneficial to subjective lexicon creation. Studies can be roughly divided into how they approach categorization of lexical expressions: by tackling intensity, emotion, and either syntactic or semantic principles for classification (including Appraisal categories).

5.1.1 Intensity-Based Categorizations. One way to improve upon simple polarity determination is to associate with each subjective entry an evaluation score or intensity level. SentiWordnet (Baccianella et al. 2010) is an extension of WordNet (Fellbaum 1998) that assigns to each synset three sentiment scores: positive (P), negative (N), and objective (O). Hence different senses of the same term may have different sentiment scores. For example, the adjective *happy* has four senses: The first two expressing joy or pleasure are highly positive ($P = 0.875$ and $P = 0.75$, respectively); the third corresponding to *eagerly disposed to act* or *to be of service* (as in *happy to help*) is ambiguous ($P = 0.5$ and $O = 0.5$); and the last sense (*a happy turn of phrase*) is likely to be objective ($P = 0.125$ and $O = 0.875$). Although SentiWordNet has been successfully deployed to derive document-level sentiment orientation (Denecke 2009; Martin-Wanton et al. 2010; Popat et al. 2013; Manoussos et al. 2014), word-sense disambiguation algorithms are often needed to find the right sense of a given term. Other intensity-based lexicons in English include Q-WordNet (Agerri and García-Serrano 2010), the MPQA Subjectivity Lexicon (Wiebe et al. 2005), and SO-CAL (Taboada et al. 2011).

Methods for automatic ordering of polar adjectives according to their intensity include pattern-based approaches, assuming one single intensity-scale for all adjectives (de Melo and Bansal 2013), or corpus-driven techniques, providing intensity levels to adjectives that bear the same semantic property (Ruppenhofer et al. 2014; Sharma et al. 2015).

5.1.2 Emotion and Affect Categorizations. A second way to enhance polarity consists of encoding, in addition to polarity and/or intensity, information about the semantic category of the evaluative expression. Categories can be defined according to psychologically based classifications of emotions and affect of various sorts that attempt to group evaluation into a set of basic emotions such as anger, fear, surprise, or love (Osgood et al. 1957; Izard 1971; Russell 1983; Ekman 1984; Ortony et al. 1988). Well-known Affect resources in English include the General Inquier (Stone et al. 1962), the Affective Norms for English Words (ANEW) (Bradley and Lang 1999), the LIWC Dictionary,[6] and the LEW list (Francisco et al. 2010). WordNet-Affect is also a resource for the lexical representation of affective knowledge (Strapparava and Valitutti 2004). It associates with each affective synset from WordNet an emotion class, following the classes defined within Ortony et al.'s (1988) model of emotions. Other interesting affective resources are SenticNet (Poria et al. 2013) and the EmotiNet knowledge base (Balahur et al. 2011), which associates polarity and affective information with affective situations such as *accomplishing a goal*, *failing an exam*, or *celebrating a special occasion*. Modeling such situations is particularly important for recognizing implicit emotions (Balahur et al.

5 See Liu (2015) Chapter 7 for an overview of existing techniques.
6 `http://liwc.wpengine.com/`.

2012). Besides the obvious usefulness of affect lexicons in emotion detection, their use in sentiment analysis tasks has shown to be helpful (Agarwal et al. 2009), especially in figurative language detection (Reyes and Rosso 2012) (see Section 6.3).

The choice of the right resources to use generally depends on the task and the corpus genre. Musto et al. (2014) compare the effectiveness of SentiWordNet, WordNet-Affect, MPQA, and SenticNet for sentiment classification of Twitter posts. Their results show that MPQA and SentiWordNet performed the best. This is interesting because although MPQA is a lexicon with small coverage, its results are comparable to a general-purpose lexicon like SentiWordNet. To overcome coverage limitation, one can consider combining several lexicons. This can, however, lead to polarity inconsistency, where the same word appears with different polarities in different dictionaries. Dragut et al. (2012) point out that sentiment dictionaries have two main problems: They exhibit substantial (intra-dictionary) inaccuracies, and have (inter-dictionary) inconsistencies. They propose a method to detect polarity assignment inconsistencies for the words and synsets within and across dictionaries.

5.1.3 Syntactic and Semantic Categorizations. Other categories have been proposed in the literature to deal with the complex lexical semantics of evaluative expressions. Some are both syntactically and semantically driven, whereas others are exclusively semantic. Levin (1993) classifies over 3,000 English verbs according to shared meaning and syntactic behavior. She examines verb behavior with respect to a wide range of syntactic alternations that reflect verb meaning. Several classes are sentiment relevant such as Judgment Verbs or Verbs of Psychological State. Mathieu (2005) offers a semantic classification of sentiment in which verbs and nouns are split into 38 semantic classes, according to their meaning (Love, Fear, Astonish, etc.). She points out that syntactic structure influences the interpretation of evaluative expressions and distinguishes between three classes of verbs that mean: The experience or the causation of a rather unpleasant feeling, pleasant feeling, and neither pleasant nor unpleasant. Semantic classes are linked by meaning, intensity, and antonymy relationships. She associates a set of linguistic properties with words and classes and builds semantic representations, described by means of feature structures. Mathieu and Fellbaum (2010) extended this classification to English verbs of emotion.

SentiFrameNet (Ruppenhofer and Rehbein 2012) is probably the best example of the syntactico-semantic categorization of evaluative expressions. It extends FrameNet (Baker et al. 1998) to connect opinion source and target to semantic roles, and to add semantic features such as polarity, intensity, and affectedness (changes of state that leave an event participant in a changed state). Evaluative language per se is not described in FrameNet, but several frames are relevant for sentiment analysis like JUDGMENT, OPINION, EMOTION_DIRECTED, and semantic roles such as JUDGE or EXPERIENCER. Each frame is associated with a set of lexical units composed of words that evoke that frame. Example (8) illustrates the frame elements of the lexical unit *splendid* (from the frame DESIRABILITY).

(8) [On clear days,]^{Circumstance} [the view]^{Evaluee} was [absolutely]^{Degree} *splendid*.

Ruppenhofer and Rehbein (2012) argue that a frame-based representation of evaluative language is suitable for capturing multi-word evaluative expressions and idioms such as *give away the store* and sentiment composition. However, apart from using semantic frames for identifying the topics (or targets) of sentiment (Kim and Hovy 2006) and deriving an intensity-based sentiment lexicon (Raksha et al. 2015), little work has been

done to show the real effectiveness of this deep representation in practical sentiment analysis systems.

Purely semantic categorizations include the categories defined within the MPQA project (Wiebe et al. 2005), and the Blogoscopy lexicon (Daille et al. 2011), which classifies opinions according to four main categories following Charaudeau (1992), and the lexicon model for subjectivity description of Dutch verbs proposed by Maks and Vossen (2012). Drawing from Levin (1993), Mathieu (2005), and Wierzbicka (1987), Asher et al. (2009) group each opinion expression into four main categories: REPORTING, which provides, at least indirectly, a judgment by the author on the opinion expressed; JUDGMENT, which contains normative evaluations of objects and actions; ADVICE, which describes an opinion on a course of action for the reader; and SENTIMENT-APPRECIATION, containing feelings and appreciations. Subcategories include, for example, INFORM, ASSERT, EVALUATION, FEAR, ASTONISHMENT, BLAME, and so forth. These categories have been successfully deployed as features for consensus detection in book reviews (Benamara et al. 2014) and for studying opinion in discourse (Benamara et al. 2016). Also, the ADVICE category, which is decomposed into SUGGEST, HOPE, and RECOMMEND, has been used for extracting customer suggestions from reviews (Negi and Buitelaar 2015). We further discuss discourse-based sentiment analysis and suggestion detection in Section 6.

Even though efforts have been made to move beyond positive/negative classification, the resulting lexicons are not widely used in the sentiment analysis community. We do believe that the semantic categorizations discussed in this section are necessary for a better understanding of evaluative language and hope that further studies will leverage such categorizations to enhance current bag-of-words approaches. One categorization that deserves separate treatment, because of its comprehensiveness, is Appraisal. After an introduction to the theory itself in Section 2.6, in the next section we discuss computational treatments of Appraisal.

5.1.4 Semantic Categorizations: Appraisal in Sentiment Analysis. An early surge of interest in Appraisal led to publications proposing how to automatically identify expressions that could be characterized as belonging to the three subtypes of Attitude (Affect, Judgment, and Appreciation). The potential gains are obvious: A positive or a negative expression is informative in itself, but even more so if we know whether it refers to personal feelings, opinions about others, or evaluations of objects. Taboada and Grieve (2004) first proposed a classification of adjectives as to whether they mostly conveyed one of the three main Attitude categories (Affect, Judgment, Appreciation). They then used this classification of adjectives to determine what type of Attitude was predominant in a text.

Whitelaw et al. (2005) investigated the use of Appraisal groups or phrases for sentiment analysis. They also classified adjectives according to three Attitude categories. Using a combination of manual methods and thesauri crawling, they built a list of 1,329 adjectives. They then used this information to extract Appraisal groups, in particular, adjective groups that may include intensifiers and negation (e.g., *very good*, *not terribly funny*). The resulting Appraisal groups are fed into a classifier to train a system that identifies movie reviews as positive or negative. Whitelaw et al. found that using Appraisal groups improved the classification over baselines that utilized bag-of-words features. It is clear from their results that using adjectives that can be classified as conveying appraisal values is beneficial in sentiment analysis. In particular, and unsurprisingly, adjectives labeled as Appreciation are some of the most useful features for the classifier. Further analysis showed that bag-of-words features contribute

positively to the classification in part because they contain sentiment words that are not adjectives.

Although it was a breakthrough in terms of using sentiment-specific adjectives, and using linguistic insights (phrases rather than isolated adjectives; intensification and negation), the work of Whitelaw and colleagues did not explicitly utilize the information contained in the Appraisal groups. That is, whether a text contained more Appreciation than Judgment, for instance, was not part of the classification or the information extracted from the text. The authors do point out potential benefits of Appraisal analysis, including identifying the Appraiser and Appraised (what elsewhere have been termed *source* and *target*). Similar work by Read and Carroll uses a corpus annotated for Appraisal (not only Attitude, but also Graduation and Engagement) to train a classifier that detects whether unseen words and phrases are instances of Appraisal and, when they are, their polarity (Read and Carroll 2012b, 2012a).

Work by Argamon, Bloom, and colleagues (Bloom et al. 2007; Argamon et al. 2009; Bloom and Argamon 2010) also focused on using Appraisal to build lexicons. Of note is the effort in Bloom et al. (2007) to extract adjectives according to Appraisal categories (the three top-level Attitude types) and present them as output (rather than just use them as input to build a lexicon). Furthermore, their system assigns a target to each adjective, specifying whether, for example, it refers to an actor, a character, or the plot or special effects of a movie. This linking of Appraisal expression and target is achieved via dependency parsing. Their method seems to perform well, as shown by a manual evaluation of the extracted expressions, and by its potential usefulness in creating rules for opinion patterns. In a follow-up paper, Bloom and Argamon (2010) present an automatic method for linking Appraisal expressions and patterns.

Such linking of Appraisal expressions and targets precedes later work in feature-based sentiment extraction or opinion mining (Titov and McDonald 2008; Brody and Elhadad 2010; Liu 2012), which seems to have developed independently, and without taking into account the possibility of bootstrapping the identification of features with Appraisal expressions. Information on whether an Appraisal expression is likely to be Judgement rather than Appreciation will help determine whether its target is human or not, and vice versa.

A notable exception in the practical application of Appraisal is the Attitude Analysis Model of Neviarouskaya and colleagues (Neviarouskaya 2010; Neviarouskaya et al. 2010b, 2010a; Neviarouskaya and Aono 2013). They propose a method of assigning what are essentially Attitude values to adjectives. In their system, a classifier is trained to determine whether a particular word or phrase expresses the three basic types of Attitude, with the determination being done in context, that is, taking into account the sentence in which the word appears.

In summary, we see Appraisal analysis as a richer, more detailed analysis that goes beyond simple polarity labels, and that can help characterize texts across several categories. Once more data and resources become available, automatic analysis of Appraisal is possible. This would enable the presentation of Appraisal information as part of the process of sentiment analysis. Just as some systems break down lengthy reviews and provide information on features or aspects (such as service, food or ambiance for a restaurant), a review can be further characterized according to Appraisal categories, specifying, for instance, whether it contains more Affect than Judgment, or an unusally high frequency of graduated terms (indicating a particularly strong opinion). Current research, however, does not seem to be making use of Appraisal, and it is unclear how significant the improvements may be for simple binary classification systems.

5.2 Valence Shifters

The term **valence shifters** was first used by Polanyi and Zaenen (2006) to describe how an evaluation expression can be modified by context, including extra-propositional aspects of meaning, intensification, downtoning, presuppositions, discourse, and irony. This section focuses on valence shifters that may impact evaluative expressions at the sentence or the sub-sentential level.

Dealing with valence shifters roughly involves three sub-tasks: identifying these expressions and their scope, analyzing their effect on evaluation, and computing sentiment composition, leveraging this effect to update the prior polarity of opinion expressions. The first task makes use of full sentence parsing or dependency parsing to identify the scope of cues (Councill et al. 2010; Wiegand et al. 2010; Velldall et al. 2012). We detail in the following sections the latter two tasks.

5.2.1 Effect of Valence Shifters on Sentiment Analysis

Intensification and downtoning. Whatever parts of speech are identified as conveying sentiment, they can be intensified and downtoned by being modified. The general term **intensifier** is used for devices that change the intensity of an individual word, whether by bringing it up or down. Many devices intensify; for instance, adjectives may intensify or downtone the noun they accompany (*a definite success*). Periphrastic expressions and hedges also change the intensity of other words, as is the case with *in a way* or *for the most part*.

Intensification, however, is mostly expressed via adverbs. Syntactically, adverbs may appear in different positions in a sentence. For example, they could occur as complements or modifiers of verbs (*he behaved badly*), modifiers of nouns (*only adults*), adjectives (*a very dangerous trip*), adverbs (*very nicely*), and clauses (*undoubtedly, he was right*). Adverbs of degree and manner have been the most studied, as they are most sentiment relevant. Taking them into account has consistently been shown to improve the performance of sentiment analysis systems (Kennedy and Inkpen 2006; Benamara et al. 2007; Taboada et al. 2011).

The effect of intensifiers and downtoners on evaluative language is generally modeled as a linear model using addition or subtraction. For example, if a positive adjective has a value of 2, an amplified (or positively intensified) adjective would become 3, and the downtoned version a 1. Intensifiers, however, do not all intensify at the same level. For instance, consider the difference between *extraordinarily* and *rather*. The value of the word being intensified also plays a role. A word at the higher end of the scale is probably intensified more intensely, as can be seen in the difference between *truly fantastic* and *truly okay*. In fact, the latter is probably often used ironically. A method of modeling these differences is to use multiplication rather than addition. For example, Taboada et al. (2011) place intensifiers on a percentage scale, proposing values such as the following: *most* +100%, *really* +25%, *very* +15%, *somewhat* −30%, and *arguably* −20%.

Extra-propositional aspects of meaning. These are aspects of meaning that convey information beyond the propositional content of a clause or sentence (i.e., beyond simple, categorical assertions), but are still within the realm of syntax, not discourse.

Nonveridicality is an example of such aspects (see Section 2.3). It can be used to express possibility, necessity, permission, obligation, or desire, and it is grammatically expressed via adverbial phrases (*perhaps, maybe, certainly*), conditional verb mood, some modals (*must, can, may*), and intensional verbs (*think, believe*). Adjectives and nouns

can also express modality (*a probable cause; It remains a possibility*). A general consensus in sentiment analysis is that nonveridicality and irrealis result in the unreliability of any expression of sentiment in the sentences containing it (Wilson et al. 2009; Taboada et al. 2011; Benamara et al. 2012; Morante and Sporleder 2012b; Denis et al. 2014). Consider the effect of the intensional verb *thought* and the modal *would* in Example (9) and the modal plus question in Example (10), which completely discount any positive evaluation that may be present in *good* and *more suitable*. In some cases, however, evaluative expressions under the scope of modality do not have to be ignored (Benamara et al. 2012). This is true for the cumulative modality in Example (11) and the use of negation in Example (12), where the deontic modal *should* strengthens the negative recommendation.

(9) I thought this movie would be as good as the Grinch.

