Computational Intelligence, Volume 22, Number 2, 2006

RECOGNIZING STRONG AND WEAK OPINION CLAUSES

THERESA WILSON
Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260

JANYCE WIEBE AND REBECCA HWA
Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260

There has been a recent swell of interest in the automatic identification and extraction of opinions and emotions in text. In this paper, we present the first experimental results classifying the intensity of opinions and other types of subjectivity and classifying the subjectivity of deeply nested clauses. We use a wide range of features, including new syntactic features developed for opinion recognition. We vary the learning algorithm and the feature organization to explore the effect this has on the classification task. In 10-fold cross-validation experiments using support vector regression, we achieve improvements in mean-squared error over baseline ranging from 49% to 51%. Using boosting, we achieve improvements in accuracy ranging from 23% to 96%.

Key words: opinion recognition, subjectivity.

1. INTRODUCTION

In the past few years, interest in the automatic identification and extraction of attitudes, opinions, and sentiments in text has been growing rapidly. This task is motivated by the desire to provide tools and support for information analysts in government, commercial, and political domains, who want to automatically track attitudes and feelings in the news and online forums. How do people feel about recent events in the Middle East? Is the rhetoric from a particular opposition group intensifying? Is there a change in the attitudes being expressed toward the war in Iraq? A system that could automatically identify and extract opinions and emotions from text would be an enormous help to someone trying to answer these kinds of questions.

To date, the majority of work on subjectivity and sentiment analysis has focused on classification at the document or sentence level. Document classification tasks include, for example, distinguishing editorials from news articles and classifying reviews as positive or negative. A common sentence-level task is to classify sentences as subjective or objective. However, for many applications, simply recognizing which documents or sentences are opinionated may not be sufficient. Opinions vary in their *intensity*, and many applications would benefit from being able to determine not only if an opinion is being presented, but how strong is the opinion. Flame detection systems, for example, seek to identify strong rants and emotional tirades, while letting milder opinions pass through. Information analysts need tools that will help them to recognize changes over time in the virulence expressed by persons or groups of interest, and to detect when rhetoric is heating up, or cooling down.

A further challenge with automatic opinion identification is that it is not uncommon to find two or more opinions in a single sentence, or to find a sentence containing opinions as well as factual information. Information extraction (IE) systems are natural language processing (NLP) systems that extract from text any information relevant to a prespecified topic. An IE system trying to distinguish between factual information (which should be extracted) and non-factual information (which should be discarded or labeled uncertain) would benefit from the ability to pinpoint the particular clauses that contain opinions. This ability would also be important for multi-perspective question answering systems, which aim to present multiple answers to non-factual questions based on opinions derived from different sources, and for

multi-document summarization systems, which need to summarize differing opinions and perspectives.

With this article, our goal is to present research on the automatic classification of the *intensity* of opinions and emotions being expressed in text. Intensity classification includes detecting the absence of opinion, as well as detecting the strength of opinion. Thus, the intensity classification task subsumes the task of classifying language as subjective versus objective. An important aspect of the intensity classification task presented in this paper is that we focus not only on sentences, but on clauses within sentences. In this way, we begin to address the challenge of identifying when multiple opinions are being presented within a single sentence.

Our approach to this task is to use supervised machine learning techniques to train classifiers to predict the intensity of clauses and sentences. The learning algorithms use a large lexicon of *subjectivity clues*, summarized in Section 6. Subjectivity clues are words and phrases that may be used to express opinions and emotions. The clues in the lexicon are diverse. Many were learned automatically or collected from manual resources in previous studies of subjective language. The lexicon also contains new syntactic clues, which we introduce in this work. People use a staggering variety of words and phrases to express opinions. With the new syntactic clues, one goal is to capture common dependencies between words that may be relevant for recognizing intensity, such as intensifying adverbs modifying adjectives (e.g., *quite good* and *very bad*).

We want the learning algorithms to take full advantage of the subjectivity clues in the lexicon, but there are two major challenges. One is the sheer volume of clues; the other is that many of the words and phrases in the lexicon occur very infrequently. This raises the question of how best to organize the clues in the lexicon into features for the learning algorithms. The approach we use is to organize the clues into sets and to create one feature per set. Section 7 describes the two different methods we use for organizing clues into sets, and how features for the learning algorithms are defined based on these sets.

The data we use for both training and testing contain detailed annotations of the words and phrases being used to express opinions and emotions. These annotations are used to define the intensity of the sentences and clauses for the experiments. We perform 10-fold cross-validation experiments using three different learning algorithms: boosting, rule learning, and support vector regression. The experiments and their results are described in Section 8. We show that many clues of subjective language, including the new syntactic clues and those from the literature, can be adapted to the task of intensity recognition. We further show that the best results for intensity classification are achieved when the widest variety of clues is used.

The remainder of this article is organized as follows. Section 2 reviews the current state of research in subjectivity and sentiment analysis. Section 3 introduces the notion of *private state*, a general covering term that we use for opinions and emotions. Section 4 describes the corpus of opinion annotations used for the experiments in this work, as well as an inter-annotator agreement study that measures the reliability of the human intensity judgments in the corpus. Section 5 briefly describes the division of the annotated corpus into data sets for experiments. Section 6 describes the lexicon of subjectivity clues used for the intensity classification experiments, and Section 7 describes the feature organizations that are used. Related work and conclusions follow in Sections 9 and 10.

2. RESEARCH IN SUBJECTIVITY AND SENTIMENT ANALYSIS

Research in automatic subjectivity and sentiment analysis falls into three main areas. The first is identifying words and phrases that are associated a priori with subjectivity or

sentiment, for example, *believe*, which is associated with the expression of opinions, and *beautiful*, which is associated with positive sentiments (e.g., Hatzivassiloglou and McKeown 1997; Wiebe 2000; Kamps and Marx 2002; Turney 2002; Wiebe et al. 2004; Esuli and Sebastiani 2005).

The second area is identifying subjective language and its associated properties in context. This includes identifying expressions or sentences that are subjective in the context of a particular text or conversation (e.g., Riloff and Wiebe 2003; Yu and Hatzivassiloglou 2003; Nasukawa and Yi 2003; Popescu and Etzioni 2005), identifying particular types of attitudes (e.g., Gordon et al. 2003; Liu, Lieberman, and Selker 2003), recognizing the polarity or sentiment of phrases or sentences (e.g., Morinaga et al. 2002; Yu and Hatzivassiloglou 2003; Nasukawa and Yi 2003; Yi et al. 2003; Kim and Hovy 2004; Hu and Liu 2004; Popescu and Etzioni 2005; Wilson, Wiebe, and Hoffman 2005), identifying who is expressing an opinion (Kim and Hovy 2004; Choi et al. 2005), and identifying levels of attributions (e.g., that it is according to China that the U.S. believes something) (Breck and Cardie 2004).

The third area exploits automatic subjectivity analysis in NLP applications. Examples of such applications are recognizing inflammatory messages (Spertus 1997), tracking sentiment timelines in online discussions (Tong 2001), extracting investor sentiment from stock message boards (Das and Chen 2001), distinguishing editorials from news articles (e.g., Wiebe, Wilson, and Bell 2001; Yu and Hatzivassiloglou 2003), review classification (e.g., Turney 2002; Pang, Lee, and Vaithyanathan 2002; Morinaga et al. 2002; Dave, Lawrence, and Pennock 2003; Nasukawa and Yi 2003; Beineke, Hastie, and Vaithyanathan 2004; Mullen and Collier 2004; Kudo and Matsumoto 2004; Pang and Lee 2005; Whitelaw, Garg, and Argamon 2005), mining opinions from product reviews (e.g., Morinaga et al. 2002; Nasukawa and Yi 2003; Yi et al. 2003; Hu and Liu 2004; Popescu and Etzioni 2005), automatic expressive text-to-speech synthesis (Alm, Roth, and Sproat 2005), information extraction (e.g., Riloff, Wiebe, and Phillips 2005), and question answering (e.g., Yu and Hatzivassiloglou 2003; Stoyanov, Cardie, and Wiebe 2005).