(10) Couldn't you find a more suitable ending?

(11) You definitely must see this movie.

(12) You should not go see this movie.

Not enough research has explored exactly how evaluative expressions are affected in the presence of nonveridical operators. In de Marneffe et al. (2012), nonveridicality is characterized as a distribution over veridicality categories, rather than a binary classification, which would make accurate identification of nonveridical statements even more challenging. Liu et al. (2014) automatically determine whether opinions expressed in sentences with modality are positive, negative, or neutral.

Negation is another linguistic phenomenon that affects evaluative expressions locally. In Example (13), the negation *not* conveys a mild positive evaluation by negating a negative item (*bad*), or the opposite, using a negated positive to express a negative evaluation (cf. Example (14)). The effect of negation seems to be one of downtoning the overall effect of the evaluation, whether positive or negative. This is a form of **litotes**, the negation of the opposite meaning to the one intended, often for rhetorical effect.

(13) This student is not bad.

(14) It hasn't been my best day.

Negation can be used to deny or reject statements. It is grammatically expressed via a variety of forms: prefixes (*un-, il-*), suffixes (*-less*), content word negators such as *not*, and negative polarity items (NPIs) like *any, anything, ever*. NPIs are words or idioms that appear in negative sentences, but not in their affirmative counterparts, or in questions but not in assertions (which also makes them nonveridical markers). Negation can be expressed using nouns or verbs that have negation as part of their lexical semantics (*abate* or *eliminate*). It can also be expressed implicitly without using any negative words, as in *This restaurant was below my expectations*. Negation can aggregate in a variety of ways. In some languages, multiple negatives cancel the effect of negation (*This restaurant never fails to disappoint on flavor*), whereas in negative-concord languages like French, multiple negations usually intensify the effect of negation. Compared to negators and content word negators, NPIs and multiple negatives have received less attention in the sentiment analysis literature. Taboada et al. (2011) treat NPIs (as well as modalities) as **irrealis blockers** by ignoring the semantic orientation of sentiment words in their scope. For example, the adjective *good* will just be ignored in *Any good movie in this theater*. In contrast, Benamara et al. (2012) consider that NPIs strengthen expressions under their scope. They observe that most multiple negatives preserve polarity, except for those

composed of content word negators and NPIs that cancel the effect of lexical negations. For example in the French expression *manque de goût* ('lack of taste'), the polarity is negative, while in *ne manque pas de goût* (roughly, 'no lack of taste'), the opinion is positive.

Another notable aspect of negation is its markedness. Negative statements tend to be perceived as more marked than their affirmative counterparts, both pragmatically and psychologically (Osgood and Richards 1973; Horn 1989). Potts (2010) posits an *emergent expressivity* for negation and negative polarity, observing that negative statements are less frequent and pragmatically more negative, with emphatic and attenuating polarity items modulating such negativity in a systematic way. Research in sentiment analysis has found that accurately identifying negative sentiment is more difficult, perhaps because we use fewer negative terms and because negative evaluation is couched in positive terms (Pang and Lee 2008, Chapter 3). One way to solve this problem is to, in a sense, follow the Negativity Bias (Rozin and Royzman 2001; Jing-Schmidt 2007): If a negative word appears, then it has more impact. This has been achieved by weighing negative words more heavily than positives in aggregation (Taboada et al. 2011).

Most approaches treat negation as polarity reversal (Wilson et al. 2005; Polanyi and Zaenen 2006; Moilanen and Pulman 2007; Choi and Cardie 2008). However, negation cannot be reduced to reversing polarity. For example, if we assume that the score of the adjective *excellent* is $+3$, then the opinion score in *This student is not excellent* cannot be -3. The sentence probably means that the student is not good enough. It is thus difficult to negate a strongly positive word without implying that a less positive one is to some extent possible (*not excellent, but not horrible either*). A possible solution is to use shift negation, in which the effect of a negator is to shift the negated term in the scale by a certain amount, but without making it the polar opposite of the original term (Liu and Seneff 2009; Taboada et al. 2011; Chardon et al. 2013a).

5.2.2 Sentiment Composition. Sentiment composition aims at computing the sentiment orientation of an expression or a sentence (in terms of polarity and/or strength) on the basis of the sentiment orientation of its constituents. This process, based on the *principle of compositionality* (Dowty et al. 1981), captures how opinion expressions interact with each other and with specific linguistic operators such as intensifiers, negations, or modalities. For instance, the sentiment expressed in the sentence *This restaurant is good but expensive* is a combination of the prior sentiment orientation of the words *good*, *but*, and *expensive*.

A prior step to this process is to perform syntactic parsing to determine the scope of valence shifters and then use syntax to do compositional sentiment analysis. Jia et al. (2009) propose a set of complex heuristic rules to determine the scope of negation and then study the impact of different scope models for sentence and document polarity analysis. Their results show that incorporating linguistic insights into negation modeling is meaningful. Another way to model composition is to rely on a sentiment lexicon and predefined set of heuristics that predicts how shifters affect the sentiment of a phrase/sentence (Moilanen and Pulman 2007; Choi and Cardie 2008; Nakagawa et al. 2010; Taboada et al. 2011; Benamara et al. 2012). For example, Moilanen and Pulman (2007) propose three types of rules to deal with negation and intensifiers: sentiment propagation (the polarity of a neutral constituent is overridden by that of the evaluative constituent), polarity conflict resolution (a non-neutral polarity value is changed to another non-neutral polarity value), and polarity reversal. Lexicon-based sentiment composition has been shown to outperform bag-of-words learning classification of sentiment at the sentence level (Choi and Cardie 2008).

Rules being language-and context-dependent, another alternative approach is to represent each node in a parse tree with a vector, and then learn how to compose leaf vectors in a bottom–up fashion. The composition process is modeled as a function learned from the training data, which can be standard data sets that only have document-level sentiment annotation (e.g., star ratings [Yessenalina and Cardie 2011; Socher et al. 2011, 2012] or sentiment treebanks with fine-grained annotations for every single node of the top parse tree (Johansson and Moschitti 2013; Socher et al. 2013; Dong et al. 2014; Hall et al. 2014; Zhu et al. 2015). The Stanford Sentiment Treebank is probably the best known example (Socher et al. 2013). It is composed of 215,154 phrases annotated by three human judges according to six sentiment values, ranging from very negative to very positive, with neutral in the middle. Those phrase annotations are then combined with a dependency parse to propagate sentiment up through the nodes of the tree.

Various composition functions have been proposed in the literature. For example, Yessenalina and Cardie (2011) represent each word as a matrix and combine words using iterated matrix multiplication, which allows for modeling both additive (for negation) and multiplicative (for intensifiers) semantic effects. Wu et al. (2011) propose a graph-based method for computing a sentence-level sentiment representation. The vertices of the graph are the opinion targets, opinion expressions, and modifiers of opinion; the edges represent relations among them (mainly, opinion restriction and opinion expansion). However, using recursive neural tensor networks trained over the Stanford Sentiment Treebank yields significant improvements over approaches based on token-level features. Haas and Versley (2015) observe that this gain in accuracy comes at the cost of the huge effort needed to built such treebanks, which are necessarily limited to certain domains and languages. To overcome this difficulty, they propose an alternative cross-lingual and cross-domain approach.

6. New Approaches: From Discourse to Pragmatic Phenomena

Valence shifters capture important linguistic behavior. Relying on the principle of compositionality, researchers take for granted that the sentiment of a document, a sentence, or a tweet is the sum of its parts. Some of the parts contribute more than others and some reduce or cancel out the sentiment, but the assumption is often that components can be added up, subtracted, or multiplied to yield a reliable result. Shifting phenomena discussed in the last section cannot account for the phenomena of contextual valence shifting in general because they are necessarily bounded at the sentence level and contextual sentiment assignment occurs at the discourse and the pragmatic levels.

In this section, we discuss five contextual phenomena that we believe constitute the keys to the future of sentiment analysis systems: discourse, implicit evaluation, figurative language, extra-linguistic information, and intent detection.

6.1 Discourse-Level Phenomena

Texts and conversations are not mere juxtapositions of words and sentences. They are, rather, organized in a structure in which discourse units are related to each other so as to ensure both discourse coherence and cohesion. Coherence refers to the logical structure of discourse where every part of a text has a function, a role to play, with respect to other parts in the text (Taboada and Mann 2006b). Coherence has to do with semantic or pragmatic relations among units to produce the overall meaning of a discourse (Hobbs 1979; Mann and Thompson 1988; Grosz et al. 1995). The impression of coherence in text (that it is organized, that it hangs together) is also aided by cohesion, the linking

of entities in discourse (Halliday and Hasan 1976). Linking across entities happens through grammatical and lexical connections such as anaphoric expressions and lexical relations (synonymy, meronymy, hyponymy) appearing across sentences.

In sentiment analysis, discourse structure provides a crucial link between the local sentence level and the entire document (article, conversation, blog post, tweet, headline) and is needed for a better understanding of the opinions expressed in text. In particular, discourse can help in three main tasks: (1) identifying the subjectivity and polarity orientation of evaluative expressions; (2) furnishing important clues for recognizing implicit opinions; and (3) assessing the overall stance of texts. Two main directions have been explored to deal with discourse phenomena: top–down and bottom–up.

6.1.1 Top–Down Approaches. Top–down discourse analysis captures the macro-organization of a text or high-level textual patterns, following previous studies dealing with the signalling of text organization (Chafe 1994; Fries 1995; Goutsos 1996), the linearization problem (Levelt 1981), and the multi-dimensionality of discourse (Halliday and Hasan 1976). Top–down approaches define discourse segments as being units higher than the sentence (e.g., paragraph, topic units) and focus on building either a topic structure or a functional structure (see Purver 2011 and Stede 2011 for comprehensive surveys). Organizing discourse by topicality consists of splitting discourse into a linear sequence of segments, each of which focuses on a distinct subtopic occurring in the context of one or more main topics. Topic segmentation is generally guided by local discourse continuity or the lexical cohesion assumption (Halliday and Hasan 1976) that stipulates that topic and lexical usage (such as word repetition, discourse connectives, and paradigmatic relations) are strongly related (Hearst 1994). Functional structure, on the other hand, analyzes discourse from the point of view of the communicative roles or intentions of discourse units in genre-specific texts or from the speaker's (or writer's) communicative intention perspective (Grosz and Sidner 1986; Moore and Paris 1993; Lochbaum 1998). Genre-induced text structure aims at segmenting discourse into different parts that serve different functions. This segmentation is achieved through a conventionalized set of building blocks that contribute to the overall text function. These building blocks are called content, functional, or discourse zones. Discourse zones are specific to particular genres, such as the communicative roles played by the introduction, background, and conclusion sections in a scientific paper (Swales 1990; Teufel and Moens 2002). Other genres studied include law texts (Palau and Moens 2009), biomedical articles (Agarwal and Yu 2009), and movie reviews (Bieler et al. 2007).

Top–down approaches in sentiment analysis assume that in a subjective document only some parts are relevant to the overall sentiment. Irrelevant parts thus have to be filtered out or de-emphasized, and the remaining parts are used to infer an overall evaluation at the document level. For example, a recent psycholinguistic and psychological study shows that polarity classification should concentrate on messages in the final position of the text (Becker and Aharonson 2010). Pang et al. (2002) were the first to empirically investigate positional features. Specifically, depending on the position at which a token appears (first quarter, last quarter, or middle half of the document), the same unigram is treated as different features. However, the outcomes did not result in a significant improvement. An error analysis showed that this low improvement is due to the positional features that fail to adequately handle the "thwarted expectation" phenomenon, very common in such text genre (cf. Section 4). Pang et al. (2002) argued that a more sophisticated form of discourse analysis is needed. Taboada and Grieve (2004) also used positional features, but focused on adjectives and found that adjectives

at the beginning of the text are not as relevant, and that an opinion on the main topic tends to be found towards the end of the text.

Instead of keeping relevant segments on the basis of their relative position in a document, other studies suggest selecting them according to topic criteria. Inspired by Wiebe's (1994) assumption that objectivity and subjectivity are usually consistent between adjacent sentences, Pang and Lee (2004) built an initial model to classify each sentence as being subjective or objective and then used the top subjective sentences as input for a standard document level polarity classifier. The proposed cascade model was shown to be more accurate than one using the whole document. To reduce errors relative to models trained in isolation, McDonald et al. (2007), Mao and Lebanon (2007), Yessenalina et al. (2010), Täckström and McDonald (2011), Paltoglou and Thelwall (2013), and Yogatama and Smith (2014) use a joint structured model for both sentence- and document-level that captures sentiment dependencies between adjacent sentences. These models outperform a plain bag-of-words representation for document polarity classification.

Another line of research explores the functional role played by some parts or zones in a text. Smith and Lee (2014) focused on two discourse function roles—expressive and persuasive—and showed that training a supervised polarity classifier on persuasive documents can have a negative effect when testing on expressive documents. Based on a corpus study of German film reviews, Bieler et al. (2007) implemented a hybrid algorithm that automatically divided a document into its formal and functional constituents, with the former being constituents whose presence is characteristic for the genre (factual information such as date of the review, name of the author, or cast of the film), and the latter being longer paragraphs making contributions to the communicative goal of the author. Functional constituents are further divided roughly into Description and Comment, the former containing background information about the film or a description of the plot, and the latter contains the main evaluative content of the text, the part from which sentiment should be extracted. Using 5-gram SVM classifiers, the authors report a precision ranging from 70% for formal zones to 79% for functional zones. Later, Taboada et al. (2009) extended this approach and used the output of a paragraph classifier to weigh paragraphs in a sentiment analysis system. Results show that weighing Comment paragraphs higher than Description paragraphs boosts the accuracy of classifying reviews as either positive or negative from 65% to 79%. Roberto et al. (2015) follow the same idea for hotel reviews.

Argumentation is another top–down aspect that can play a functional role in evaluative texts. In fact, evaluative language may also serve a role in building arguments. Hunston and Thompson (2000a) suggest that evaluation helps organize the discourse, in addition to its role of strictly conveying an opinion (see Section 2.5). Argumentation is a process by which arguments are constructed by the writer to clarify or defend their opinions. An argument is generally defined as a set of premises that provide the evidence or the reasons for or against a **conclusion**, also known as a claim (Walton 2009). When using arguments, holders are able to tell not just *what* views are being expressed, but also *why* those particular views are held (Lawrence and Reed 2015).

Premises can be introduced in texts by specific markers or cue phrases such as *for example, but,* or *because* (Knott and Dale 1994). The claim is a proposition stating the general feeling or recommendation of the writer. It can be supported or attacked through various statements making a holder reveal preferences and priorities. For instance, in *The movie is good because the characters were great*, the second clause is evidence that supports the conclusion in the first clause, whereas in *The movie is good but the script was bad*, the same conclusion is attacked. Tracking arguments in text consists of identifying

its argumentative structure, including the premises, conclusion, and the connections between them such as the argument and counter-argument relationships.

Argumentation mining is a relatively new area in NLP (Mochales and Moens 2011; Peldszus and Stede 2013; Hasan and Ng 2014). Roughly, three main approaches have been proposed to extract arguments (Lawrence and Reed 2015). The first one relies on a list of discourse markers connecting adjacent premises split into two groups according to their effect on argumentation: support markers and attack markers. For example, adversative connectives such as *but* and *although* connect opposing arguments, whereas conjunctives like *and*, *or*, and *then* link independent arguments that target the same goal. The second approach in argumentation uses supervised learning to classify a given statement as being an argument or not. Then each argument can be classified as being either a premise or a conclusion, or whether it fits within a predefined argumentative scheme that takes the form of a number of premises working together to support or attack a conclusion. Finally, the last approach makes use of topic changes as indicators of a change in the argumentation line. For instance, if the topic of a given proposition is similar to the one discussed in the preceding propositions, then one can assume that these propositions are connected and are following the same line of reasoning.

In sentiment analysis, the role of argumentation has been investigated for document polarity classification (Hogenboom et al. 2010; Vincent and Winterstein 2014; Wachsmuth et al. 2014). For example, Vincent and Winterstein (2014) noticed in their corpus of movie reviews that a persuasive argumentation in positive reviews often consists of several independent arguments in favor of its conclusion. In negative reviews, however, a single negative argument appears to be enough. To date, most existing studies use a predefined list of markers to extract arguments and build the document's argumentative structure. Although the approach is rather simple (only a few arguments are marked), it has shown to improve document classification.

6.1.2 Bottom–Up Approaches. Bottom–up parsing defines hierarchical structures by constructing complex discourse units from Elementary Discourse Units (EDUs) in a recursive fashion. It aims at identifying the rhetorical relations holding between EDUs, which are mainly non-overlapping clauses, and also between larger units recursively built up from EDUs and the relations connecting them.[7] Identifying rhetorical relations is a crucial step in discourse analysis. Given two discourse units that are deemed to be related, this step labels the attachment between the two units with relations such as ELABORATION, EXPLANATION, or CONDITIONAL, as in [*This is the best book*]$_1$ [*that I have read in a long time*]$_2$, where the second argument expands or elaborates on the first. Some relations are explicitly marked, that is, they contain overt markers to clearly signal the type of connection between the arguments, such as *but, although, as a consequence*. Others are implicit, that is, they do not have clear indicators, as in *I didn't go to the beach. It was raining*. In this last example, in order to infer the EXPLANATION relation (or, more generally, a causal relation) between the clauses, we need detailed lexical knowledge and probably domain knowledge as well (that beaches are usually avoided when it is raining).

The study of discourse relations in language can be broadly characterized as falling under two main approaches: the lexically grounded approach and an approach that aims at complete discourse coverage. Perhaps the best example of the first approach is

7 For an introduction to rhetorical, coherence, or discourse relations, see Asher and Lascarides (2003) and Taboada and Mann (2006b).

the Penn Discourse Treebank (Prasad et al. 2008). The annotation starts that specific lexical items that signal the relation explicitly, most of them conjunctions, and includes two arguments for each conjunction. This leads to partial discourse coverage: There is no guarantee that the entire text is annotated, because parts of the text not related through a conjunction would be excluded. Complete discourse coverage requires annotation of the entire text, with most of the propositions in the text integrated in a structure. It includes work from two theoretical perspectives, either intentionally driven, such as Rhetorical Structure Theory (RST; Mann and Thompson 1988), or semantically driven, such as Segmented Discourse Representation Theory (SDRT; Asher and Lascaride 2003). RST proposes a tree-based representation, with relations between adjacent segments, and emphasizes a differential status for discourse components (the nucleus vs. satellite distinction). Captured in a graph-based representation with long-distance attachments, SDRT proposes relations between abstract objects using a relatively small set of relations.

Manually annotated resources following the aforementioned approaches have contributed to a number of applications, most notably discourse segmentation into elementary discourse units, identification of explicit and implicit relations for the purpose of discourse parsing, and development of end-to-end discourse parsers (Hernault et al. 2010; Feng and Hirst 2014; Joty et al. 2015; Surdeanu et al. 2015).

Efforts to incorporate discourse information into sentiment analysis can be grouped into two categories: those that rely on local discourse relations at the inter-sentential or intra-sentential level, and those that rely on the structure over the entire document, as given by a discourse parser or manually annotated data. We discuss each of these approaches in turn.

Leveraging local discourse relations. The idea is that among the set of relations, only some are sentiment-relevant. The simplest way to identify them is to take into account discourse connectives or cue phrases. Polanyi and Zaenen (2006) first noticed that connectives can reverse polarity, and they can also conceal subjectivity. For example, the CONCESSION relation in *Although Boris is brilliant at math, he is a horrible teacher* shows that the positivity of *brilliant* is neutralized, downtoned at best. A CONDITION relation will also limit the extent of a positive evaluation, as observed by Trnavac and Taboada (2012) in a corpus study of Appraisal in movie and book reviews. For instance, in *It is an interesting book if you can look at it without expecting the Grisham "law and order" style*, the positive evaluation in *interesting* is tempered by the condition that readers have to be able to change their expectations about the author's typical style and previous books. Narayanan et al. (2009) also focused on conditionals marked by connectives such as *if, unless,* and *even if*, and proposed a supervised learning algorithm to determine if sentiment expressed on different topics in a conditional sentence is positive, negative, or neutral. Instead of extracting specific connectives, some researchers use a compiled list of connectives and incorporate it as features in a bag-of-words model to improve sentiment classification accuracy (Mittal et al. 2013; Trivedi and Eisenstein 2013). Others identify discourse connectives automatically, relying on a discourse tagger trained on the Penn Discourse Treebank (Yang and Cardie 2014).

Relations that have been used in sentiment analysis are either relations proposed under various theories of discourse (e.g., RST, SDRT), or a set of relations built specifically to be used in sentiment analysis. Asher et al. (2008) considered five types of SDRT-like rhetorical relations, both explicit and implicit (CONTRAST, CORRECTION, RESULT, CONTINUATION, and SUPPORT), and conducted a manual study in which they represented opinions in text as shallow semantic feature structures. These are combined into

an overall opinion using hand-written rules based on manually annotated discourse relations. Benamara et al. (2016) extended this study by assessing the impact of 17 relations on both subjectivity and polarity analysis in movie reviews in French and in English, as well as letters to the editor in French. Zhou et al. (2011) focused on five RST relations (CONTRAST, CONDITION, CONTINUATION, CAUSE, and PURPOSE). Instead of relying on cue phrases, they proposed an unsupervised method for discovering these relations and eliminating polarity ambiguities at the sentence level. Zirn et al. (2011) grouped RST relations into Contrast vs. Non-Contrast and integrated them as features in a Markov Logic Network to encode information between neighboring segments. Somasundaran et al. (2009) proposed the notion of opinion frames as a representation of documents at the discourse level in order to improve sentence-based polarity classification and to recognize the overall stance. Two sets of relations were used: relations between targets (SAME and ALTERNATIVE) and relations between opinion expressions (REINFORCING and NON-REINFORCING). Lazaridou et al. (2013) also use a specific scheme of discourse relations. However, rather than relying on gold discourse annotations, they jointly predict sentiment, aspect, and discourse relations and show that the model improves accuracy of both aspect and sentiment polarity at the sub-sentential level.

Leveraging overall discourse structure. Coping with discourse relations at the local level has three main disadvantages: (1) the local approach captures explicitly marked relations—indeed, most approaches do not handle cases where a signal can trigger different relations or does not have a discourse use; (2) it accounts for the phenomena of contextual valence shifting only at the sentence level; and finally, (3) it does not account for long-distance discourse dependency. Polanyi and van den Berg (2011) argue that sentiment is a semantic scope phenomenon. In particular, discourse syntax encodes semantic scope; and because sentiment is a semantic phenomenon, its scope is governed by the discourse structure. Another important feature for studying the effects of discourse structure on opinion analysis is long-distance dependency. For instance, if an opinion is within the scope of an attribution that spans several EDUs, then knowing the scope of the attribution will enable us to determine who is in fact expressing the opinion. Similarly, if there is a contrast that has scope over several EDUs in its left argument, this can be important to determine the overall contribution of the opinions expressed in the arguments of the contrast. Example (15) illustrates this where complex segment [1–5] contrasts with segment [6–7].

(15) [I saw this movie on opening day.]$_1$ [Went in with mixed feelings,]$_2$ [hoping it would be good,]$_3$ [expecting a big let down]$_4$ [(such as clash of the titans (2011), watchmen etc.).]$_5$ [This movie was shockingly unique however.]$_6$ [Visuals, and characters were excellent.]$_7$

The importance of discourse structure in sentiment analysis has been empirically validated by Chardon et al. (2013b). Relying on manually annotated structures following SDRT principles, they proposed three strategies to compute the overall opinion score for a document: bag-of-segments that does not take into account the discourse structure; partial discourse that takes into account only relevant segments; and full discourse, which is based on the full use of a discourse graph, where a rule-based approach guided by the semantics of rhetorical relations aggregates segments opinion scores in a bottom–up fashion. These strategies were compared with a baseline that consisted of aggregating the strengths of opinion words within a review with respect to a given polarity and then assigning an overall rating to reflect the dominant polarity. The strategies were evaluated on 151 French movie reviews and 112 French news reactions annotated

with both opinion and discourse. Results showed that discourse-based strategies lead to significant improvements of around 10% over the baseline on both corpora. The added value of the discourse models is more impressive for newspaper comments than for movie reviews. This is probably because implicit evaluations (more frequent in comments) are well captured by the discourse graph. Another interesting result suggests that the use of full discourse is more salient for overall scale rating than for polarity rating. Wang et al. (2012) also relied on manual RST annotations in Chinese. They, however, only focus on relations triggered by explicit connectives using a strategy that weighs nuclei and satellites differently.

Extracting evaluative expressions in real scenarios requires automatic discourse representations. Voll and Taboada (2007) explored the integration of Spade (Soricut and Marcu 2003), a sentence-level RST discourse parser, into their system SO-CAL for automatic sentiment analysis. Their approach ignores adjectives outside the top nuclei sentences. The results obtained are comparable to those for the baseline that averages over all adjectives in a review. The authors argue that the loss in performance is mainly due to the parser having approximately only 80% accuracy in assigning discourse structure. Later work by Taboada et al. (2008) uses the same parser with a different approach: They first extract rhetorical structure from the texts, assign parts of the text to nucleus or satellite status, and then perform semantic orientation calculations only on the nuclei, namely, the most important parts. The authors evaluate the performance of this approach against a topic classifier that extracts topic sentences from texts. Results show that the use of weights on relevant sentences results in an improvement over word-based methods that consider the entire text equally. However, the weighing approach used only nuclei, regardless of the type of relation between nucleus and satellite. For example, a contrasting span may play a different role in conveying the overall sentiment than an elaboration on information in the nucleus does. Heerschop et al. (2011) further develop this finding and show that exploiting sentence-level RST relation types outperforms the baseline, with a sentiment classification accuracy increase of 4.5%.

The weighing scheme has also been used on document level parsing following RST (Gerani et al. 2014; Bhatia et al. 2015; Hogenboom et al. 2015). For example, Bhatia et al. (2015) propose two ways of combining RST trees with sentiment analysis: reweighing the contribution of each discourse unit based on its position in the tree, and recursively propagating sentiment up through the tree. Compared with a standard bag-of-words approach, the reweighing method substantially improves lexicon-based sentiment analysis, but the improvements for the classification-based models are poor (less that 1%). The recursive approach, on the other hand, results in a 3% accuracy increase on a large corpus of 50,000 movie reviews. Adding sensitivity to discourse relations (Contrastive vs. Non-Contrastive relations) offers further improvements.

6.2 Implicit Evaluation

Compared with explicit evaluation, implicit evaluation requires readers to make pragmatic inferences that go beyond what is literally said. Although humans are much better than automatic systems at figuring out implicit evaluation, much of it is difficult even for humans. For example, Toprak et al. (2010) reported a kappa of 0.56 for polar fact sentences in customer reviews, and Benamara et al. (2016) obtained 0.48 when annotating implicit opinions in French news reactions. In addition, when analyzing the Pearson's correlations between annotators' overall opinion score of a document and the scores given to subjective segments, Benamara et al. (2016) showed that implicit opinions are

better correlated with the global opinion score when negative opinions are concerned. This could indicate a tendency to "conceal" negative opinions as seemingly objective statements, which can be related to social conventions (politeness, in particular).

Grice (1975) made a clear distinction between what is said by an utterance (i.e., meaning out of context) and what is implied or meant by an utterance (i.e., meaning in context). In his theory of conversational implicature, Grice considers that to capture the speaker's meaning, the hearer needs to rely on the meaning of the sentence uttered, contextual assumptions, and the Cooperative Principle, which speakers are expected to observe. The **Cooperative Principle** states that speakers make contributions to the conversation that are cooperative. The Cooperative Principle is expressed in four maxims that the communication participants are supposed to follow. The maxims ask the speaker to say what they believe to be the truth (Quality), to be as informative as possible (Quantity), to say the utterance at the appropriate point in the interaction (Relevance), and in the appropriate manner (Manner). Maxims are, in a sense, ideals, and Grice provided examples of violations of maxims for various reasons. The violation of a maxim may result in the speaker conveying, in addition to the literal meaning of the utterance, an additional meaning that does not contribute to the truth-conditional content of the utterance, which leads to conversational implicature. Implicatures are thus inferences that can defeat literal and compositional meaning. Example (16) is a typical example of relevance violation: B conveys to A that he will not be accepting A's invitation for dinner although he has not literally said so.

(16) A. Let's have dinner tonight.
B. I have to finish my homework.

Borrowing from Grice's conversational implicature, Wilson and Wiebe (2005) view implicit evaluation as **opinion implicatures**, which are "the default inferences that may not go through in context." Hence, subjectivity is part of *what is said* while private-state inferences is part of *what is implied*. Example (17), taken from the MPQA corpus (Wiebe et al. 2005), illustrates the inter-dependencies among explicit and implicit sentiment. The explicit sentiment *happy* clearly indicates a positive sentiment but, at the same time, a negative sentiment toward Chavez himself may be inferred (somebody's fall is a negative thing; being happy about it implies that they deserved it, or that they are not worthy of sympathy).

(17) I think people are happy because Chavez has fallen.

Implicit evaluation is sometimes conveyed through an elusive process of **discourse prosody**, the positive or negative character that a few explicit items can infuse a text with. Martin and White (2005, page 21) define (one form of) prosody as meanings that are realized locally, but that color a longer stretch of text by dominating meanings in their domain. For instance, a review that starts out negatively, or that we see has a low number of stars associated with it, will lead us to interpret many of the meanings in the review as negative, even if negative opinion is not always explicitly stated. Bednarek (2006) also discusses this phenomenon in news discourse, characterizing it as **evaluative prosody**. Discourse prosody is also related to the sentence proximity assumption of Wiebe's (1994), whereby subjective or objective sentences are assumed to cluster together (see Section 3.2).

In general terms, there are three ways to make an evaluation implicit or invoked. The first one is to describe desirable or undesirable situations (states or events). Wilson (2008) refers to these as polar facts, that is, they are facts (as opposed to opinions), but they convey polarity because of the way such states or events are conventionally

associated with a positive or negative evaluation. Van de Kauter et al. (2015) state that the polarity of such facts is inferred using common sense, world knowledge, or context. Situations can be conveyed through verb phrases like those in italics in Examples (18) and (19), or noun phrases like the word *valley* in Example (20). The first two examples are translations from the French CASOAR corpus (Benamara et al. 2016); Example (20) comes from Zhang and Liu (2011).

(18) The movie is not bad, *although some persons left the auditorium.*

(19) This movie is poignant, and the actors excellent. *It will remain in your DVD closet.*

(20) Within a month, *a valley formed in the middle of the mattress*.

Situations that affect the evaluation of entities can be automatically identified relying either on co-occurrence assumptions, a set of rules, or patterns enlarged via bootstrapping (Goyal et al. 2010; Benamara et al. 2011; Zhang and Liu 2011; Riloff et al. 2013; Deng et al. 2014b; Wiebe and Deng 2014). For example, Riloff et al. (2013) learn from tweets patterns of the form $[VP_+]$.$[Situation_-]$ that correspond to a contrast between positive sentiment in a verb phrase and a negative situation. Tweets following this pattern are more likely to be sarcastic (see Section 6.3 on sarcasm and figurative language). Zhang and Liu (2011) exploit context to identify nouns and noun phrases that imply sentiment. They hypothesize that such phrases often have a single polarity (either positive or negative, but not both) and tend to co-occur in an explicit negative (or positive) context. In addition to co-occurrence, Benamara et al. (2011) use discursive constraints to find the subjective orientation of EDUs in movie reviews. An EDU can belong to four classes: explicit evaluative, subjective non-evaluative, implicit, and objective. These constraints were guided by the effect certain discourse relations may have on evaluative language: Contrasts usually reverse polarity, as in Example (18), whereas parallels and continuations preserve subjectivity, as in Example (19).

We believe that leveraging positive and negative situations would improve sentiment detection, especially in blogs or corpora about politics or economy, which tend to contain more implicit evaluation. Recent results are very encouraging. For example, using a global optimization framework, Deng et al. (2014b) achieve a 20-point increase over local sentiment detection that does not account for such situations.

The second type of implicit evaluation concerns objective words that have positive or negative connotations. Whereas denotation is the precise, literal definition of a word that might be found in a dictionary, connotation refers to emotional suggestive meanings surrounding a word. Consider the words in italics in the following three sentences:

(21) a. Jim is a *vagrant*.
b. Jim has *no fixed address*.
c. Jim is *homeless*.