There are a number of other topics in automatic subjectivity analysis that have yet to be explored. For example, very little work has been done on recognizing what an opinion is about, recognizing the boundaries of subjective expressions, summarizing opinions expressed in multiple documents, and extrapolating opinions from individuals to groups and vice versa.

3. PRIVATE STATES AND INTENSITY

Although we use the terms "opinion" and "emotion" for their accessibility, our research focuses not just on opinions and emotions, but also on speculations, evaluations, sentiments, beliefs, and other mental and emotional states. A general covering term for such states is *private state* (Quirk et al. 1985), an internal state that cannot be directly observed by others.

3.1. Expressing Private States

Subjective expressions are the words and phrases used to express private states (Wiebe, Wilson, and Cardie 2005). Different kinds of subjective expressions express private states in different ways. A subjective expression may explicitly mention a private state, as with "fears" in (1).

(1) "The U.S. fears a spillover," said Xirao-Nima.

Another kind of subjective expression expresses a mixture of speech and private state. For example, with "praised" in (2), we know that something was said and that a positive evaluation was expressed in what was said, even without the exact words being given.

(2) Italian senator Renzo Gubert praised the Chinese Government's efforts.

A subjective expression may indirectly express a private state, through the way something is described or through a particular wording. This is called an *expressive subjective element* (Banfield 1982). The phrase "full of absurdities" in (3) is an example of an expressive subjective element. Expressive subjective elements are used by people to express their frustration, anger, wonder, positive sentiment, mirth, etc., without specifically stating that they are frustrated, angry, etc. Sarcasm and irony often involve expressive subjective elements.

(3) "The report is full of absurdities," he said.

Finally, a subjective expression may describe a *private state action* (Wiebe 1994), as with "applaud" in (4). With private state actions, private states are expressed by direct physical actions. Booing someone, sighing heavily, shaking one's fist angrily, waving one's hand dismissively, smirking, and frowning are all examples of private state actions.

(4) As the long line of would-be voters marched in, those near the front of the queue began to spontaneously applaud those who were far behind them.

Private states are often expressed in speech events. Both "praised" in (2) and "said" in (3) are examples of speech events expressing private states. However, it is important to note that speech events may also be *objective* and not express a private state. For example, the speech event "said" in (5) does not express a private state.

(5) Sergeant O'Leary said the incident took place at 2:00 p.m.

3.2. Intensity

Intensity is a measure of the strength of a private state. When no private state is being expressed, intensity is *neutral*; however, a precise definition of what is a low-intensity private state or what is a high-intensity private state is more difficult. In spite of this, as language users, we intuitively perceive distinctions in the intensity levels of different private states. For example, *outraged* is a more intensely negative emotion than *annoyed*, *mildly outraged* is less intense than *outraged*, and *extremely annoyed* is more intense than *annoyed*.

4. A CORPUS OF OPINION ANNOTATIONS

For the experiments in this work, we use the Multi-Perspective Question Answering Opinion Corpus (MPQA Corpus).[1] The corpus is a collection of English-language versions of news documents from the world press. The documents are from 187 different news sources in a variety of countries. They date from June 2001 to May 2002. The corpus was collected and annotated as part of the summer 2002 Northeast Regional Research Center (NRRC) Workshop on Multi-Perspective Question Answering (Wiebe et al. 2003) sponsored by ARDA (Advanced Research and Development Activity).

The annotations in the MPQA Corpus are described in detail in Wiebe et al. (2005). Below, we review the aspects of the annotation scheme that are relevant for this work and present a new inter-annotator agreement study for intensity judgments.

[1]The MPQA Corpus used for the experiments in this article is available at nrrc.mitre.org/NRRC/publications.htm. A newer version of the corpus with contextual polarity judgments is available at www.cs.pitt.edu/mpqa/databaserelease.

Expressive Subjective Element Frame
- source
- text span
- intensity

Direct Subjective Frame
- source
- text span
- intensity
- expression intensity
- implicit

Objective Speech Event Frame
- source
- text span
- implicit

FIGURE 1. Annotation frames for private states and speech events in the MPQA Corpus. Attributes relevant for this work are listed.

4.1. Annotation Scheme

The MPQA Corpus contains annotations of private states and speech events, with speech events including both speaking and writing events. Private states expressed using expressive subjective elements are marked using *expressive subjective element* frames. *Direct subjective* frames are used to mark explicit mentions of private states, private state actions, mixtures of speech and private state, and other speech events in which private states are expressed in what is said. Speech events that do not express private states are marked using *objective speech event* frames.

Each annotation frame is characterized by a number of different attributes. Those relevant for this work are listed in Figure 1.

All annotation frames have a *source* and a *text span* attribute. The source is the experiencer of the private state or the speaker or writer of the speech event. Obviously, the writer of an article is a source, because he or she wrote the sentences composing the article, but the writer may also write about other people's private states and speech events, leading to multiple sources in a single sentence. For example, in sentence (1) above, there are three sources: the writer, Xirao-Nima (the speaker of the speech event "said"), and the United States (the experiencer of the private state "fears"). A key aspect of sources is that they are nested to capture the levels of attribution. In (1), the United States does not directly state its fear. Rather, according to the writer, according to Xirao-Nima, the United States fears a spillover. The full source of the private state expressed by "fears" is thus the *nested source* ⟨*writer, Xirao-Nima, United States*⟩.

The *text span* attribute is a pointer to the span of text that represents the private state or speech event, with one exception: speech events that are marked with the *implicit* attribute. Not all speech events are marked by discourse parentheticals, such as "he said" in (3). For example, every sentence is a speech event for the writer of the document, but there is no explicit phrase such as "I write" to mark the speech event. A similar situation can arise with direct quotations, such as with the second sentence in the following passage:

(6) "We think this is an example of the United States using human rights as a pretext to interfere in other countries' internal affairs," Kong said. "We have repeatedly stressed that no double standard should be employed in the fight against terrorism."

For these speech events, the speech event phrase is implicit, and the annotation is merely anchored to the sentence or quoted string with the text of the speech event.

The *intensity* attribute is used to mark the overall intensity of the private state that is represented by the direct subjective or expressive subjective element frame. The values are *low*, *medium*, *high*, and *extreme*.

For direct subjective frames, there is an additional intensity rating, the *expression intensity*. The *expression intensity* attribute represents the contribution to the intensity made specifically by the private state or speech event phrase. The values are *neutral*, *low*, *medium*, *high*, and *extreme*. For example, *say* is often neutral, even if what is uttered is not neutral, while *excoriate* by itself implies a very strong private state.

To help to clarify the differences among the various intensity attributes, Figure 2 gives the annotations for some of the example sentences used above. In sentence (1) there are three annotations: an objective speech event for the writer (because the writer wrote the sentence, and presents it as true that Xirao-Nima uttered the quoted string), an objective speech event frame for "said," and a direct subjective frame for "fears." The intensity of "fears" was marked as medium, as was the expression intensity. In sentence (2), there are two annotations: an objective speech event for the writer and a direct subjective frame for the speech event "praised." The intensity for "praised" is high, as is the expression intensity. Sentence (3) also has three annotations: an objective speech event for the writer, an expressive subjective element for "full of absurdities," and a direct subjective frame for the speech event "said." The intensity of "full of absurdities" was marked as medium. The intensity for "said" was also marked as medium. This is because for direct subjective frames, everything inside the scope of the speech event or private state, as well as the speech event or private state phrase itself, is considered when judging the overall intensity. The expression intensity for "said" is neutral: the word "said" does not itself contribute to the intensity of the private state.

4.2. Inter-Annotator Agreement for Intensity Judgments

This section reports on an agreement study that was conducted to measure the reliability of the various intensity judgments in the annotation scheme. For the experiments presented later in Section 8, we chose to merge the *high* and *extreme* intensity classes because of the rarity of the *extreme* class (only 2% of sentences in the corpus contain an annotation with extreme intensity). Thus, when calculating agreement, we also merge the *high* and *extreme* ratings, to mirror their treatment in the experiments.

Included in the judgment of intensity is a determination of whether a private state is being expressed at all. That is, when an annotator chooses to mark an expression as an objective speech event as opposed to a direct subjective annotation, the annotator is in essence making a judgment that intensity is *neutral*. Thus, to accurately measure agreement for intensity, the direct subjective and objective speech events must be considered together. The value of the intensity for all objective speech events is neutral. For all objective speech events that are not implicit, expression intensity is also neutral.

For the agreement study, three annotators (A, M, and S) independently annotated 13 documents with a total of 210 sentences. We first measure agreement for the intensity of expressive subjective elements and for the intensity and expression intensity of direct subjective and speech event frames. We then measure agreement for the overall intensity of sentences.

(1) "The U.S. fears a spillover," said Xirao-Nima.

Objective Speech:
source: ⟨*writer*⟩
implicit: true

Objective Speech:
source: ⟨*writer, Xirao-Nima*⟩
text span: said

Direct Subjective:
source: ⟨*writer, Xirao-Nima, U.S.*⟩
text span: fears
intensity: medium
expression intensity: medium

---

(2) Italian senator Renzo Gubert praised the Chinese Government's efforts.

Objective Speech:
source: ⟨*writer*⟩
implicit: true

Direct Subjective:
source: ⟨*writer, Gubert*⟩
text span: praised
intensity: high
expression intensity: high

---

(3) "The report is full of absurdities," Xirao-Nima said.

Objective Speech:
source: ⟨*writer*⟩
implicit: true

Expressive Subjective Element:
source: ⟨*writer, Xirao-Nima*⟩
text span: full of absurdities
intensity: medium

Direct Subjective:
source: ⟨*writer, Xirao-Nima*⟩
text span: said
intensity: medium
expression intensity: neutral

FIGURE 2. Example sentences (1), (2), and (3) with their private state and speech event annotations.

4.2.1. Agreement for Intensity and Expression Intensity. When measuring agreement, we first need to identify the units that will be evaluated. One thing that was not specified in the annotation scheme was a particular syntactic unit that the annotators should judge. A private state may be expressed by a single word (e.g., *odious*) or by a long phrase involving several syntactic constituents (e.g., *so conservative that it makes Pat Buchanan look vegetarian*). Therefore, it was decided that the annotators themselves would choose which words and phrases to annotate. Because of this, each annotator identified a different set of expressions. Thus, the first step in evaluating agreement for the intensity judgments made by a pair of annotators is to identify those expressions that were marked by both annotators. It is these expressions with *matching annotations* that are the units we evaluate.

For expressive subjective elements, matching annotations and their corresponding units of evaluation are determined based on overlapping text spans. For example, in (7) below, the expressive subjective elements for annotators A and M are marked in bold.

(7)
A: We applauded this move because it was **not only just**, but it made us begin to feel that we, as Arabs, were an **integral** part of Israeli society.
M: We applauded this move **because** it was **not only just**, **but** it made us begin to feel that we, as Arabs, were an **integral part** of Israeli society.

In (7), the expression "not only just" was marked by both annotators, making one unit for evaluation. Although their boundaries do not match exactly, "integral" and "integral part" do overlap; therefore, they are also considered matching annotations. Annotator M also marked "because" and "but" as expressive subjective elements. However, because there are no matching annotations marked by A, they cannot be considered when evaluating agreement for the intensity of expressive subjective elements.

For direct subjective and objective speech event frames that are not implicit (e.g., "fears," "said"), matching annotations are also determined based on overlapping text spans. For example, if one annotator marked "feel" in (7) as a direct subjective annotation and the other annotator marked "begin to feel" as a direct subjective annotation, these two annotations would be matching. Most sentences have only one direct subjective or objective speech event frame that is implicit (i.e., the one for the writer of the sentence). In the event that there is more than one implicit direct subjective or objective speech event frame in a sentence, the *source* attribute is used to determine which annotations are matching.

In Wiebe et al. (2005), we reported on the agreement for text spans for the same documents and annotations used in this study. Agreement for direct subjective and speech event text spans was 82%, and agreement for expressive subjective element text spans was 72%.

Now that the units for evaluation have been determined, the next issue is which metric should be used to evaluate inter-annotator agreement for intensity judgments. There are a number of considerations.

First, the classes used for intensity judgments are not discrete; instead, they represent an ordinal scale. For the combined direct subjective and objective speech event annotations, the rating scale for both intensity and expression intensity is *neutral*, *low*, *medium*, and *high*. For expressive subjective elements, the rating scale for intensity is *low*, *medium*, and *high*. Because intensity judgments are ordinal in nature, we do not want to treat all disagreements the same. A disagreement about whether intensity is neutral or high is more severe than a disagreement about whether it is medium or high. Thus, metrics such as Cohen's kappa (Cohen 1960), which treat all disagreements equally, are not appropriate.

There is an adaptation of Cohen's kappa called weighted kappa (Cohen 1968), which is for use with ordinal data. Weighted kappa assigns weights that allow for partial agreement. However, the weights are calculated based on the number of categories. In our case, expressive

subjective elements range in intensity from low to high, while the combined direct subjective and speech event annotations range in intensity from neutral to high. With weighted kappa, the weights for expressive subjective elements will be different than the weights for direct subjective and speech event annotations. Therefore, weighted kappa is also inappropriate.

The metric for agreement that we use is Krippendorff's α (Krippendorff 1980; Krippendorff 2004). Like kappa, Krippendorff's α takes into account chance agreement between annotators, but it is more general. It can be used to calculate agreement for both discrete and ordinal judgments, and its method of weighting disagreements does not depend on the number of categories. In its most general form, α is defined to be

$$\alpha = 1 - \frac{D_o}{D_e},$$

where D_o is a measure of the observed disagreement and D_e is a measure of the disagreement that can be expected by chance. Krippendorff's α ranges between 0 and 1, with $\alpha = 1$ indicating perfect agreement and $\alpha = 0$ indicating agreement that is no better than chance.

With α, a distance metric is used to weight disagreements. Different distance metrics are used for different types of data. For intensity, the ratings map naturally to the scale [0, 1, 2, 3], where 0 represents neutral and 3 represents high. Using this scale, we can use the distance metric that squares the difference between any two disagreements. Thus, the distance weight is 1 for any disagreement that differs by one (e.g., neutral-low), the distance weight is 4 for any disagreement that differs by two (e.g., neutral-medium), and the distance weight is 9 for any disagreement that differs by three (e.g., neutral-high).

Table 1 gives the pairwise α-agreement values for the intensity and expression intensity judgments of the combined direct subjective and speech event annotations. For comparison, the absolute percent agreement is also given. In interpreting α, Krippendorff (2004) suggests that values above 0.8 indicate strong reliability and values above 0.67 are enough for at least tentative conclusions. Using this scale, we see that the α scores for the intensity judgments of direct subjective and speech events are good.

Table 2 gives the pairwise α-agreement for the intensity of expressive subjective elements, along with absolute percent agreement for comparison. Unlike the agreement for the intensity judgments of direct subjective and speech event annotations, agreement for the intensity judgments of expressive subjective elements is not high. When we look at the disagreements, we find that many of them are influenced by differences in boundary judgments. Although annotations are considered matching as long as they have overlapping text spans, differences in boundaries can affect how intensity is judged. For example, expression (8) below shows how the same subjective expression was judged by two annotators.