All these expressions refer to exactly the same social situation, but they will evoke different associations in the reader's mind: Vagrancy has a significant negative connotation in English. It is used to describe those who live on the street and are perceived as a public nuisance. For example, there are laws against vagrancy in many locations. A homeless person, on the other hand, is not necessarily perceived as a nuisance, and the expression can connote sympathy and be used to appeal to charity. Taboada et al. (2011) noticed that some nouns and verbs often have both neutral and non-neutral connotations. For instance, *inspire* has a very positive meaning (*The teacher inspired her students to pursue their dreams*), as well as a rather neutral meaning (*This movie was inspired by*

real events). Some instances of different connotations can be addressed through word-sense disambiguation (Akkaya et al. 2011; Sumath and Inkpen 2015). In other cases, the problem is framed as domain dependency. What is considered positive in one domain may be negative in another. Example (22) was seen on the Toulouse transit system. The word *volume* changes its connotation, or polarity, with different domains of application: Volume is good for hair; (loud) volume is bad for public transit.

(22) Le volume c'est bien dans les cheveux. . . moins dans les transports.
'Volume is good in hair. . . less so in transportation.'

Although connotation has a strong impact on sentiment analysis, most current subjective lexicons contain words that are intrinsically positive or negative. To overcome this limitation, Feng et al. (2013) propose a set of induction algorithms to automatically build the first broad-coverage connotation lexicon.[8] This lexicon performs better than denotation lexicons in binary sentiment classification on SemEval and Tweet corpora. Later, Kang et al. (2014) extended this lexicon to deal with polysemous words and introduced ConnotationWordNet, a connotation lexicon over words in conjunction with senses. Connotation is also being explored in other NLP tasks like machine translation (Carpuat 2015).

The third way in which implicit evaluation can arise is when one expresses an evaluation towards an implicit aspect of an entity. This has been more frequently observed in aspect-based sentiment analysis (Liu 2012). For example, the adjective *heavy* in *The cell phone is heavy* implicitly provides a negative opinion about the aspect weight. Similarly, the verb *last* in *My new phone lasted three days* suggests that the aspect durability is assigned a negative opinion. Some of these implicit evaluations arise out of connotations, some of them because of polysemy (the problem to be solved in word sense disambiguation), and some of them because of domain dependence, as pointed out earlier. Implicit aspects can be expressed by nouns, noun phrases, or verb phrases, as in *The camera fits in my pocket*, which expresses a positive evaluation towards the size of the camera. Inferring implicit aspects from sentences first requires detecting implicit aspect clues that are often assumed to be subjective words (adjective, adverb, or noun expressions). Once clues are found, clustering methods are used to map them to their corresponding aspects (Popescu and Etzioni 2005; Su et al. 2008; Hai et al. 2011; Fei et al. 2012; Zeng and Li 2013). Although recent studies are addressing verb expressions that imply negative opinions (Li et al. 2015), identifying implicit aspects that are not triggered by sentiment words is still an open problem.

6.3 Dealing with Figurative Language

Figurative language makes use of figures of speech to convey non-literal meaning, that is, meaning that is not strictly the conventional or intended meaning of the individual words in the figurative expression. Figurative language encompasses a variety of phenomena, including metaphor, oxymoron, idiomatic expressions, puns, irony, and sarcasm.

Metaphors equate two different entities, concepts, or ideas, referred to as source and target. It has traditionally been viewed as the domain of expressive and poetic language, but studies in cognitive linguistics have clearly shown that it is pervasive in language. In cognitive linguistics, the all-encompassing view of metaphor states that

8 The lexicon is available from: www3.cs.stonybrook.edu/∼ychoi/connotation/.

it is fundamental to our conceptual system, and that metaphors in language are simply a reflection of our conceptual framework, whereby we conceptualize one domain by using language from a more familiar or basic one. For instance, political and other debates often borrow the language of war and conflict to characterize their antagonistic nature (we *win* arguments; *attack* or *shoot down* the opponent's points; *defend* our point of view). Lakoff and Johnson (1980) constitutes the foundational work in this area, and Shutova et al. (2013) and Shutova (2015) provide an excellent overview from a computational perspective.

Far from it being a phenomenon restricted to literary text, metaphor and figurative language are present in all kinds of language and at all levels of formality. According to Shutova and Teufel (2010), approximately one in three sentences in regular general text contains a metaphorical expression. Irony and sarcasm can be viewed as forms of metaphorical and figurative language, because they convey more than what is literally expressed.

Irony detection has gained relevance recently because of its importance for efficient sentiment analysis (Maynard and Greenwood 2014; Ghosh et al. 2015). Irony is a complex linguistic phenomenon widely studied in philosophy and linguistics (Grice 1975; Sperber and Wilson 1981; Utsumi 1996; Attardo 2000). The reader can refer to Barbe (1995), Winokur (2005) and Wallace (2015) for linguistic models of verbal irony. Glossing over differences across approaches, irony can be defined as an incongruity between the literal meaning of an utterance and its intended meaning. For example, to express a negative opinion towards a cell phone, one can either use a literal form using a negative opinion word, as in *This phone is a disaster*, or a non-literal form by using a positive word, as in *What an excellent phone!!* For many researchers, irony overlaps with a variety of other figurative devices such as satire, parody, and sarcasm (Clark and Gerrig 1984; Gibbs 2000). In computational linguistics, irony is often used as an umbrella term that includes sarcasm, although some researchers make a distinction between irony and sarcasm, considering that sarcasm tends to be harsher, humiliating, degrading, and more aggressive (Lee and Katz 1998; Clift 1999).

In social media, such as Twitter, users tend to utilize specific hashtags (#*irony*, #*sarcasm*, #*sarcastic*) to help readers understand that their message is ironic. These hashtags are often used as gold labels to detect irony in a supervised learning setting (i.e., learning whether a text span is ironic/sarcastic or not). In doing so, systems are not be able to detect irony without explicit hashtags, but on the positive side, it provides researchers with positive examples with high precision. There are, however, some dangers in that kind of approach. Kunneman et al. (2015) show that tweets with and without the hashtag have different characteristics. The main difference between tagged tweets and tweets labeled by humans as sarcastic (but without a hashtag) is that the tagged tweets have fewer intensified words and fewer exclamations (at least in Dutch, the language of their corpus). Kunneman et al. hypothesize that the intensification, a form of hyperbole, helps in the identification of sarcasm by readers, and that the explicit hashtag is the equivalent of non-verbal expressions in face-to-face interaction, which are used to convey nuances of meaning.

A comparative study on the use of irony and sarcasm in Twitter suggests two main findings (Wang 2013). First, sarcastic tweets tend to be more positive whereas ironic tweets are more neutral. Indeed, sarcasm being more aggressive, users seem to soften their message with the use of more positive words. Second, sarcastic tweets are more likely to contain subjective expressions. The automatic distinction between irony and sarcasm seems rather difficult, nevertheless. For example, Barbieri and Saggion (2014) report an F-score of 0.6 on sarcasm vs. irony, around 25% less than the scores obtained on

sarcasm vs. non-ironic hashtags. Besides these results, we believe that such a distinction can have an impact on both polarity analysis and rating prediction. A first step in this direction has been carried out by Hernández Farías et al. (2015) with the use of a specific feature that reverses the polarity of tweets that include #*sarcasm* or #*not*.

There are roughly two ways to infer irony or sarcasm from text: Rely exclusively on the lexical cues internal to the utterance, or combine these cues with an additional pragmatic context external to the utterance. In the first case, the speaker intentionally creates an explicit juxtaposition of incompatible actions or words that can either have opposite polarities (cf. Example (23)), or can be semantically unrelated. Explicit opposition can also arise from an explicit positive/negative contrast between a subjective proposition and a situation that describes an undesirable activity or state. The irony is inferred from the assumption that the writer and the reader share common knowledge about this situation, which is judged as being negative by cultural or social norms. Raining on summer holidays or growing older are examples of such situations.

(23) I love when my phone fails when I need it.

To detect irony in explicit and implicit oppositions, most state-of-the-art approaches rely on a variety of features gleaned from the utterance-internal context going from *n*-gram models, stylistic (punctuation, emoticons, quotations), to dictionary-based features (sentiment and affect dictionaries, slang language) (Kreuz and Caucci 2007; Burfoot and Baldwin 2009; Davidov et al. 2010; Tsur et al. 2010; González-Ibáñez et al. 2011; Gianti et al. 2012; Liebrecht et al. 2013; Reyes et al. 2013; Barbieri and Saggion 2014). Kreuz and Caucci (2007) conclude that the presence of interjections is an important indication for irony detection whereas word frequency, the presence of adjectives, adverbs, words in bold font, and punctuation do not have a strong impact. Gianti et al. (2012) found that verb tenses can be another way to study linguistic differences between humorous and objective text. Carvalho et al. (2009) proposed a set of eight patterns among which the ones based on the presence of quotations and emoticons achieved the best with accuracy of 85.4% and 68.3%, respectively, when testing on Portuguese newspaper articles. Veale and Hao (2010) focused on simile, a specific form of irony in which an element is provided with special attributes through a comparison with something quite different (e.g., *As tough as a marshmallow cardigan*). This is a form of metaphor, and is often marked by specific cues such as *like*, *about*, or *as* in English. Veale and Hao (2010) used patterns of the form *about as X as Y* or *as ADJ as*, and semantic similarities to detect ironic simile. They conclude that ironic similarities often express negative feelings using positive terms. The evaluation of this model achieves an F-measure of 73% for the irony class and 93% for the non-ironic. Qadir et al. (2015) extend this approach to learn to recognize affective polarity in similes.

In addition to these more lexical features, many authors point out the necessity of pragmatic features in the detection of this complex phenomenon. Utsumi (2004) shows that opposition, rhetorical questions, and politeness level are relevant. Burfoot and Baldwin (2009) focus on satire detection in newswire articles and introduce the notion of validity, which models absurdity by identifying a conjunction of named entities present in a given document and queries the web for the conjunction of those entities. González-Ibáñez et al. (2011) exploit the common ground between speaker and hearer by checking whether a tweet is a reply to another tweet. Reyes et al. (2013) use opposition in time and context imbalance to estimate the semantic similarity of concepts in a text to each other. Barbieri and Saggion (2014) capture the gap between rare and common words as well as the use of common vs. rare synonyms. Finally, Buschmeier et al. (2014) measure the imbalance between the overall polarity of words in a review and its star rating.

Most of these pragmatic features still rely on linguistic aspects of the tweet by using only the text of the tweet. Recent work explores other ways to go further by capturing the context outside of the utterance that is needed to infer irony. Bamman and Smith (2015) explore properties of the author (like profile information and historical salient terms), the audience (such as author/addressee interactional topics), and the immediate communicative environment (previous tweets). Wallace et al. (2015) exploit signals extracted from the conversational threads to which comments belong. Finally, Karoui et al. (2015) propose a model that detects irony in tweets containing an asserted fact of the form $Not(P)$. They hypothesize that such tweets are ironic if and only if one can prove the validity of P in reliable external sources, such as Wikipedia or online newspapers.

6.4 Extra-Linguistic Information

In the previous section, we saw how irony classification can gain in accuracy when extra-linguistic or extra-textual features are taken into account. In this section, we further discuss how sentiment analysis can benefit from these features. We focus in particular on demographic information and social network structure.

Demographic information refers to statistical data used in marketing and business to classify an audience into age, gender, race, income, location, political orientation, and other categories. Several studies have found strong correlations between the expression of subjectivity and gender and leverage these correlations for gender identification (Rao et al. 2010; Thelwall et al. 2010; Burger et al. 2011; Volkova et al. 2015). For example, women tend to use more emotion features (emoticons, exasperation, etc.) than men, or different writing styles. Recently, Volkova et al. (2014) propose an approach to exploit gender differences to improve multilingual sentiment classification in social media. The method relies on the assumption that some subjective words will be used by men, but never by women, and vice versa. Polarity may also be gender-influenced. A combination of lexical features and features representing gender-dependent sentiment terms improve subjectivity and polarity classification by 2.5% and 5% for English, 1.5% and 1% for Spanish, and 1.5% and 1% for Russian, respectively. In addition to gender, Persing and Ng (2014) explore 15 other types of demographic information to predict votes from comments posted in a popular social polling Web site. This information is found in the user's profile and includes, for example, political views (conservative, moderate, or progressive), relationship status (single, married, etc.), and whether the user is a drinker or smoker. Not all information is known, but, when it is, it is modeled as features in a voting prediction system. Results show that combining these features with inter-comment constraints improves over a baseline that uses only textual information.

Another interesting source of extra-linguistic information can be extracted from the structure of social networks. Indeed, while on review Web site reviews are typically written independently of each other, comments posted in social media are usually connected in such a way that enables the grouping of users according to specific communities. A community is often not identified in advance, but its users are expected to share common goals: circles of friends, business associates, political party members, groups of topically related conversations, and so forth. Hence, users in a given community may have similar subjective orientations. This observation has been empirically validated in several recent studies showing that sentiment can enhance community detection (Xu et al. 2011; Deitrick and Hu 2013), and users' social relationships sentiment analysis (Tan et al. 2011; Hasan and Ng 2013; Deng et al. 2014a; Vanzo et al. 2014; West et al. 2014; Naskar et al. 2016; Ren et al. 2016).

6.5 Intent Detection

6.5.1 From Sentiment to Intent Analysis. Discourse and different pragmatic context can enhance sentiment analysis systems. However, knowing what a holder likes and dislikes is only a first step in the decision making process. Consider the statements in Examples (24), (25), (26), and (27):

(24) I don't like Apple's policy overall, and will never own any Mac products.

(25) I wish to buy a beautiful house with a swimming pool.

(26) How big is the screen on the Apple iPhone 4S?

(27) I am giving birth in a month.

If we look at these examples from a sentiment analysis point of view, only the first sentence, in Example (24), would be classified as negative, the other examples being objective.[9] However, in addition to a negative opinion, the writer in Example (24) explicitly states their intention not to buy Mac products, which is not good news for Apple. In Example (25), the writer wishes for a change in their existing situation, but there is no guarantee that this wish will lead to forming an intention to buy a new house in the future. In Example (26), the writer wants to know about others' opinions and, based on these opinions, they may or may not be inclined to buy an iPhone. Finally, in Example (27), one can infer that the writer may want to buy baby products that may help Web sites to provide the most appropriate ads to display. These last two examples are typical of implicit intentions.

Knowing about the holder's future actions or plans from texts is crucial for decision makers: Does the writer intend to stop using a service after a negative experience? Do they desire to purchase a product or service? Do they prefer buying one product over another? *Intent analysis* attempts to answer these questions, focusing on the detection of future states of affairs that a holder wants to achieve.

We use the term intent as a broader term that covers *desires*, *preferences*, and *intentions*, which are mental attitudes contributing to the rational behavior of an agent. These attitudes play a motivational role and it is in concert with beliefs that they can move us to act (Bratman 1990). Indeed, before deciding to perform an action, an agent considers various desires, which are states of affairs that the agent, in an ideal world, would wish to be brought about. Desires may be in conflict and are thus subject to inconsistencies. Among these desires, only some can be potentially satisfied. The chosen desires that the agent has committed to achieving are called **intentions** (Bratman 1990; Wooldridge 2000; Perugini and Bagozzi 2004). Intentions cannot conflict with each other and have to be consistent. This constitutes an important difference between desires and intentions. This distinction has been formalized in the Belief-Desire-Intention model (Bratman 1990), an intention-based theory of practical reasoning, namely, reasoning directed toward actions.

Desires may be ordered according to preferences. A preference is commonly defined as an asymmetric, transitive ordering by an agent over outcomes, which are understood as actions that the agent can perform or goal states that are the direct result of an action of the agent. For instance, an agent's preferences may be defined over actions like *buy a new car* or by its end result like *have a new car*. Among these outcomes, some are acceptable for the agent (i.e., the agent is ready to act in such a way as to

9 A standard system that does not account for the volitive modality of *wish* would also classify Example (25) as positive.

realize them) and some outcomes are not. Among the acceptable outcomes, the agent will typically prefer some to others. Preferences are not opinions. Whereas opinions are defined as a point of view, a belief, a sentiment, or a judgment that *an agent may have about an object or a person*, preferences involve an ordering on behalf of an agent and thus are *relational and comparative*. Opinions concern absolute judgments towards objects or persons (positive, negative, or neutral), and preferences concern relative judgments towards actions (preferring them or not over others). The following examples illustrate this.