TABLE 1. α-Agreement and Percent Agreement Scores for Intensity Judgments for the Combined Direct Subjective and Objective Speech Annotations

	Intensity		Expression Intensity	
Annotator Pair	α	%	α	%
A & M	0.79	0.73	0.76	0.66
A & S	0.81	0.75	0.76	0.63
M & S	0.76	0.76	0.73	0.59
Average	0.79	0.75	0.75	0.62

TABLE 2. α-Agreement and Percent Agreement Scores for Expressive Subjective Element Intensity Judgments

	Intensity	
Annotator Pair	α	%
A & M	0.40	0.49
A & S	0.52	0.56
M & S	0.46	0.54
Average	0.46	0.53

TABLE 3. α-Agreement and Percent Agreement Scores for Sentence-Level Intensity Judgments

	Intensity	
Annotator Pair	α	%
A & M	0.74	0.56
A & S	0.83	0.73
M & S	0.72	0.57
Average	0.77	0.62

(8)
A: <*high*>imperative for harmonious society</>
M: <*medium*>imperative</> for <*medium*>harmonious</> society

Both annotators recognized that the above phrase is subjective. However, while the first annotator marked the entire phrase as a single expressive subjective element with high intensity, the second annotator marked particular words and smaller phrases as expressive subjective elements and judged the intensity of each separately.

A severe type of disagreement between annotators would be a difference in intensity ordering, that is, annotator A rating expression 1 more intense than expression 2, and annnotator B rating expression 2 more intense than expression 1. Fortunately, there are few such disagreements. On average, only 5% of all possible pairings of matching annotations result in disagreements in the ordering of intensity.

4.2.2. Agreement for Intensity of Sentences. Although intensity judgments were not made at the sentence level, sentence-level judgments can be derived from the expression-level intensity judgments. In this section, we measure agreement for those judgments.

Evaluating intensity agreement at the sentence level is important for two reasons. First, annotations that were previously excluded from consideration because they were identified by only one annotator may now be included. Second, in the experiments in Section 8, the units of evaluation are sentences and clauses, and it is important to know what the agreement is for intensity judgments at this higher level.

An annotator's intensity judgment for a sentence is defined as the highest intensity or expression-intensity rating of any annotation marked by that annotator in the sentence. Pairwise agreement scores for sentence-level intensity judgments are given in Table 3. The average α-agreement for sentences is 0.77.

TABLE 4. Sample of Subjective Expressions with High- and Extreme-Intensity Ratings

victory of justice and freedom	such a disadvantageous situation
will not be a game without risk	breeding terrorism
grown tremendously	menace
such animosity	not true at all
throttling the voice	imperative for harmonious society
tainted with a significant degree of hypocrisy	power at all costs
so exciting	glorious
violence and intimidation	disastrous consequences
could not have wished for a better situation	did not exactly cover himself in glory
freak show	exalted
if you're not with us, you're against us	the embodiment of two-sided justice
vehemently denied	appalling
everything good and nice	very definitely
under no circumstances	diametrically opposed
justice-seeking cries	shameful mum
powder keg	purposes of intimidation and exaggeration
most fraudulent, terrorist and extremist	should be an eye-opener for the whole world
number one democracy	enthusiastically asked
apocalyptic savagery	hate
odious	gross misstatement
indulging in blood-shed and their lunaticism	increasingly tyrannical
many absurdities, exaggerations, and fabrications	surprised, to put it mildly
take justice to prehistoric times	disdain and wrath
lost the reputation of commitment to principles of human justice	great fanfare
	unconditionally and without delay
ultimately the demon they have reared will eat up their own vitals	those digging graves for others, get engraved themselves
	so conservative that it makes Pat Buchanan look vegetarian

4.3. Exploring Intensity

An examination of a portion of the annotated data held out for development shows not only that an extreme variety of expressions have been marked, but that higher-intensity private states in particular are expressed in many different ways. Table 4 gives a sample of some subjective expressions with high and extreme intensity. Of course there are obvious words that almost always express more intense private states, such as "exciting" and "hate." These are easy to list, as are some obvious modifications that increase or decrease their intensity: "**very** exciting," "**really** hate," and "**don't** hate." However, it is unlikely that expressions such as "powder keg," "freak show," "prehistoric," and "tyrannical" readily come to mind, all of which are marked in the MPQA Corpus.

Higher-intensity expressions often contain words that are very infrequent. For example, the words "prehistoric," "tyrannical," and "lunaticism" each appear only once in the corpus. Because subjective words are often less frequent (Wiebe et al. 2004), it is important to have knowledge of patterns such as "expressed <direct-object>," which can generalize to many different phrases, such as "expressed hope," "expressed concern," "expressed gratitude," and

"expressed some understanding." Collocations such as "at all" add punch to an expression, as in, "at all costs" and "not true at all." There are also syntactic modifications and syntactic patterns that have subjective force. In addition to those patterns that merely intensify a subjective word, for example, "very <ADJECTIVE>," we find patterns that have a cumulative effect on intensity: "justice and freedom," "terrorist and extremist," "violence and intimidation," "exaggerations and fabrications," and "disdain and wrath." The clues used later in the intensity classification experiments contain examples of all these kinds of subjective phenomena.

Sentences in which private states are expressed are often complex, with subjective expressions of differing intensities being expressed by perhaps two or more agents. This is the case in (9) below.

> (9) President Mohammad Khatami of Iran, whose attempt at reforms have gotten American <low>support</>, <high>accused</> the United States of "<high>warmongering</>."

In this sentence, there is low-intensity support being expressed by the United States, as well as high-intensity negative accusations coming from Khatami. In the MPQA Corpus, 31% of sentences are made up of clauses that differ in intensity by two or more intensity ratings. This highlights the need to identify opinions at the clause level, as we do in the experiments below.

Many researchers have argued for the importance of recognizing polarity, that is, whether a positive or negative sentiment is being expressed. Such knowledge is needed to determine whether a favorable or unfavorable opinion is being expressed in a review, or to determine whether support for a policy or idea is being expressed in an editorial. We find some interesting interactions between polarity and intensity in the data. The annotators were asked to judge the polarity of expressions that they marked, using an *attitude-type* attribute, which has values *positive*, *negative*, and *other*. The annotations show that a number of *attitude-type* labels are neither *positive* or *negative*: 22% are *other*. However, the annotations also reveal that the stronger the expression, the clearer the polarity. Only 8% of the high-intensity annotations are marked as *other*, while 39% of the low-intensity annotations are so marked. In addition to stronger expressions having clearer polarity, stronger expressions of opinions and emotions also tend to be more negative in this corpus. Only 33% of low-intensity annotations are negative, compared to 78% of high-intensity annotations. These observations lead us to believe that the intensity of subjective expressions will be informative for recognizing polarity, and vice versa.

5. DATA SETS

For the experiments below, the documents in the MPQA Corpus are divided into two data sets. The first data set (66 documents/1,344 sentences) is a development set, used for data exploration, feature development, and parameter tuning. The second data set (469 documents/9,313 sentences) is an evaluation set, used to identify and evaluate the new syntactic clues presented below in Section 6.2 and in the experiments in Section 8 The sentences in the evaluation set are further divided into 10-folds, which are used to define training and testing sets for cross-validation.

6. SUBJECTIVITY CLUES

In this section, we describe the knowledge that we use for automatic intensity classification, namely, a broad collection of *subjectivity clues*. Subjectivity clues are words and

phrases that may be used to express private states. In other words, they have subjective usages, although they may have objective usages as well.

First, we review the wide variety of clues in our established subjectivity lexicon, which was developed through previous work. Then we introduce a collection of new syntactic clues that are correlated with subjective language.

6.1. Previously Established Types of Clues

Through previous work in subjectivity identification, we have developed a large lexicon of subjectivity clues, which we will refer to as PREV clues. The PREV clues include words and phrases culled from manually developed resources and learned from annotated and unannotated data. An interesting aspect of the set of PREV clues is that, because of the wide variety of sources from which they were compiled, the lexicon is quite varied and not limited to a fixed word list or to words of a particular part of speech.

The clues from manually developed resources include:

- Verbs of judgment (e.g., *commend*, *reprove*, *vilify*), desire (e.g., *fancy*, *pine*, *want*), and psych (e.g., *dread*, *love*, *vex*) from Levin's (1993) English verb classes.
- Words and phrases culled from Ballmer and Brennenstuhl's (1981) speech act verb classes (e.g., *advocate*, *grumble about*, *vow*).
- Verbs and adjectives listed in FrameNet (Baker, Fillmore, and Lowe 1998) with frame element *experiencer*. These include words from the Emotion_active (e.g., *fuss*, *worry*), Emotion_directed (e.g., *pleased*, *upset*), Emotion_heat (e.g., *burn*, *seethe*), Experiencer_obj (e.g., *embarrass*, *thrill*), Experiencer_subj (e.g., *dislike*, *sympathize*), and Perception_body (e.g., *ache*, *tickle*) frames.
- Adjectives manually annotated for polarity from Hatzivassiloglou and McKeown (1997). The list of Positive adjectives includes the words *appealing*, *brilliant*, *luxurious*, and *nifty*. Included in the list of negative adjectives are the words *bizarre*, *dismal*, *hypocritical*, and *tedious*.
- Subjectivity clues listed in Wiebe (1990) (e.g., *absurdly*, *funny*, *grin*, *stench*, *truly*, *wonder*).

Clues learned from annotated data include distributionally similar adjectives and verbs, and n-grams from Wiebe et al. (2004). The adjectives and verbs were learned from *Wall Street Journal* (WSJ) data using Dekang Lin's (1998) method for clustering words according to their distributional similarity. The seed words for this process were the adjectives and verbs in editorials and other opinion-piece articles. The n-gram clues were learned from WSJ data annotated for subjective expressions. They range from two to four words in length. Some examples of 3-grams are *worst of all*, *of the century*, and *do something about*. Examples of 4-grams are *on the other hand* and *price you have to*.

From unannotated data, extraction patterns and subjective nouns were learned using two different bootstrapping algorithms and a set of seed words (Riloff and Wiebe 2003; Riloff, Wiebe, and Wilson 2003). Extraction patterns are lexico-syntactic patterns typically used by information extraction systems to identify relevant information. For example, the pattern *<subject> was hired* would apply to sentences that contain the verb *hired* in the passive voice and would extract the subject as the hiree. In Riloff and Wiebe (2003), AutoSlog-TS, an algorithm for automatically generating extraction patterns, is used to find extraction patterns that are correlated with subjectivity. An example of a subjective extraction pattern is *<subj> dealt blow*, which matches phrases such as "the mistake dealt a stiff blow to his pride." In Riloff et al. (2003), the Meta-Bootstrapping (Riloff and Jones 1999) and Basilisk

(Thelen and Riloff 2002) bootstrapping algorithms were used to learn sets of subjective nouns.

Finally, although not explicitly part of the lexicon, low-frequency words, which are informative for subjectivity recognition and require no training to identify (Wiebe et al. 2004), are also used as clues. A word is considered to be of low frequency if it appears $\leq$3 times in the document containing it plus a one-million word corpus of news articles. In addition, we use n-gram clues from Wiebe et al. (2004) that have fillers matching low-frequency words. When these clues were learned, the fillers matched low-frequency words in the training data. When used during testing, the fillers are matched against low-frequency words in the test data. Examples of such n-grams are *<LowFreq-verb> and <LowFreq-verb>*, matching the phrases *bleat and bore* and *womanize and booze*, and *so <LowFreq-adj>*, matching the phrases *so enthusiastic* and *so cumbersome*.

Most of the above clues were collected as part of the work reported in Riloff, Wiebe, and Wilson (2003).

6.2. Syntax Clues

The new syntactic clues (SYNTAX clues) are developed by using, a mostly supervised, learning procedure. The training data are based on both the annotations in the MPQA Corpus and a large unannotated corpus of automatically identified subjective and objective sentence from Riloff and Wiebe (2003). The learning procedure consists of three steps.

First, we parse the training sentences in the MPQA Corpus with a broad-coverage lexicalized English parser (Collins 1997). The output constituent trees are automatically converted into their dependency representations (Hwa and Lopez 2004). In a dependency representation, every node in the tree structure is a surface word (i.e., there are no abstract nodes such as NP or VP), but each word may have additional attributes such as its part-of-speech (POS) tag. The parent word is known as the *head*, and its children are its *modifiers*. The edge between a parent and a child node specifies the grammatical relationship between the two words (e.g., *subj*, *obj*, and *adj*). Figure 3 shows the dependency parse tree for a sentence, along with the corresponding constituent representation, for comparison. For this study, we use 48 POS tags and 24 grammatical relationships.

Next, for each word in every dependency parse tree, we exhaustively generate potential syntactic clues. There are five classes of syntactic clues. For each class, we generate clues that include specific words (indicated with **lex**) as well as less specific variants that back off to only POS tags (indicated with **backoff**).

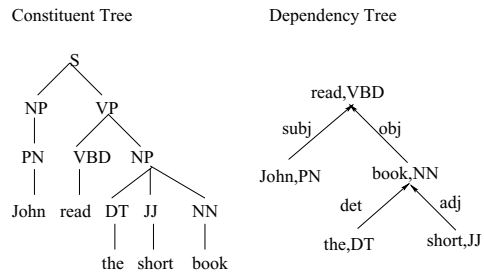


FIGURE 3. The constituent tree for "People are happy because Chavez has fallen" is on the left, and the dependency representation is on the right.

root

root-lex(w, t): word w with POS tag t is the root of a dependency tree (i.e., the main verb of the sentence).

root-backoff(t): a word with POS tag t is the root of a dependency tree.

leaf

leaf-lex(w, t): word w with POS tag t is a leaf in a dependency tree (i.e., it has no modifiers).

leaf-backoff(t): a word with POS tag t is a leaf in a dependency tree.

node

node-lex(w, t): word w with POS tag t.

node-backoff(t): a word with POS tag t.

bilex

bilex-lex(w, t, r, w_c, t_c): word w with POS tag t is modified by word w_c with POS tag t_c, and the grammatical relationship between them is r.

bilex-backoff(t, r, t_c): a word with POS tag t is modified by a word with POS tag t_c, and the grammatical relationship between them is r.

allkids

allkids-lex(w, t, r_1, w_1, t_1, ..., r_n, w_n, t_n): word w with POS tag t has n children. Each child word w_i has POS tag t_i and modifies w with grammatical relationship r_i, where $1 \leq i \leq n$.

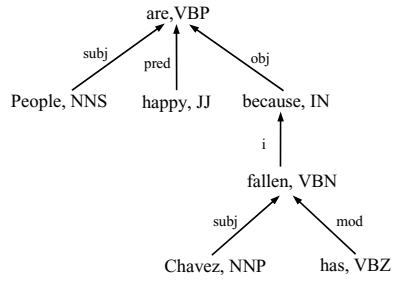
allkids-backoff(t, r_1, t_1, ..., r_n, t_n): a word with POS tag t has n children. The ith child word has POS tag t_i and modifies the parent word with grammatical relationship r_i.