(28) The movie is not bad.

(29) The script for the first season is better than the second one.

(30) I would like to go to the cinema. Let's go and see *Madagascar 3*.

Example (28) expresses a direct positive opinion towards the movie, but we do not know if this movie is the most preferred. Example (29) expresses a comparative opinion about two movies with respect to their shared features (script). If actions involving these movies (e.g., seeing them) are clear in the context, such a comparative opinion will imply a preference, an ordering the first season scenario over the second. Finally, Example (30) expresses two preferences, one depending on the other. The first is that the speaker prefers to go to the cinema over other alternative actions; the second is: Given the option of going to the cinema, they want to see *Madagascar 3* over other possible movies.

Reasoning about preferences is also distinct from reasoning about opinions. An agent's preferences determine an order over outcomes that predicts how the agent, if they are rational, will act. This is not true for opinions. Opinions have at best an indirect link to action: I may not absolutely love what I am doing right now, but do it anyway because I prefer that outcome to any of the alternatives.

6.5.2 Intent Detection: Main Approaches. Acquiring, modeling, and reasoning with desires, preferences, and intentions are well-established fields in artificial intelligence (Cohen and Levesque 1990; Georgeff et al. 1999; Brafman and Domshlak 2009; Kaci 2011). Predicting user intentions from search queries and/or the user's click behavior has also been extensively studied in the Web search community to assist the user to search what they want more efficiently (Chen et al. 2002; Wang and Zhang 2013). There is, however, little research that investigates how to extract desires, preferences, and intentions from users' linguistic actions using NLP techniques. We survey here some existing work.

Desire extraction. Wish and desire detection from text have been explored by Goldberg et al. (2009). They define a wish as "a desire or hope for something to happen" and propose an unsupervised approach that learns if a given sentence is a wish or not. Given that the expression of wishes is domain-dependent, they first exploit redundancy in how wishes are expressed to automatically discover wish templates from a source domain. These templates are then used to predict wishes in two target domains: product reviews and political discussions. The source domain is a subset of the WISH corpus composed of about 100,000 multilingual wish sentences collected over a period of 10 days in December 2007, when Web users sent in their wishes for the new year. *Peace on earth*, *To be financially stable*, and *I wish for health and happiness for my family*, are typical sentences. Extraction suggestions for products using templates has also been explored for tweets (Dong et al. 2013). Using a small set of hand-crafted rules, Ramanand et al. (2010) focus on two specific kinds of wishes characteristic of product reviews: sentences

that make suggestions about existing products, and sentences that indicate the writer is interested in purchasing a product. The same approach has been used in Brun and Hagège (2013) to improve feature-based sentiment analysis of product reviews. It is, however, limited, since the system only detects those wishes that match previously defined rules.

Preference extraction. Preference extraction from text has been investigated with the study of comparative opinions (Jindal and Liu 2006a, 2006b; Ganapathibhotla and Liu 2008; Yang and Ko 2011; Li et al. 2013a). Given a comparison within a sentence, this task involves two steps. First extract entities, comparative words, and entity features that are being compared; then, identify the most preferred entity. In Example (29), *the first season* and *second season* are the entities, *better than* the comparative, *script* the entity feature, and *the first season* the preferred entity. This approach is quite limited, because it either only focuses on the task of identifying comparative sentences without extracting the comparative relations within the sentences, or when it does, it only considers comparisons at the sentence level, even sometimes with the assumption that there is only one comparative relation in a sentence. However, for reasoning with preferences, it is unavoidable to consider more complex comparisons with more than one dependency at a time and with a higher level than just the sentence, in order to manage all the preference complexity. Cadilhac et al. (2012) explore such an approach to automatically extract the preferences and their dependencies within each dialogue move in negotiation dialogues. They perform the extraction in two steps: first the set of outcomes; then, how these outcomes are ordered. Those extracted preferences are then used to predict trades in the win–lose game Settlers of Catan (Cadilhac et al. 2013).

Intention extraction. As for desires and preferences, intention extraction is also formulated as a classification problem: deciding whether a sentence expresses an intention or not. Sujay and Yalamanchi (2012) focus on explicit intentions and propose to categorize text according to the type of intentions it expresses among wish, praise, complain, buy, and so on. Using a naive bag-of-words approach, they achieve an accuracy of almost 67% on a social media corpus. Chen et al. (2013) also focus on explicit intentions in discussion forums such as *I am looking for a brand new car to replace my old Ford Focus*. The authors observe that this classification problem suffers from noisy data (only a few sentences express intentions) and domain-dependency of features indicating the negative class (i.e., non-intention). To deal with these issues, Chen et al. propose a transfer learning method that first classifies sentences using labeled data from a given source domain, and then applies the classifier to classify the target unlabeled data. Transfer learning has also been applied to detect implicit intentions in tweets following a two-step procedure (Ding et al. 2015): First, determine whether the sentence involves a consumption intention. If it does, extract intention words.

In summary, we see intent analysis as orthogonal and supplementary to sentiment analysis, which focuses on past/present holder's states. This is why we believe that intent detection would benefit from being built on top of sentiment analysis systems, since positive or negative sentiments are often expressed prior to future actions.

7. When Linguistics Meets Computational Linguistics: Future Directions

We firmly believe that future developments in sentiment analysis need to be grounded in linguistic knowledge (and also extra-linguistic information). In particular, discourse and pragmatic phenomena play such an important role in the interpretation of

evaluative language that they need to be taken into account if our goal is to accurately capture sentiment. The dynamic definition of sentiment that we have presented includes update functions that allow for different contextual aspects to be incorporated into the calculation of sentiment for evaluative words and expressions, and can be applied at all levels of language. We see the use of linguistic and statistical methods not as mutually exclusive, but as contributing to each other. For instance, rather than general *n*-gram bag-of-words features, other features from discourse can be used to train classifiers for sentiment analysis. Contextual features can be deployed to detect implicit evaluation, and to accurately capture the meaning in figurative expressions.

We showed in this survey that including discourse information into opinion analysis is definitively beneficial. Discourse has also been successfully deployed in machine translation (Hardmeier 2013), natural language generation (Ashar and Indukhya 2010), and language technology in general (Taboada and Mann 2006a; Webber et al. 2012). Incorporating discourse into sentiment analysis can be done by relying either on *shallow discourse processing* (using specific discourse markers, leveraging the notion of topicality, zoning, and social network structure), or through *full discourse parsing*, exploiting the entire discourse structure of a document. The shallow approach has been shown to be effective when experimented on movie/product review data, and there is an increasing amount of work on other kinds of data, such as blogs (Liu et al. 2010; Chenlo et al. 2013) and tweets (Mukherjee and Bhattacharyya 2012), where links across posts and the stream of related posts are investigated. The effectiveness of full discourse, however, strongly depends on the availability of powerful tools, such as discourse parsers. Compared with syntactic parsing and shallow semantic analysis, discourse parsing is not as mature. To date, the performance of parsers is still considerably inferior compared with the human gold standard, although significant advances have been made in the last few years (Muller et al. 2012; Ji and Eisenstein 2014; Feng 2015; Joty et al. 2015; Surdeanu et al. 2015; Perret et al. 2016), and we expect improvements to continue. Automatic discourse segmentation has attained high accuracy. For example, Fisher and Roark (2007) report an F-score of 90.5% in English. Discourse relations remain nonetheless hard to detect, due in part to the ambiguity of discourse markers, and to implicit relations. End-to-end parsing involving structured prediction methods from machine learning is also still in development. For example, Ji and Eisenstein (2014) report an accuracy of 60% for discourse relation detection and Joty et al. (2015) achieve above 55% for text-level relation detection in the RST Treebank. Muller et al. (2012) also achieve between 47% and 66% accuracy on the ANNODIS corpus, annotated following SDRT. This may explain why most state-of-the-art NLP applications that rely on discourse do not yet offer a substantial boost compared with discourse-unaware systems.

An additional problem is the domain dependence of many of the existing parsers, which have been trained on newspaper articles, mostly versions of the Penn corpus of *Wall Street Journal* articles, either in its RST annotation (Carlson et al. 2002), SDRT annotation (Afantenos et al. 2012), or the Penn Discourse TreeBank annotation (Prasad et al. 2008). It is no surprise, then, that they do not perform very well on reviews. A possible solution would be to train a parser on gold discourse structure annotations and sentiment labels, such as the SFU Review Corpus (Taboada et al. 2008) or the CASOAR Corpus (Benamara et al. 2016). On the negative side, such corpora are too small to train a discourse parser and a competitive sentiment analysis system. On the positive side, review-style documents are relatively short , which can make parsers less sensitive to errors due to long dependency attachments. Also, given that not all discourse relations are sentiment relevant, the number of relations to be predicted can be reduced. This might concern framing relations like BACKGROUND and CIRCUMSTANCE but also

some temporal relations such as SEQUENCE or SYNCHRONOUS. On the other hand, relations can be grouped according to their similar effects on both subjectivity and polarity analysis. One possible grouping could be *argumentative relations* that are used to support (e.g., MOTIVATION, JUSTIFICATION, INTERPRETATION) or oppose (e.g., CONTRAST, CONCESSION, ANTITHESIS) claims and theses, *causal relations* (e.g., RESULT, CONDITION), and *structural relations* (e.g., ALTERNATIVE, CONTINUATION). *Thematic relations* like ELABORATION, SUMMARY, and RESTATEMENT have also a strong impact on evaluative discourse. Their effect is, however, close to support relations. We believe that discarding certain relations and grouping others will make discourse parsers more reliable.

Besides domain dependency, parsing is by definition theory-dependent, which means that a system trained to learn RST relations fails to predict SDRT or PDTB relations. Indeed, each theory has its own hierarchy of discourse relations, but relations tend to overlap or be related in a few specific ways: A relation *R* in one approach can correspond to several relations in another approach and vice versa; a relation may be defined in one approach but not taken into account in another; and, finally, relations across approaches may have similar names, but different definitions. One solution to this problem is to map relations across approaches to a unified hierarchy. Merging different discourse relation taxonomies has several advantages. First of all, for classification tasks such as discourse parsing, access to larger amounts of data is likely to yield better results. Secondly, and from a more theoretical point of view, we think that differences across approaches are minimal, and a unified set of relations is possible. Third, a unified set of discourse relations would allow us to compile a list of discourse markers and other signals for those relations, which would also benefit discourse annotation. Recent efforts to merge existing discourse relation taxonomies and annotations should help improve discourse parsing (Benamara and Taboada 2015; Rehbein et al. 2015). Notable also is work being carried out within the COST Action TextLink, a pan-European initiative to unify definitions of relations and their signals across languages (`http://www.textlink.ii.metu.edu.tr/`).

Once powerful discourse parsers are developed, the argumentative structure of evaluative text can be fully exploited. Processing arguments for sentiment analysis is still at an early stage and we feel that recent progress in argument mining will likely spur new research in this direction (Bex et al. 2013; Stab and Gurevych 2014; Peldszus and Stede 2015).

We see sentiment analysis not as an aim per se but as a first step in processing and understanding large amounts of data. Indeed, sentiment analysis has strong interactions with social media (Farzindar and Inkpen 2015), big data (Arora and Malik 2015), and, more importantly, with modeling human behavior, that is, how sentiment translates into action. We defined "the sentiment to action" process as intent detection (cf. Section 6.5), an area which gives linguistic objects a predictive power such as predicting voter behavior and election results (Yano et al. 2013; Qiu et al. 2015), predicting deception (Fitzpatrick et al. 2015), or intention to buy (Ding et al. 2015). Predictions can also be derived on the basis of extra-linguistic sources of information such as characteristics of the author and their online interactions (Qiu et al. 2015). State-of-the-art approaches are still heavily dependent on bag-of-words representations. We believe that predicting a user's future actions from text (and speech) needs to integrate models from artificial intelligence with NLP techniques to find specific *intent signals*, such as changes in the argumentation chain; the social relationship between discourse participants; topic changes; user's beliefs; the sudden use of sentiments or emotions of a certain type (like aggressive expressions); or the correlation between genre and the use of specific

linguistic devices. Intent detection is an emerging research area with great potential in business applications (Wang et al. 2015).

In summary, we believe that the *discourse turn* that computational linguistics is experiencing can be successfully combined with data-driven methods as part of the effort to accurately capture sentiment and evaluative language.

Acknowledgments

This research was funded by a Discovery Grant from the Natural Sciences and Engineering Council of Canada to Maite Taboada, and ERC grant 269427 (STAC). We thank Nicholas Asher, Paola Merlo, and the two anonymous reviewers for their comments or suggestions. We take responsibility for any errors that may remain.

References

Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems*, 26(3):1–34.

Afantenos, Stergos, Nicholas Asher, Farah Benamara, Myriam Bras, Cécile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paule Péry-Woodley, Laurent Prévot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: The ANNODIS corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, LREC 2012, pages 2727–2734, Istanbul.

Agarwal, Shashank and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23):3174–3180.

Agarwal, Apoorv, Fadi Biadsy, and Kathleen R. McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic *n*-grams. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2009, pages 24–32, Athens.

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38, Portland, OR.

Agerri, Rodrigo and Ana García-Serrano. 2010. Q-WordNet: Extracting polarity from WordNet senses. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2300–2305, Malta.

Akkaya, Cem, Janyce Wiebe, and Rada Mihalcea. 2009. Subjectivity word sense disambiguation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 190–199, Singapore.

Akkaya, Cem, Janyce Wiebe, Alexander Conrad, and Rada Mihalcea. 2011. Improving the impact of subjectivity word sense disambiguation on contextual opinion analysis. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 87–96, Portland, OR.

Andreevskaia, Alina and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 290–298, Columbus, OH.

Argamon, Shlomo, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2009. Automatically determining attitude type and force for sentiment analysis. In Zygmunt Vetulani and Hans Uszkoreit, editors, *Human Language Technology: Challengies of the Information Society*. Springer, Berlin, pages 218–231.

Arora, Deepali and Piyush Malik. 2015. Analitics: Key to go from generating big data to deriving business value. In *Proceedings of the IEEE First International Conference on Big Data Computing Service and Applications*, pages 446–452, San Francisco, CA.

Ashar, Jayen and Nitin Indukhya. 2010. A unifying view of computational discourse and natural language generation. *ACM Computing Surveys*, pages 1–30.

Asher, Nicholas and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Asher, Nicholas, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling

opinion in discourse: A preliminary study. In *Proceedings of the Computational Linguistics Conference*, pages 7–10, Manchester.

Asher, Nicholas, Farah Benamara, and Yannick Mathieu. 2009. Appraisal of opinion expressions in discourse. *Linguisticae Investigationes*, 32(2):279–292.

Attardo, Salvatore. 2000. Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826.

Aue, Anthony and Michael Gamon. 2005. Customizing sentiment classifiers to new domains: A case study. In *Proceedings of Recent Advances in Natural Language Processing*, Borovets.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Conference on International Language Resources and Evaluation*, pages 2200–2204, Malta.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the International Conference on Computational Linguistics - Volume 1*, pages 86–90, Montréal.

Bakhtin, Mikhail. 1981. Discourse in the novel. In Michael Holquist, editor, *The Dialogic Imagination: Four Essays by M. M. Bakhtin*. University of Texas Press, Austin, pages 259–422.

Balahur, Alexandra and José M. Perea-Ortega. 2015. Sentiment analysis system adaptation for multilingual processing. *Information Processing Management*, 51(4):547–556.

Balahur, Alexandra, Jesús M. Hermida, Andrés Montoyo, and Rafael Muñoz. 2011. EmotiNet: A knowledge base for emotion detection in text built on the Appraisal theories. In *Natural Language Processing and Information Systems*, volume 6716 of *Lecture Notes in Computer Science*. Springer, pages 27–39.

Balahur, Alexandra, Jesús M. Hermida, and Andrés Montoyo. 2012. Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4):742–753.

Bamman, David and Noah A. Smith. 2015. Contextualized sarcasm detection on Twitter. In *Proceedings of the International Conference on Web and Social Media*, pages 574–577, Oxford.

Banfield, Ann. 1982. *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. Routledge and Kegan Paul, Boston.

Barbe, Katharina. 1995. *Irony in Context*. John Benjamins, Amsterdam.