One thing that can determine the intensity of a private state being expressed is the patterning of a word together with its modifiers. For example, in the phrase *really quite nice*, the adverbial modifiers "really" and "quite" are working to intensify the positive evaluation expressed by "nice." With the *allkids* clues, our aim was to try to capture these types of patterns. One problem with the *allkids* clues, however, is the sparsity of their occurrences. This led us to include the *bilex* clues, which focus on the patterning found between a word and just one of its modifiers.

Examples of the different classes of syntactic clues are given in Figure 4. The top of Figure 4 gives the dependency parse tree for the sentence, *People are happy because Chavez has fallen*. The bottom half of the figure lists the potential syntactic-lex clues that would be generated for the sentence.

Finally, we evaluate the potential clues to determine which clues to retain for later experiments. A clue is considered to be *potentially useful* if more than x% of its occurrences are in subjective expressions in the training data, where x is a parameter tuned on the development set. For our experiments, we chose $x = 70\%$. Potentially useful clues are further categorized into one of three *reliability* levels. First, a clue is considered *highly reliable* if it occurs five or more times in the training data. For those that occur fewer than five times, we check their reliability on the larger corpus of automatically identified subjective and objective sentences. Clues that do not occur in the larger unannotated corpus are considered *not very reliable*. Clues that occur in the subjective set at least y times more than in the objective set are considered *somewhat reliable*. The parameter y is tuned on the development set and is set to 4 in our experiments. The remaining clues are rejected as not useful.

After filtering the potential syntax clues, 16,168 are retained on average per fold: 6.1% highly reliable, 42.9% somewhat reliable, and 51% not very reliable. Table 5 gives the distribution of clues based on type and reliability level. Table 6 gives a few examples of *allkids-backoff* clues from the different reliability levels.



root(are,VBP)
leaf(people,NNS)
leaf(happy,JJ)
leaf(chavez,NNP)
leaf(has,VBZ)
node(are,VBP)
node(people,NNS)
node(happy,JJ)
node(because,IN)
node(fallen,VBN)
node(chavez,NNP)
node(has,VBZ)
bilex(are,VBP,people,NNS,subj)
bilex(are,VBP,happy,JJ,pred)
bilex(are,VBP,because,IN,obj)
bilex(because,IN,fallen,VBN,i)
bilex(fallen,VBN,chavez,NNP,subj)
bilex(fallen,VBN,has,VBZ,mod)
allkids(are,VBP,people,NNS,subj,happy,JJ,pred,because,IN,pred)
allkids(because,IN,fallen,VBN,i)
allkids(fallen,VBN,chavez,NNP,subj,has,VBZ,mod)

FIGURE 4. Dependency parse tree and potential syntax-lex clues that would be generated from the tree for the sentence "People are happy because Chavez has fallen."

TABLE 5. Distribution of Retained Syntax Clues by Type and Reliability Level*

	Reliability Level		
Type	Highly Reliable	Somewhat Reliable	Not Very Reliable
root	0.2	0.6	0.6
leaf	0.6	2.5	2.1
node	2.1	5.9	4.0
bilex	3.1	32.8	41.8
allkids	0.2	1.2	2.5

*The values in the table are in percentages.

7. FEATURE ORGANIZATION

Given the large number of PREV and SYNTAX clues, we are faced with the question of how best to organize them into features for intensity classification. A feature representation in which each clue is treated as a separate feature was tried, but this gave poor results. We believe this is because so many of the individual clues are of low frequency. Of the PREV

TABLE 6. Examples of *Allkids-Backoff* Clues from the Different Reliability Levels and Instances that They Match in the Corpus*

Highly reliable	
CC,RB,mod,JJ,conj,JJ,conj	very precious *and* (very) sophisticated
	awfully grave *and* pressing
	only natural *and* rational
	quite neat *and* tidy
	thoroughly disgraceful *and* unacceptable
VB,DT,subj,JJ,pred	this *was* effective
	this *was* essential
	this *is* crazy
	those (who want to devalue) *are* shameless
	this *is* (no) different
Somewhat reliable	
CC,JJR,conj,NN,conj,NN,conj	better governance *and* democracy
	greater speed *and* strength
WRB,JJ,adj,VB,i	*how* good (they) were
	how long (it can still) justify
	(no matter) *how* cynical (this may) appear
Not very reliable	
VB,MD,mod,RP,mod,NN,obj,PREP,p	would *turn* back (the) clock on
WRB,NN,amod,VB,i	*where* (the) hell (it) is

*In the instances, the word that is being modified is in italics, and words that are not its direct modifiers are in parentheses.

clues with instances in the corpus, 32% only occur once and an additional 16% occur twice. With the SYNTAX clues, a full 57% have a frequency of 1. Instead of treating each clue as a separate feature, we adopt the strategy of aggregating clues into sets and creating one feature for each set (Cohen 1996; Wiebe, McKeever, and Bruce 1998). The value of each feature is the number of instances in the sentence or clause of all the members of the set. The motivation for this type of organization is twofold. First, it increases the probability that a feature in the test set would have been observed in the training data: even if a clue in the test set did not appear in the training data, other members of that clue's set may have appeared in the training data. Second, because clues are aggregated, feature frequencies are higher. We experiment with two strategies for aggregating clues into sets: organizing clues by their type and organizing clues by their intensity.

7.1. Organizing Clues by Type

To organize clues by their type, we define 29 sets for the PREV clues and 15 sets for the SYNTAX clues. The sets created for the PREV clues reflect how the clues were presented in the original research. For example, there are three sets created for the three classes of Levin (1993) verbs, and there are two sets created for the polar adjectives from Hatzivassiloglou and McKeown (1997), one for the positive adjectives and one for the negative adjectives. The SYNTAX clues are aggregated into sets based on the class of clue and reliability level. For example, highly reliable *bilex* clues form one set; somewhat-reliable *node* clues form another set.

In the experiments below, when features are used that correspond to sets of clues organized by type, they are referred to as TYPE features.

7.2. Organizing Clues by Intensity

Although the sets of subjectivity clues being used were selected because of their correlation with subjective language, they are not necessarily geared to discriminate subjective language of differing intensities. Also, the groupings of clues into sets was not done with intensity in mind. We hypothesized that a feature organization that takes into consideration the potential intensity of clues would be better for intensity classification.

To adapt the clues for intensity classification, we use the annotations in the training data to filter the clues and organize them into four new sets, one for each intensity rating. Clues are placed into sets based on intensity as follows: For each clue c and intensity rating s, calculate $P(intensity(c) = s)$, the probability of c being in a subjective expression with intensity s. For $s = neutral$, this is the probability of c being in the text span of an objective speech event, in the text span of a direct subjective annotation with neutral expression-intensity, or in no annotation at all. Then, if $P(intensity(c) = s) \geq T(s)$, where $T(s)$ is the threshold determined for intensity s, place c in the set of clues with intensity s. In our experiments, we set $T(s) = P(intensity(word) = s) + 0.25$ or 0.95 if $P(intensity(word) = s) + 0.25 \geq 1$. $P(intensity(word) = s)$ is the probability of any given word being in a subjective expression with intensity s. The value 0.25 was determined using experiments on the development set. Note that with this method of organizing clues into sets, it is possible for a clue to be in more than one set.

In the experiments below, when features are used that correspond to sets of clues organized by intensity, they are referred to as INTENSITY features.

8. EXPERIMENTS IN INTENSITY CLASSIFICATION

It is important to classify the intensity of clauses as well as sentences, but pinpointing subjectivity at deeper levels can be challenging because there is less information to use for classification. To study the feasibility of automatically classifying clauses by their intensity, we conducted a suite of experiments in which an intensity classifier is trained based on the features previously described. We wished to confirm three hypotheses. First, it is possible to classify the intensity of clauses, for those that are deeply nested as well as for those at the sentence level. Second, classifying the intensity of subjectivity depends on a wide variety of features, including both lexical and syntactic clues. Third, a feature organization based on intensity is beneficial.

To test our hypotheses, we performed the experiments under different settings, varying four factors: (1) the learning algorithm used to train the classifiers, (2) the depth of the clauses to be classified, (3) the types of clues used, and (4) the feature organization (TYPE vs. INTENSITY). We vary the learning algorithm to explore its effect on the classification task. In our studies, the three machine learning algorithms are boosting, rule learning, and support vector regression. For boosting, we use BoosTexter (Schapire and Singer 2000) AdaBoost.HM with 1,000 rounds of boosting. For rule learning, we use Ripper (Cohen 1996). For support vector regression we use SVMlight (Joachims 1999) and discretize the resulting output into the ordinal intensity classes. These algorithms were chosen because they have successfully been used for a number of natural language processing tasks.

In the sections below, we first describe how clauses are determined, and how the gold-standard intensity classes are defined for sentences and clauses. We then describe the training-testing setup used for the experiments, followed by the experimental results.

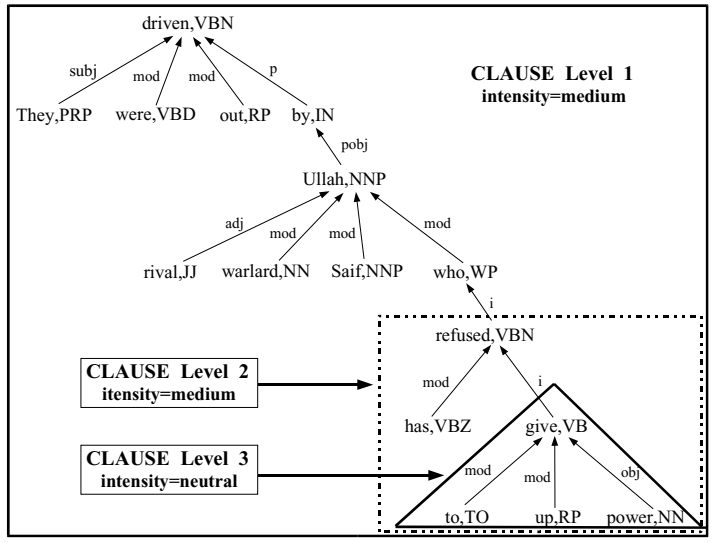


FIGURE 5. Dependency parse tree for the sentence "They were driven out by rival warlord Saif Ullah, who has refused to give up power." Clause level 1 is the whole sentence, clause level 2 is the subtree headed by "refused," and clause level 3 is the subtree headed by "give."

8.1. Determining Clauses and Defining the Gold Standard

Clauses were chosen as our unit of evaluation because they can be determined automatically, and because they give us different levels of nesting to vary in our experiments. Clauses are determined based on the non-leaf verbs in the parse tree, parsed using the Collins parser and converted to the dependency representation as described earlier in Section 6.2. For example, sentence (10) has three clauses, corresponding to the verbs "driven," "refused," and "give."

(10) They were driven out by rival warlord Saif Ullah, who has refused to give up power.

The parse tree for sentence (10) is given in Figure 5. The clause defined for "driven" (level 1) is the entire sentence; the clause for "refused" (level 2) is "has refused to give up power"; and the clause for "give" (level 3) is "to give up power." Determining clauses in this way results in 9,817 level-2 clauses, 6,264 level-3 clauses, and 2,992 level-4 clauses in the experiment data set.

The gold-standard intensity ratings of sentences and clauses are based on individual expression annotations: the intensity of a sentence or clause is defined as the highest intensity rating of any expression in that sentence or clause. For example, in sentence (10), "refused" is the annotation with the highest intensity in the sentence. It was marked as a direct subjective annotation with medium intensity. Thus, the level-1 clause (the entire sentence, headed by "driven") and the level-2 clause (headed by "refused") both have a gold-standard intensity of medium. However, the gold-standard intensity for the level-3 clause (headed by "give") is neutral, because the annotation for "refused" lies outside of the clause and there are no other annotations within the clause.

8.2. Experimental Setup

In setting up experiments for classifying nested clauses, there is a choice to be made for training: (1) either clauses from the same nested level may be used for training, or (2) clauses from a different level may be used for training. In the experiments in this article, the training examples are always entire sentences, regardless of the clause level being classified during testing. Experimental results showed that this configuration is better than training on clauses from the same level. We believe this is because whole sentences contain more information.

8.3. Classification Results

All results reported are averages over 10-fold cross-validation experiments using the 9,313 sentences from the experiment data set. Significance is measured using a one-tailed t-test. For each experiment, both mean-squared error (MSE) and classification accuracy are given. Although raw accuracy is important, it treats a misclassification that is off by 1 the same as a misclassification that is off by 3. As with disagreements in annotator intensity judgments, treating all intensity misclassifications equally does not reflect the ordinal nature of the intensity classes. MSE captures this distinction, and, for this task, it is perhaps more important than accuracy as a metric for evaluation. If t_i is the true intensity of sentence i, and p_i is the predicted intensity of sentence i,

$$\text{MSE} = \frac{1}{n}\sum_{i}^{n}(t_i - p_i)^2,$$

where n is the number of sentences or clauses being classified. Note that the distance metric used in the α-agreement score from Section 4.2 is the same as MSE.

Table 7 gives the baselines and the results for experiments using all clues (PREV and TYPE) as well as experiments using BAG. The question of what to use for a baseline is not straightforward. A common strategy is to use a baseline classifier that always chooses the most frequent class. However, the most frequent class for sentences is medium, which is different than the most frequent class for nested clauses, neutral. Thus, in Table 7 we chose to give both baselines, one for a classifier that always chooses neutral, and one for a classifier that always chooses medium. Note that there is quite a difference between the performance of the baselines with respect to MSE and accuracy. Because medium is closer to the midpoint on the intensity scale that we are using, the medium-class baseline performs better for MSE. The neutral-class baseline, on the other hand, performs better for accuracy, except for at the sentence level.

In Table 7, results for the same five experiments are given for each of the three classification algorithms. The experiments differ in which features and feature organizations are used. Experiment (1) in the table uses BAG, where the words in each sentence are given to the classification algorithm as features. Experiments (2) and (3) use all the subjectivity clues described in Section 6. For experiment (2), the TYPE organization is used; for experiment (3), the INTENSITY organization is used. For experiments (4) and (5), BAG is used along with the subjectivity clues in their two different feature organizations.