Barbieri, Francesco and Horacio Saggion. 2014. Modelling irony in Twitter: Feature analysis and evaluation. In *Proceedings of Language Resources and Evaluation Conference*, pages 4258–4264, Reykjavik.

Becker, Israela and Vered Aharonson. 2010. Last but definitely not least: On the role of the last sentence in automatic polarity-classification. In *Proceedings of the Annual Association of Computational linguistics (Volume 2)*, pages 331–335, Upsala.

Bednarek, Monika. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus*. Continuum, London.

Benamara, Farah and Maite Taboada. 2015. Mapping different rhetorical relation annotations: A proposal. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 147–152, Denver, CO.

Benamara, Farah, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO.

Benamara, Farah, Baptiste Chardon, Yannick Mathieu, and Vladimir Popescu. 2011. Towards context-based subjectivity analysis. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 1180–1188, Chian Mai.

Benamara, Farah, Baptiste Chardon, Yvette Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2012. How do negation and modality impact opinions? In *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18, Jeju Island.

Benamara, Farah, Véronique Moriceau, and Yvette Yannick Mathieu. 2014. Fine-grained semantic categorization of opinion expressions for consensus detection (Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus) [in French]. In *TALN-RECITAL 2014 Workshop DEFT 2014: DÉfi Fouille de Textes (DEFT 2014 Workshop: Text Mining Challenge)*, pages 36–44, Marseille.

Benamara, Farah, Nicholas Asher, Yannick Mathieu, Vladimir Popescu, and Baptiste

Chardon. 2016. Evaluation in discourse: A corpus-based study. *Dialogue and Discourse*, 7(1):1–49.

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, pages 22–24, Palo Alto, CA.

Bex, Floris, John Lawrence, Mark Snaith, and Chris Reed. 2013. Implementing the argument Web. *Communications of the ACM*, 56(10):66–73.

Bhatia, Parminder, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon.

Biber, Douglas and Edward Finegan. 1988. Adverbial stance types in English. *Discourse Processes*, 11(1):1–34.

Biber, Douglas and Edward Finegan. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text*, 9(1):93–124.

Bieler, Heike, Stefanie Dipper, and Manfred Stede. 2007. Identifying formal and functional zones in film reviews. In *Proceedings of the 8th SIGDIAL Workshop*, pages 75–78, Antwerp.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Prague.

Bloom, Kenneth and Shlomo Argamon. 2010. Automated learning of Appraisals expression patterns. In Stefanie Wulff, Stefan Th. Gries, and Mark Davies, editors, *Corpus-linguistic Applications: Current Studies, New Directions*. Rodopi, Amsterdam, pages 249–260.

Bloom, Kenneth, Navendu Garg, and Shlomo Argamon. 2007. Extracting Appraisal expressions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics / Human Language Technologies*, pages 308–315, Rochester, NY.

Bollegala, Danushka, David Weir, and John Carroll. 2011. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 132–141, Portland, OR.

Bollegala, Danushka, David J. Weir, and John A. Carroll. 2013. Cross-domain sentiment classification using a sentiment sensitive thesaurus. *IEEE Transaction on Knowledge Data Engeneering*, 25(8): 1719–1731.

Boyd-Graber, Jordan and Philip Resnik. 2010. Holistic sentiment analysis across languages: Multilingual supervised Latent Dirichlet Allocation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 45–55, Cambridge, MA.

Boye, Kasper and Peter Harder. 2009. Linguistic categories and grammaticalization. *Functions of Language*, 16(1):9–43.

Bradley, Margaret M. and Peter J. Lang. 1999. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical Report C-1, Center for Research in Psychophysiology, University of Florida.

Brafman, Ronen I. and Carmel Domshlak. 2009. Preference handling: An introductory tutorial. *AI Magazine*, 30(1):58–86.

Bratman, Michael. 1990. Dretske's desires. *Philosophy and Phenomenological Research*, 50:795–800.

Brody, Samuel and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of the Annual Conference of the North American Association for Computational Linguistics*, pages 804–812, Upsala.

Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. In *Recent Advances in Natural Language Processing*, pages 50–54, Borovets.

Bruce, Rebecca F. and Janyce M. Wiebe. 1999. Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

Brun, Caroline and Caroline Hagège. 2013. Suggestion mining: Detecting suggestions for improvement in users' comments. *Research in Computing Science*, 70:199–209.

Burfoot, Clint and Clint Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 161–164, Singapore.

Burger, John D., John Henderson, George Kim, and Guido Zarrella. 2011.

Discriminating gender on Twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2011, pages 1301–1309, Edinburgh.

Buschmeier, Konstantin, Philipp Cimiano, and Roman Klinger. 2014. An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 42–49, Baltimore, MD.

Cadilhac, Anaïs, Nicholas Asher, Farah Benamara, Vladimir Popescu, and Mohamadou Seck. 2012. Preference extraction from negotiation dialogues. In *European Conference on Artificial Intelligence*, pages 211–216, Montpellier.

Cadilhac, Anaïs, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 357–368, Seattle, WA.

Cambria, Erik and Amir Hussain. 2012. *Sentic Computing: Techniques, Tools, and Applications*. Springer, Berlin.

Cambria, Erik and Amir Hussain. 2015. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer, Berlin.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank. Linguistic Data Consortium. Philadelphia, PA.

Carpuat, Marine. 2015. Connotation in translation. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, Lisbon.

Carvalho, Paula, Luís Sarmento, Mário J. Silva, and Eugénio De Oliveira. 2009. Clues for detecting irony in user-generated contents: Oh...!! It's so easy;-). In *Proceedings of the 1st International CIKM workshop on Topic-sentiment Analysis for Mass Opinion*, pages 53–56, Hong Kong.

Chafe, Wallace and Johanna Nichols. 1986. *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, Norwood, NJ.

Chafe, Wallace. 1986. Evidentiality in English conversation and academic writing. In Wallace Chafe and Johanna Nichols, editors, *Evidentiality: The Linguistic Coding of Epistemology*. Ablex, Norwood, NJ, pages 261–272.

Chafe, Wallace. 1994. *Discourse Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, Chicago.

Charaudeau, Patrick. 1992. *Grammaire du sens et de L'expression*. Hachette, Paris.

Chardon, Baptiste, Farah Benamara, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013a. Sentiment composition using a parabolic model. In *Proceedings of the International Conference on Computational Semantics*, pages 47–58, Potsdam.

Chardon, Baptiste, Farah Benamara, Yvette Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013b. Measuring the effect of discourse structure on sentiment analysis. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing*, pages 25–37, Samos.

Chen, Zheng, Fan Lin, Huan Liu, Yin Liu, Wei-Ying Ma, and Liu Wenyin. 2002. User intention modeling in web applications using data mining. *World Wide Web*, 5(3):181–191.

Chen, Zhiyuan, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. 2013. Identifying intention posts in discussion forums. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 1041–1050, Atlanta, GA.

Chenlo, Jose M., Alexander Hogenboom, and David E. Losada. 2013. Sentiment-based ranking of blog posts using Rhetorical Structure Theory. In *Natural Language Processing and Information Systems*, volume 7934 of *Lecture Notes in Computer Science*. Springer, Berlin, pages 13–24.

Choi, Yejin and Claire Cardie. 2008. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of Empirical Methods in Natural Language Processing*, pages 793–801, Waikiki, HI.

Choi, Yejin, Eric Breck, and Claire Cardie. 2006. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Sydney.

Clark, Herbert H. and Richard J. Gerrig. 1984. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126.

Clift, Rebecca. 1999. Irony in conversation. *Language in Society*, 28:523–553.

Clough, Patricia Ticineto and Jean O'Malley Halley. 2007. *The Affective Turn: Theorizing the Social*. Duke University Press, Durham, NC.

Cohen, Philip R. and Hector J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence*, 42(2-3):213–261.

Conrad, Susan and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford University Press, Oxford, pages 56–73.

Councill, Isaac G., Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala.

Cruz, Noa, Maite Taboada, and Ruslan Mitkov. 2016. A machine learning approach to negation and speculation detection for sentiment analysis. *Journal of the Association for Information Science and Technology*, 67:2218–2136.

Daille, Béatrice, Estelle Dubreil, Laura Monceaux, and Mathieu Vernier. 2011. Annotating opinion—Evaluation of blogs: The Blogoscopy corpus. *Language Resources and Evaluation*, 45(4):409–437.

Dave, Kushal, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the International Conference on World Wide Web*, pages 519–528, Budapest.

Davidov, Dmitry, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 107–116, Uppsala.

Davidson, Richard J., Klaus R. Scherer, and H. Hill Goldsmith. 2003. *Handbook of Affective Sciences*. Oxford University Press, Oxford.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

de Melo, Gerard and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Deitrick, William and Wei Hu. 2013. Mutually enhancing community detection and sentiment analysis on Twitter networks. *Journal of Data Analysis and Information Processing*, 1(3):19–29.

Denecke, Kerstin. 2009. Are SentiWordNet scores suited for multi-domain sentiment classification? In *IEEE International Conference on Digital Information Management*, pages 33–38, Ann Arbor, MI.

Deng, Hongbo, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu. 2014a. Exploring and inferring user-user pseudo-friendship for sentiment analysis with heterogeneous networks. *Statistical Analysis and Data Mining*, 7(4):308–321.

Deng, Lingjia, Janyce Wiebe, and Yoonjung Choi. 2014b. Joint inference and disambiguation of implicit sentiments via implicature constraints. In *Proceedings of the International Conference on Computational Linguistics*, pages 79–88, Dublin.

Denis, Alexandre, Samuel Cruz-Lara, Nadia Bellalem, and Lofti Bellalem. 2014. Synalp-Empathic: A valence shifting hybrid system for sentiment analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 605–609, Dublin.

Ding, Xiao, Ting Liu, Junwen Duan, and Jian-Yun Nie. 2015. Mining user consumption intention from social media using domain adaptive convolutional neural network. In *Proceedings of the Conference on the American Association of Artificial Intelligence*, pages 2389–2395.

Dong, Li, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu. 2013. The automated acquisition of suggestions from tweets. In *Proceedings of the Twenty-Seventh American Association for Artificial Intelligence*, pages 239–245, Bellevue, WA.

Dong, Li, Furu Wei, Ming Zhou, and Ke Xu. 2014. Adaptive multi-compositionality for recursive neural models with applications to sentiment analysis. In *Proceedings of the American Association for Artificial Intelligence*, pages 1537–1543, Québec City.

Dowty, David R., Robert E. Wall, and Stanley Peters. 1981. *Introduction to Montague Semantics*. Kluwer, Dordrecht.

Dragut, Eduard, Hong Wang, Clement Yu, Prasad Sistla, and Weiyi Meng. 2012. Polarity consistency checking for sentiment dictionaries. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 997–1005, Jeju Island.

Ekman, Paul. 1984. Expression and the nature of emotion. In K. Scherer and

P. Ekman, editors, *Approaches to Emotion*. Erlbaum, Hillsdale, NJ, pages 319–344.
Esuli, Andrea and Fabrizio Sebastiani. 2015. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery from Data*, 10(1):Article 27.
Farzindar, Atefeh and Diana Inkpen. 2015. *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool publishers.
Fei, Geli, Bing Liu, Meichun Hsu, Malú Castellanos, and Riddhiman Ghosh. 2012. A dictionary-based approach to identifying aspects implied by adjectives for opinion mining. In *Proceedings of the International Conference on Computational Linguistics*, pages 309–318, Mumbai.
Feldman, Ronen. 2013. Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM*, 56(4):82–89.
Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
Feng, Vanessa Wei and Graeme Hirst. 2014. A linear-time bottom–up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 511–521, Baltimore, MD.
Feng, Song, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1774–1784, Sofia.
Feng, Vanessa Wei. 2015. *RST-style discourse parsing and its applications in discourse analysis*. Ph.D. thesis, Computer Science, University of Toronto, Canada.
Fisher, Seeger and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 488–495, Prague.
Fitzpatrick, Eileen, Joan Bachenko, and Tommaso Fornaciari. 2015. *Automatic Detection of Verbal Deception*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
Francisco, Virginia, Pablo Gervàs, and Federico Peinado. 2010. Ontological reasoning for improving the treatment of emotions in text. *Knowledge and Information Systems*, 25(3):421–443.
Fries, Peter H. 1995. Themes method of development and texts. In R. Hasan and P. Fries, editors, *On Subject and Theme: A Discourse Functional Perspective*. John Benjamins, Amsterdam/Philadelphia, pages 317–359.
Gamon, Michael. 2004. 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*, pages 841–847, Geneva.
Ganapathibhotla, Murthy and Bing Liu. 2008. Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 241–248, Manchester.
Gangemi, Aldo, Valentina Presutti, and Diego Reforgiato Recupero. 2014. Frame-based detection of opinion holders and topics: A model and a tool. *IEEE Computational Intelligence Magazine*, 9(1):20–30.
Ganu, Gayatree, Yogesh Kakodkar, and Amélie Marian. 2013. Improving the quality of predictions using textual information in online user reviews. *Information Systems*, 38(1):1–15.
Gao, Dehong, Furu Wei, Wenjie Li, Xiaohua Liu, and Ming Zhou. 2015. Cross-lingual sentiment lexicon learning with bilingual word graph label propagation. *Computational Linguistics*, 41(1):21–40.
Georgeff, Michael P., Barney Pell, Martha E. Pollack, Milind Tambe, and Michael Wooldridge. 1999. The belief-desire-intention model of agency. In *Proceedings of the 5th International Workshop on Intelligent Agents V, Agent Theories, Architectures, and Languages*, pages 1–10.
Gerani, Shima, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613.
Ghosh, Aniruddha, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation, Co-located with NAACL*, pages 470–478, Denver, CO.

Giannakidou, Anastasia. 1995. On the semantic licensing of polarity items. In Anastasios-Phoevos Christidis, Maria Margariti-Roga, and Argyris Arhakis, editors, *Studies in Greek Linguistics 15: Proceedings of the Annual Meeting of the Department of Linguistics*, pages 406–418, University of Thessaloniki.

Gianti, Andrea, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel Italian corpus for sentiment analysis. In *Proceedings of the Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, pages 1–7, Istanbul.

Gibbs, Raymond W. 2000. Irony in talk among friends. *Metaphor and Symbol*, 15(1–2):5–27.

Ginsberg, Benjamin. 2011. *The Fall of the Faculty: The Rise of the All-administrative University and Why it Matters*. Oxford University Press, Oxford.

Goldberg, Andrew B. and Xiaojin Zhu. 2006. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the Workshop on Graph Based Methods for Natural Language Processing*, pages 45–52, New York City, NY.

Goldberg, Andrew B., Nathanael Fillmore, David Andrzejewski, Zhiting Xu, Bryan Gibson, and Xiaojin Zhu. 2009. May all your wishes come true: A study of wishes and how to recognize them. In *Proceedings of Human Language Technologies and the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 263–271, Boulder, CO.

González-Ibáñez, Roberto, Smaranda Muresan, and Nina Wacholde. 2011. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–586, Portland, OR.

Goutsos, Dyonisos. 1996. A model of sequential relations in expository test. *Text*, 16(4):501–533.

Goyal, Amit, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Speech Acts. Syntax and Semantics, Volume 3*. Academic Press, New York, pages 41–58.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics, 21(2)*: 203–225.

Haas, Michael and Yannick Versley. 2015. Subsentential sentiment on a shoestring: A crosslingual analysis of compositional classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 694–704, Denver, CO.

Hai, Zhen, Kuiyu Chang, and Jung-jae Kim. 2011. Implicit feature identification via co-occurrence association rule mining. In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing*, pages 393–404, Tokyo.

Hai, Zhen, Kuiyu Chang, and Gao Cong. 2012. One seed to find them all: Mining opinion features via association. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, CIKM, pages 255–264, Maui, HI.

Hall, David Leo Wright, Greg Durrett, and Dan Klein. 2014. Less grammar, more features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 228–237, Baltimore, MD.

Halliday, Michael A. K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Halliday, Michael A. K. and Christian M. I. M. Matthiessen. 2014. *An Introduction to Functional Grammar*. Arnold, London, 4th edition.

Halliday, Michael A. K. 1985. *An Introduction to Functional Grammar*. Arnold, London, 1st edition.

Hardmeier, Christian. 2013. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

Hasan, Kazi Saidul and Vincent Ng. 2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 816–821, Sofia.

Hasan, Kazi Saidul and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Doha.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic

orientation of adjectives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–181, Madrid.

Haviland, Susan E. and Herbert H. Clark. 1974. What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour*, 13:512–521.

He, Yulan and Deyu Zhou. 2011. Self-training from labeled features for sentiment analysis. *Information Processing Management*, 47(4):606–616.

He, Yulan, Chenghua Lin, Wei Gao, and Kam-Fai Wong. 2013. Dynamic joint sentiment-topic model. *ACM Transactions on Intelligent Systems and Technology*, 5(1):6.

Hearst, Marti A. 1992. Direction-based text interpretation as an information access refinement. In Paul S. Jacobs, editor, *Text-based Intelligent Systems*. Erlbaum, Hillsdale, NJ, pages 257–274.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, NM.

Heerschop, Bas, Frank Goossen, Alexander Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska de Jong. 2011. Polarity analysis of texts using discourse structure. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 1061–1070, Glasgow.

Hernández Farías, Delia Irazú, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, and Cristina Bosco. 2015. Valento: Sentiment analysis of figurative language tweets with irony and sarcasm. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 694–698, Denver, CO.

Hernault, Hugo, Helmut Prendinger, David A. duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using Support Vector Machine classification. *Dialogue and Discourse*, 1(3):1–33.

Hobbs, Jerry. 1979. Coherence and coreference. *Cognitive Science (3)*, 8:67–90.

Hogenboom, Alexander, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska de Jong. 2010. Mining economic sentiment using argumentation structures. In *Advances in Conceptual Modeling, Applications and Challenges*, volume 6413 of *Lecture Notes in Computer Science*. Springer, Berlin, pages 200–209.

Hogenboom, Alexander, Flavius Frasincar, Franciska de Jong, and Uzay Kaymak. 2015. Using rhetorical structure in sentiment analysis. *Communications of the ACM*, 58(7):69–77.

Horn, Laurence. 1989. *A Natural History of Negation*. University of Chicago Press.

Hu, Minqing and Bing Liu. 2004a. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA.

Hu, Minqing and Bing Liu. 2004b. Mining opinion features in customer reviews. In *Proceedings of the American Association for Artificial Intelligence*, pages 755–760, San Jose, CA.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. John Benjamins, Amsterdam.

Hunston, Susan and Geoff Thompson. 2000a. Evaluation: An introduction. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford University Press, Oxford, pages 1–27.

Hunston, Susan and Geoff Thompson, editors. 2000b. *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford University Press, Oxford.

Hunston, Susan. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. Routledge, New York.

Izard, Carroll Ellis. 1971. *The Face of Emotion*. Appleton Century Crofts, New York.

Jakob, Niklas and Iryna Gurevych. 2010. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045, Cambridge, MA.

Ji, Yangfeng and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 13–24, Baltimore, MD.

Jia, Lifeng, Clement Yu, and Weiyi Meng. 2009. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 1827–1830, Hong Kong.

Jiang, Jing and ChengXiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *Proceedings of the Annual*

Meeting of the Association for Computational Linguistics, pages 264–271, Prague.
Jindal, Nitin and Bing Liu. 2006a. Identifying comparative sentences in text documents. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, pages 244–251, Seattle, WA.
Jindal, Nitin and Bing Liu. 2006b. Mining comparative sentences and relations. In *Proceedings of the American Association for Artificial Intelligence*, pages 1331–1336, Boston, MA.
Jing-Schmidt, Zhuo. 2007. Negativity bias in language: A cognitive-affective model of emotive intensifiers. *Cognitive Linguistics*, 18(3):417–443.
Johansson, Richard and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.
Joshi, Mahesh and Carolyn Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 313–316, Singapore.
Joty, Shafiq, Giuseppe Carenini, and Raymond Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.
Kaci, Souhila. 2011. *Working with Preferences: Less Is More*. Cognitive Technologies. Springer.
Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer, Dordrecht.
Kang, Jun Seok, Song Feng, Leman Akoglu, and Yejin Choi. 2014. ConnotationWordNet: Learning connotation over the word+sense network. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1544–1554, Baltimore, MD.
Karoui, Jihen, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 644–650, Beijing.
Kennedy, Alistair and Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.
Khurshid, Ahmad. 2013. *Affective Computing and Sentiment Analysis: Emotion, Metaphor and Terminology*. Springer, Berlin.
Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 1367–1373, Geneva.
Kim, Soo-Min and Eduard Hovy. 2005. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI Workshop on Question Answering in Restricted Domains*, Pittsburgh, PA.
Kim, Soo-Min and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text*, pages 1–8, Sydney.
Knott, Alistair and Robert Dale. 1994. Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes*, 18:35–62.
Koppel, Moshe and Jonathan Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.
Kreuz, Roger J. and Gina M. Caucci. 2007. Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pages 1–4, Rochester, NY.
Kunneman, Florian, Christine Liebrecht, Margot van Mulken, and Antal van den Bosch. 2015. Signaling sarcasm: From hyperbole to hashtag. *Information Processing and Management*, 51:500–509.
Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.
Langacker, Ronald W. 1990. Subjectification. *Cognitive Linguistics*, 1(1):5–38.
Lawrence, John and Chris Reed. 2015. Combining argument mining techniques. In *Working Notes of the 2nd Argumentation Mining Workshop*, pages 127–136, Denver, CO.
Lazaridou, Angeliki, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1630–1639, Sofia.
Lee, Christopher J. and Albert N. Katz. 1998. The differential role of ridicule in sarcasm and irony. *Metaphor and Symbol*, 13(1):1–15.

Leung, Cane Wing-Ki, Chi-Fai Chan Stephen, Fu-Lai Chung, and Grace Ngai. 2011. A probabilistic rating inference framework for mining user preferences from reviews. *World Wide Web*, 14(2):187–215.

Levelt, Willem J. M. 1981. The speaker's linearization problem. *Philosophical Transactions Royal Society London. Biological Sciences*, 295:305–315.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Li, Shoushan, Sophia Y. M. Lee, Ying Chen, Chu-Ren Huang, and Guodong Zhou. 2010. Sentiment classification and polarity shifting. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 635–643, Beijing.

Li, Shoushan, Zhongqing Wang, Guodong Zhou, and Sophia Yat Mei Lee. 2011. Semi-supervised learning for imbalanced sentiment classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1826–1831, Barcelona.

Li, Shasha, Chin-Yew Lin, Young-In Song, and Zhoujun Li. 2013a. Comparable entity mining from comparative questions. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1498–1509.

Li, Shoushan, Yunxia Xue, Zhongqing Wang, and Guodong Zhou. 2013b. Active learning for cross-domain sentiment classification. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, pages 2127–2133, Beijing.

Li, Huayi, Arjun Mukherjee, Jianfeng Si, and Bing Liu. 2015. Extracting verb expressions implying negative opinions. In *Proceedings of the Twenty-Ninth American Association on Artificial Intelligence*, pages 2411–2417, Austin, TX.

Liang, Jiguang, Xiaofei Zhou, Li Guo, and Shuo Bai. 2015. Feature selection for sentiment classification using matrix factorization. In *Proceedings of the International Conference on World Wide Web*, pages 63–64, Florence.

Liebrecht, Christine, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets #not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, Atlanta, GA.

Lin, Chenghua and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 375–384, Hong Kong.

Liu, Jingjing and Stephanie Seneff. 2009. Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Singapore.

Liu, Feifan, Wang Dong, Li Bin, and Liu Yang. 2010. Improving blog polarity classification via topic analysis and adaptive methods. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*, pages 309–312, Los Angeles, CA.

Liu, Yang, Xiaohui Yu, Bing Liu, and Zhongshuai Chen. 2014. Sentence-level sentiment analysis in the presence of modalities. In *Computational Linguistics and Intelligent Text Processing*, volume 8404 of *CICLing 2014*, Kathmandu, pages 1–16.

Liu, Bing. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.

Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge.

Lizhen, Qu, Georgiana Ifrim, and Gerhard Weikum. 2010. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the International Conference on Computational Linguistics*, pages 913–921, Beijing.

Lochbaum, Karen Elizabeth. 1998. Using collaborative plans to model the intentional structure of discourse. *Computational Linguistics*, 24:525–572.

Lu, Yue, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th International Conference on World Wide Web*, pages 131–140, Madrid.

Maks, Isa and Piek Vossen. 2012. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688.

Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manoussos, Katakis Ioannis, Varlamis Iraklis, and Tsatsaronis George. 2014. Pythia: Employing lexical and semantic features for sentiment analysis. In *Machine Learning and Knowledge Discovery in Databases*, volume 8726 of *Lecture Notes in Computer Science*, pages 448–451. Springer, Berlin.

Mao, Yi and Guy Lebanon. 2007. Isotonic conditional random fields and local

sentiment flow. In *Proceedings of the 21th Conference on Neural Information Processing Systems*, pages 961–968, Vancouver.
Martin, James R. and Peter White. 2005. *The Language of Evaluation: Appraisal in English*. Palgrave, New York.
Martin-Wanton, Tamara, Aurora Pons-Porrata, Andrés Montoyo, and Alexandra Balahur. 2010. Word sense disambiguation in opinion mining: Pros and cons. *Journal on Research in Computing Science*, 46:119–130.
Martin, James R. 2000. Beyond exchange: Appraisal systems in English. In Susan Hunston and Geoff Thompson, editors, *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford University Press, Oxford, pages 142–175.
Martin, James R. 2014. Evolving Systemic Functional Linguistics: Beyond the clause. *Functional Linguistics*, 1(3):1–24.
Mathieu, Yvette Yannick and Christiane Fellbaum. 2010. Verbs of emotion in French and English. In *The 5th International Conference of the Global WordNet Association*, Mumbai.
Mathieu, Yvette Yannick. 2005. Annotation of emotions and feelings in texts. In *Affective Computing and Intelligent Interaction*, volume 3784 of *Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pages 350–357.
Matsumoto, Shotaro, Hiroya Takamura, and Manabu Okumura. 2005. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pages 301–311, Hanoi.
Maynard, Diana and Mark A. Greenwood. 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pages 4238–4243, Reykjavik.
McDonald, Ryan, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. 2007. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of the Association for Computational Linguistics*, pages 432–439, Prague.
Mihalcea, Rada, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Association for Computational Linguistics*, pages 976–983, Prague, Czech Republic.
Mitchell, Margaret, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, WA.
Mittal, Namita, Basant Agarwal, Garvit Chouhan, Prateek Pareek, and Nitin Bania. 2013. Discourse based sentiment analysis for Hindi reviews. In *Pattern Recognition and Machine Intelligence*, volume 8251 of *Lecture Notes in Computer Science*. Springer, Berlin, pages 720–725.
Mochales, Raquel and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.
Moghaddam, Samaneh and Martin Ester. 2011. ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–674, Beijing
Mohammad, Saif, Parinaz Sobhani, and Svetlana Kiritchenko. 2016. Stance and sentiment in tweets. *ACM Transactions on Internet Technology*.
Mohammad, Saif M. 2016. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herb Meiselman, editor, *Emotion Measurement*, pages 201–237. Elsevier, Amsterdam.
Moilanen, Karo and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing*, pages 378–382, Borovets.
Moore, Johanna D. and Cecile L. Paris. 1993. Planning text for advisory dialogues: Capturing intentional and rhetorical information. *Computational Linguistics*, 19:651–694.
Morante, Roser and Caroline Sporleder. 2012a. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.
Morante, Roser and Caroline Sporleder. 2012b. Special issue on modality and negation. *Computational Linguistics*, 38(2):223–260.
Mukherjee, Subhabrata and Pushpak Bhattacharyya. 2012. Sentiment analysis in Twitter with lightweight discourse analysis. In *Proceedings of International Conference on Computational Linguistics*, pages 1847–1864, Mumbai.
Mullen, Tony and Collier Nigel. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 412–418, Barcelona.

Muller, Philippe, Stergos Afantenos, Denis Pascal, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of the Computational Linguistics Conference*, pages 1883–1900, Mumbai.

Musto, Cataldo, Giovanni Semeraro, and Marco Polignano. 2014. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In *Proceedings of the 8th International Workshop on Information Filtering and Retrieval Co-located with XIII AI/*IA Symposium on Artificial Intelligence (AIIA 2014)*, pages 59–68, Pisa.

Nakagawa, Tetsuji, Kentaro Tetsuji Inui, and Sadao Kurohashi. 2010. Dependency tree-based sentiment classification using CRFs with hidden variables. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794, Los Angeles, CA.

Narayanan, Ramanathan, Bing Liu, and Alok Choudhary. 2009. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 180–189, Singapore.

Naskar, Debashis, Sidahmed Mokaddem, Miguel Rebollo, and Eva Onaindia. 2016. Sentiment analysis in social networks through topic modeling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 46–53, Portoroz.

Negi, Sapna and Paul Buitelaar. 2015. Towards the extraction of customer-to-customer suggestions from reviews. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2159–2167, Lisbon.

Neviarouskaya, Alena and Masaki Aono. 2013. Sentiment word relations with affect, judgment and appreciation. *IEEE Transactions on Affective Computing*, 4(4):81–102.

Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. 2010a. @AM: Textual Attitude Analysis Model. In *Proceedings of the 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 80–88, Los Angeles, CA.

Neviarouskaya, Alena, Helmut Prendinger, and Mitsuru Ishizuka. 2010b. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 806–814, Beijing.

Neviarouskaya, Alena. 2010. *Compositional Approach for Automatic Recognition of Fine-Grained Affect, Judgment, and Appreciation in Text*. Ph.D. dissertation, University of Tokyo.

Ng, Vincentm, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pages 611–618, Sydney.

Nguyen, Thien Hai and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1354–1364, Beijing.

Ortony, Andrew, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge.

Osgood, Charles E. and Meredith Martin Richards. 1973. From yang and yin to *and* or *but*. *Language*, 49(2):380–412.

Osgood, Charles E., Georges J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana.

Palau, Raquel Mochales and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107, Barcelona.

Paltoglou, Georgios and Mike Thelwall. 2013. More than bag-of-words: Sentence-based document representation for sentiment analysis. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP 2013, pages 546–552, Hissar.

Pan, Sinno Jialin, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760, Raleigh, NC.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 271–278, Barcelona.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Barcelona.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86, Philadelphia, PA.

Peldszus, Andreas and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31.

Peldszus, Andreas and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, pages 938–948, Lisbon.

Perret, Jérémy, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, CA.

Persing, Isaac and Vincent Ng. 2014. Vote prediction on comments in social polls. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1127–1138, Doha.

Perugini, Marco and Richard P. Bagozzi. 2004. The distinction between desires and intentions. *European Journal of Social Psychology*, 34(1):69–84.

Picard, Rosemary W. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

Polanyi, Livia and Martin van den Berg. 2011. Discourse structure and sentiment. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, pages 97–102, Vancouver.

Polanyi, Livia and Annie Zaenen. 2006. Contextual valence shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Berlin, pages 1–10.

Popat, Kashyap, Balamurali A.R, Pushpak Bhattacharyya, and Gholamreza Haffari. 2013. The haves and the have-nots: Leveraging unlabelled corpora for sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 412–422, Sofia, Bulgaria.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 339–346, Vancouver.

Poria, Soujanya, Alexander Gelbukh, Amir Hussain, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38.

Potts, Christopher. 2010. On the negativity of negation. In Nan Li and David Lutz, editors, *Proceedings of Semantics and Linguistic Theory*, volume 20, pages 636–659, Ithaca, NY.

Prasad, Rashmi, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind K. Joshi, and Bonnie Webber. 2008. Penn Discourse Treebank version 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech.

Purver, Matthew. 2011. Topic segmentation. In Gokhan Tur and Renato De Mori, editors, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley, Chichester, pages 291–317.

Qadir, Ashequl, Ellen Riloff, and Marilyn Walker. 2015. Learning to recognize affective polarity in similes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 190–200, Lisbon.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, pages 1199–1204, Pasadena, CA.

Qiu, Minghui, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions and attributes for stance prediction in online debate forums. In *SIAM International Conference on Data Mining*, pages 855–863, Vancouver.

Rai, Piyush, Avishek Saha, Hal Daumé, III, and Suresh Venkatasubramanian. 2010. Domain adaptation meets active learning.

In *Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing*, pages 27–32, Los Angeles, CA.

Raksha, Sharma, Agarwal Astha, Gupta Mohit, and Bhattacharyya Pushpak. 2015. Adjective intensity and sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*, pages 2520–2526, Lisbon.

Ramanand, J., Krishna Bhavsar, and Niranjan Pedanekar. 2010. Wishful thinking: Finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 54–61, Los Angeles, CA.

Rao, Delip, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44, Toronto.

Read, Jonathon and John Carroll. 2012a. Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46:421–447.

Read, Jonathon and John Carroll. 2012b. Weakly supervised Appraisal analysis. *Linguistic Issues in Language Technology*, 8(2):1–21.

Rehbein, Ines, Merel Scholman, and Vera Demberg. 2015. Annotating discourse relations in spoken language: A comparison of the PDTB and CCR frameworks. In *Proceedings of the Workshop on Identification and Annotation of Discourse Relations in Spoken Language*, Saarbrücken.

Ren, Yong, Nobuhiro Kaji, Naoki Yoshinaga, and Masaru Kitsuregawa. 2014. Sentiment classification in under-resourced languages using graph-based semi-supervised learning methods. *IEICE Transactions on Information and Systems*, 97-D(4):790–797.

Ren, Yafeng, Yue Zhang, Meishan Zhang, and Donghong Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *Proceedings of the Thirtieth Conference on the American Association of Artificial Intelligence*, pages 215–221, Phoenix, AZ.

Reyes, Antonio and Paolo Rosso. 2012. Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4):754–760.

Reyes, Antonio, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

Ribeiro, Filipe Nunes, Matheus Araújo, Pollyanna Gonçalves, Fabrício Benevenuto, and Marcos André Gonçalves. 2016. A benchmark comparison of state-of-the-practice sentiment analysis methods. *ArXiv*, abs/1512.01818v4.

Rick, Scott and George Loewenstein. 2008. The role of emotion in economic behavior. In Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of Emotions*. Guilford, New York, pages 138–156, 3rd edition.

Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Sapporo.

Riloff, Ellen, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, WA.

Roberto, John A., Maria Salamó, and Maria Antònia Martí. 2015. Genre-based stages classification for polarity analysis. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*, pages 229–232, Hollywood, FL.

Rozin, Paul and Edward B. Royzman. 2001. Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5(4):296–320.

Ruppenhofer, Josef and Ines Rehbein. 2012. Semantic frames as an anchor representation for sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 104–109, Jeju Island.

Ruppenhofer, Josef, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the sources and targets of subjective expressions. In *Proceedings of the Sixth International Language Resources and Evaluation*, pages 2781–2788, Marrakech.

Ruppenhofer, Josef, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2014, pages 117–122.

Russell, James A. 1983. Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology*, 45:1281–1288.

Samdani, Rajhans and Wen-Tau Yih. 2011. Domain adaptation with ensemble of feature groups. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, IJCAI 2011, pages 1458–1464, Barcelona.

Saurí, Roser and James Pustejovsky. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3):227–268.

Scherer, Klaus R., Tanja Banziger, and Etienne Roesch. 2010. *A Blueprint for Affective Computing*. Oxford University Press, Oxford.

Shaikh, Mostafa Al, Helmut Prendinger, and Ishizuka Mitsuru. 2007. Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction*, ACII 2007, pages 191–202, Lisbon.

Sharma, Raksha, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP 2015, pages 2520–2526, Lisbon.

Shutova, Ekaterina and Simone Teufel. 2010. Metaphor corpus annotated for source-target domain mappings. In *Proceedings of the International Language Resources and Evaluation*, LREC 2010, pages 3255–3261, Malta.

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Shutova, Ekaterina. 2015. Design and evaluation of metaphor processing systems. *Computational Linguistics*, 41(4):579–623.

Smith, Phillip and Mark G. Lee. 2014. Acknowledging discourse function for sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, CICLing 2014, pages 45–52, Kathmandu.

Snyder, Benjamin and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 300–307, Rochester, NY.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161, Edinburgh, UK.

Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, WA.

Somasundaran, Swapna and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 226–234, Singapore.

Somasundaran, Swapna, Galileo Namata, Janyce Wiebe, and Lise Getoor. 2009. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 170–179, Singapore.

Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics in Human Language Technology*, pages 149–156, Edmonton.

Sperber, Dan and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Radical Pragmatics*, 49:295–318.

Spertus, Ellen. 1997. Smokey: Automatic recognition of hostile messages. In *Proceedings of the American Association of Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, pages 1058–1065, Providence, RI.

Stab, Christian and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha.

Stede, Manfred. 2011. *Discourse Processing*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Stone, Philip J., Bales Robert F., J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Systems Research and Behavioral Science*, 7(4):484–498.

Stoyanov, Veselin and Claire Cardie. 2008. Topic identification for fine-grained opinion analysis. In *Proceedings of the International Conference on Computational Linguistics*, pages 817–824, Manchester.

Strapparava, Carlo and Alessandro Valitutti. 2004. WordNet-Affect: An affective extension of WordNet. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1083–1086, Lisbon.

Su, Qi, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. 2008. Hidden sentiment association in Chinese Web opinion mining. In *Proceedings of the 17th International Conference on World Wide Web*, pages 959–968, Beijing.

Sujay, Cohan and Carlos M. Yalamanchi. 2012. Intention analysis for sales, marketing and customer service. In *International Conference on Computational Linguistics*, pages 33–40, Mumbai.

Sumath, Chiraag and Diana Inkpen. 2015. How much does word sense disambiguation help in sentiment analysis of micropost data? In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 115–121, Lisbon.

Surdeanu, Miahi, Thomas Hicks, and Marco Valenzuela-Escárcega. 2015. Two practical Rhetorical Structure Theory parsers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1–5, Denver, CO.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press.

Taboada, Maite and Jack Grieve. 2004. Analyzing Appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)*, pages 158–161, Palo Alto, CA.

Taboada, Maite and William C. Mann. 2006a. Applications of Rhetorical Structure Theory. *Discourse Studies*, 8(4):567–588.

Taboada, Maite and William C. Mann. 2006b. Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, 8:423–459.

Taboada, Maite and Radoslava Trnavac, editors. 2013. *Nonveridicality and Evaluation*. Brill, Leiden.

Taboada, Maite, Kimberly Voll, and Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. Technical report 2008-20, School of Computing Science.

Taboada, Maite, Julian Brooke, and Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 62–70, London, UK.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307.

Täckström, Oscar and Ryan McDonald. 2011. Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 569–574, Portland, OR.

Tan, Chenhao, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1397–1405, San Diego, CA.

Teufel, Simone and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical structure. *Computational Linguistics*, 28(4):409–445.

Thelwall, Mike, David Wilkinson, and Sukhvinder Uppal. 2010. Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, 61(1):190–199.

Thomas, Matt, Bo Pang, and Lillian Lee. 2006. 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*, pages 327–335, Sydney.

Thompson, Geoff and Laura Alba-Juez, editors. 2014. *Evaluation in Context*. John Benjamins.

Titov, Ivan and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 308–316, Columbus, OH.

Toprak, Cigdem, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584, Upsala.

Trivedi, Rakshit S. and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 808–813, Atlanta, GA.

Trnavac, Radoslava and Maite Taboada. 2012. The contribution of nonveridical rhetorical relations to evaluation in discourse. *Language Sciences*, 34 (3):301–318.

Tsur, Oren, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM-A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, pages 162–169, Washington, DC.

Turney, Peter D. and Michael L. Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical report EGB-1094, National Research Council Canada.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, PA.

Utsumi, Akira. 1996. A unified theory of irony and its computational formalization. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 962–967, Sydney.

Utsumi, Akira. 2004. Stylistic and contextual effects in irony processing. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, pages 1369–1374, Chicago, IL.

Van de Kauter, Marjan, Bart Desmet, and Véronique Hoste. 2015. The good, the bad and the implicit: A comprehensive approach to annotating explicit and implicit sentiment. *Language Resources and Evaluation*, 49(3):685–720.

Vanzo, Andrea, Danilo Croce, and Roberto Basili. 2014. A context-based model for sentiment analysis in Twitter. In *Proceedings of the International Conference on Computational Linguistics*, pages 2345–2354, Dublin.

Veale, Tony and Yanfen Hao. 2010. Detecting ironic intent in creative comparisons. In *Proceedings of the European Conference on Artificial Intelligence*, pages 765–770, Lisbon.

Velldall, Erik, Lilja Ovrelid, Jonathon Read, and Stephan Oepen. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38(2):369–410.

Vincent, Marc and Grégoire Winterstein. 2014. Argumentative insights from an opinion classification task on a French corpus. In *New Frontiers in Artificial Intelligence*, volume 8417 of *Lecture Notes in Computer Science*. Springer, Berlin, pages 125–140.

Vincze, Veronika, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: Biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Vo, Duy-Tin and Yue Zhang. 2015. Target-dependent Twitter sentiment classification with rich automatic features. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1347–1353, Buenos Aires, Argentina.

Voas, David. 2014. Towards a sociology of attitudes. *Sociological Research Online*, 19(1):12.

Volkova, Svitlana, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 186–196, Baltimore, MD.

Volkova, Svitlana, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015.

Inferring latent user properties from texts published in social media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, pages 4296–4297, Austin, TX.

Voll, Kimberly and Maite Taboada. 2007. Not all words are created equal: Extracting semantic orientation as a function of adjective relevance. In *Proceedings of the 20th Australian Joint Conference on Advances in Artificial Intelligence*, pages 337–346, Gold Coast.

Wachsmuth, Henning, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014. Modeling review argumentation for robust sentiment analysis. In *Proceedings of the International Conference on Computational Linguistics*, pages 553–564, Dublin.

Wallace, Byron C., Do Kook Choe, and Eugene Charniak. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 1035–1044, Beijing.

Wallace, Byron C. 2015. Computational irony: A survey and new perspectives. *Artificial Intelligence Review*, 43(4):467–483.

Walton, Douglas. 2009. Argumentation theory: A very short introduction. In Guillermo Simari and Iyad Rahwan, editors, *Argumentation in Artificial Intelligence*. Springer, Berlin, pages 1–22.

Wan, Xiaojun. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 235–243, Singapore.

Wang, Hao and Martin Ester. 2014. A sentiment-aligned topic model for product aspect rating prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Doha.

Wang, Jian and Yi Zhang. 2013. Opportunity model for e-commerce recommendation: Right product; right time. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2013, pages 303–312, Dublin.

Wang, Chunyan, Mao Ye, and Bernardo A. Huberman. 2012. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 244–252, Beijing.

Wang, Jinpeng, Gao Cong, Wayne Xin Zhao, and Xiaoming Li. 2015. Mining user intents in Twitter: A semi-supervised approach to inferring intent categories for tweets. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence*, pages 318–324, Austin, TX.

Wang, Po-Ya Angela. 2013. #irony or #sarcasm-a quantitative and qualitative study based on Twitter. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*, pages 349–356, Taipei.

Webber, Bonnie, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490.

West, Robert, Hristo S. Paskov, Jure Leskovec, and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.

White, Peter R. R. 2003. Beyond modality and hedging: A dialogic view of the language of intersubjective stance. *Text*, 23(2):259–284.

White, Peter R. R. 2004. Subjectivity, evaluation and point of view in media discourse. In Caroline Coffin, Ann Hewings, and Kay O'Halloran, editors, *Applying English Grammar: Corpus and Functional Approaches*. Arnold, London, pages 229–246.

White, Peter R. R. 2012. An introductory course in Appraisal analysis. http://languageofevaluation.info/appraisal/. [Accessed 28 June 2016].

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using Appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, pages 625–631, Bremen.

Wiebe, Janyce and Lingjia Deng. 2014. An account of opinion implicatures. *arXiv*, 1404.6491v1.

Wiebe, Janyce and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, volume 3406 of *Lecture Notes in Computer Science*, pages 486–497.

Wiebe, Janyce, Eric Breck, Chris Buckley, Claire Cardie, Paul Davis, Bruce Fraser,

Diane J. Litman, David R. Pierce, Ellen Riloff, Theresa Wilson, David Day, and Mark Maybury. 2003. Recognizing and organizing opinions expressed in the world press. In *Working Notes of the AAAI Spring Symposium in New Directions in Question Answering*, pages 12–19, Palo Alto, CA.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3):277–308.

Wiebe, Janyce, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2–3):165–210.

Wiebe, Janyce M. 1994. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287.

Wiegand, Michael and Dietrich Klakow. 2010. Convolution kernels for opinion holder extraction. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 795–803, Los Angeles, CA.

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala.

Wierzbicka, Anna. 1987. *Speech Act Verbs*. Academic Press, Sydney.

Wilson, Theresa and Janyce Wiebe. 2005. Annotating attributions and private states. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 53–60, Ann Arbor, MI.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence*, 22(2):73–99.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433.

Wilson, Theresas. 2008. Annotating subjective content in meetings. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2738–2745, Marrakech.

Winokur, Jon. 2005. *The Big Book of Irony*. St. Martin's Press, New York.

Wooldridge, Michael J. 2000. *Reasoning About Rational Agents*. MIT Press, Cambridge, MA.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1533–1541, Singapore.

Wu, Yuanbin, Qi Zhang, Xuanjing Huang, and Lide Wu. 2011. Structural opinion mining for graph-based sentiment representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1332–1341, Edinburgh.

Xia, Rui and Chengqing Zong. 2010. Exploring the use of word relation features for sentiment classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1336–1344, Beijing.

Xia, Rui, Chengqing Zong, Xuelei Hu, and Erik Cambria. 2015. Feature ensemble plus sample selection: Domain adaptation for sentiment classification (extended abstract). In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 4229–4233, Buenos Aires.

Xu, Kaiquan, Jiexun Li, and Stephen Shaoyi Liao. 2011. Sentiment community detection in social networks. In *Proceedings of the 2011 iConference*, pages 804–805, Seattle, WA.

Yang, Bishan and Claire Cardie. 2012. Extracting opinion expressions with semi-Markov conditional random fields. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1335–1345, Jeju Island.

Yang, Bishan and Claire Cardie. 2014. Context-aware learning for sentence-level sentiment analysis with posterior regularization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 325–335, Baltimore, MD.

Yang, Seon and Youngjoong Ko. 2011. Extracting comparative entities and predicates from texts using comparative

type classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1636–1644, Portland, OR.

Yang, Min, Wenting Tu, Ziyu Lu, Wenpeng Yin, and Kam-Pui Chow. 2015. LCCT: A semi-supervised model for sentiment classification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 546–555, Denver, CO.

Yano, Tae, Dani Yogatama, and Noah A. Smith. 2013. A penny for your tweets: Campaign contributions and Capitol Hill microblogging. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 737–740, Boston, MA.

Yessenalina, Ainur and Claire Cardie. 2011. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 172–182, Edinburgh.

Yessenalina, Ainur, Yisong Yue, and Claire Cardie. 2010. Multi-level structured models for document-level sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1046–1056, Cambridge, MA.

Yogatama, Dani and Noah Smith. 2014. Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *Proceedings of the 31st International Conference on Machine Learning*, pages 656–664, Beijing.

Yu, Hong and Hatzivassiloglou Vasileios. 2003. 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*, pages 129–136, Sapporo.

Zeng, Lingwei and Fang Li. 2013. A classification-based approach for implicit feature identification. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, volume 8202 of *Lecture Notes in Computer Science*, pages 190–202. Springer.

Zhang, Lei and Bing Liu. 2011. Identifying noun product features that imply opinions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 575–580, Portland, OR.

Zhao, Lili and Chunping Li. 2009. Ontology based opinion mining for movie reviews. In *Proceedings of the 3rd International Conference on Knowledge Science, Engineering and Management*, pages 204–214, Vienna.

Zhao, Wayne Xin, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, MA.

Zhou, Lanjun, Binyang Li, Wei Gao, Zhongyu Wei, and Kam-Fai Wong. 2011. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 162–171, Edinburgh.

Zhu, Xiaodan, Hongyu Guo, and Parinaz Sobhani. 2015. Neural networks for integrating compositional and non-compositional sentiment in sentiment composition. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 1–9, Denver, CO.

Zhuang, Li, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 43–50, Arlington,VA.

Zirn, Cäcilia, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In *the Fifth International Joint Conference on Natural Language Processing*, pages 336–344, Chian Mai.

Zwarts, Frans. 1995. Nonveridical contexts. *Linguistic Analysis*, 25(3/4):286–312.