The results for intensity classification are promising for clauses at all levels of nesting. For BoosTexter, all experiments result in significant improvements over the two baselines, as measured by both MSE and accuracy. The same is true for Ripper, with the exception of experiment (1), which uses only BAG and none of the subjectivity clue features. For SVMlight, at the sentence level (clause level 1), all experiments also result in significant improvements over the baselines for MSE and accuracy. For the nested clause levels, all

TABLE 7. Intensity Classification Results for Experiments Using All Subjectivity Clues as Well as Bag-of-Words (BAG)*

	Level 1		Level 2		Level 3		Level 4	
Baselines	MSE	Acc	MSE	Acc	MSE	Acc	MSE	Acc
Neutral-class	3.603	28.1	2.752	41.8	2.539	45.9	2.507	48.3
Medium-class	1.540	30.4	2.000	25.4	2.141	23.7	2.225	22.5
BoosTexter								
(1) BAG	1.234	50.9	1.390	53.1	1.534	53.6	1.613	53.0
(2) TYPE	1.135	50.2	1.267	53.4	1.339	54.7	1.410	55.5
(3) INTENSITY	1.060	54.1	1.180	56.9	1.258	**57.9**	1.269	**60.3**
(4) BAG + TYPE	1.069	52.0	1.178	54.8	1.267	55.9	1.321	56.8
(5) BAG + INTENSITY	**0.991**	**55.0**	**1.111**	**57.0**	**1.225**	57.5	**1.211**	59.4
Ripper								
(1) BAG	1.570	34.5	1.961	29.2	2.091	27.1	2.176	25.7
(2) TYPE	1.025	49.7	1.150	53.5	1.206	55.0	1.269	56.3
(3) INTENSITY	**0.999**	**53.2**	**1.121**	**55.6**	**1.181**	**56.1**	**1.205**	57.7
(4) BAG + TYPE	1.072	49.4	1.194	53.4	1.244	55.3	1.319	55.9
(5) BAG + INTENSITY	1.004	**53.2**	1.138	55.3	1.220	55.9	1.244	**57.8**
SVMlight								
(1) BAG	0.962	40.2	1.432	29.2	1.647	26.2	1.748	24.5
(2) TYPE	0.971	36.5	1.080	27.7	1.117	25.0	1.138	22.4
(3) INTENSITY	1.092	38.1	1.214	29.0	1.264	26.2	1.267	24.7
(4) BAG + TYPE	**0.750**	46.0	**0.926**	34.1	**1.023**	28.9	**1.065**	25.9
(5) BAG + INTENSITY	0.793	**48.3**	0.979	**36.3**	1.071	**32.1**	1.084	**29.4**

*Results are given in both mean-squared error (MSE) and accuracy (Acc). The numbers in bold type are those with the best result for a particular clause level, experiment, and algorithm.

MSE results are significantly better than the MSE results provided by the more challenging medium-class baseline classifier. The same is not true, however, for the accuracy results, which are well below the accuracy results of the neutral-class baseline classifier.

The best experiments for all classifiers use all the subjectivity clues, supporting our hypothesis that using a wide variety of clues is effective. The experiment giving the best results varies somewhat for each classifier, depending on feature organization and whether BAG features are included. For BoosTexter, experiment (5) using BAG and INTENSITY features performs the best. For Ripper, experiment (3) using just the INTENSITY features performs the best, although not significantly better than experiment (5). For SVMlight, which experiment produces the best results depends on whether MSE or accuracy is the metric for evaluation. Experiment (4) using BAG and TYPE features has the better MSE results, experiment (5) using BAG and INTENSITY features has the better accuracies; the differences between the two experiments are significant (except for level-4 MSE).

Figure 6 shows the percent improvements over baseline achieved by each classification algorithm for experiment (5). The medium-class baseline is used for MSE, and the neutral-class baseline is used for accuracy. For BoosTexter, the improvements in MSE range from 36% to 46%, and the improvements in accuracy range from 23% to 96%. The improvements over baseline for Ripper are similar. For SVMlight, the improvements over baseline for MSE are even better, close to 50% for all clause levels.

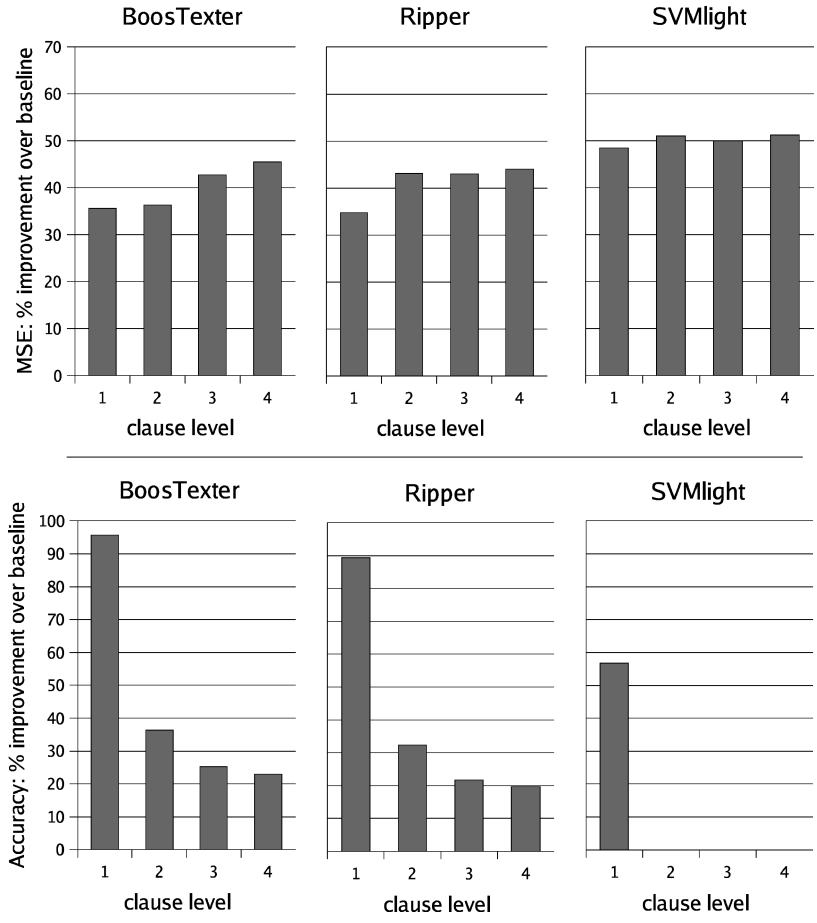


FIGURE 6. Percent improvements over baseline for each algorithm for experiment (5).

Note that BoosTexter and Ripper are non-ordinal classification algorithms, whereas support vector regression takes into account ordinal values. This difference is reflected in the results. The results are comparable for BoosTexter and Ripper (MSE is not significantly different; BoosTexter has slightly better accuracy). Although accuracies are lower, the regression algorithm achieves much better results for MSE. For experiment (5) using the BAG and INTENSITY features, SVMlight improves 10–20% over MSE results for BoosTexter and 49–51% over baseline, coming closer to the true intensity at all clause levels.

8.3.1. Contribution of SYNTAX Clues. In this section, we examine the contribution of the new syntax clues to the classification results. Table 8 shows the increases in MSE and the decreases in accuracy that result when the SYNTAX clues are omitted for experiment (5) (BAG and INTENSITY feature organization).

From Table 8, we see that the new SYNTAX clues do contribute information over and above BAG and the clues from previous work (PREV clues). For all learning algorithms and all-clause levels, omitting the SYNTAX clues results in a significant difference in MSE. The differences in accuracy are also significant, with the exception of BoosTexter levels 1 and 2

TABLE 8. Increases in MSE and Decreases in Accuracy that Result When SYNTAX Clues Are Omitted for Experiment (5)

	Level 1	Level 2	Level 3	Level 4
		Increase in MSE		
BoosTexter	0.090	0.094	0.139	0.152
Ripper	0.226	0.209	0.238	0.215
SVMlight	0.056	0.185	0.229	0.262
		Decrease in accuracy		
BoosTexter	−0.9	−1.0	−2.1	−2.4
Ripper	−2.5	−1.8	−1.6	−1.2
SVMlight	−4.8	−5.1	−4.7	−4.2

and Ripper level 4. The loss in accuracy for SVMlight, which already has lower accuracies, is particularly severe.

8.3.2. TYPE versus INTENSITY Feature Organization. To examine the difference between the TYPE and INTENSITY feature organizations, we again turn to Table 7. For boosting, the experiments using the INTENSITY organization perform better, achieving lower MSEs and higher accuracies. Comparing experiments (2) and (3), the INTENSITY organization performs significantly better at all-clause levels. For experiments (4) and (5), improvements are again significant, with the exception of MSE levels 3 and 4. For Ripper, experiments using the INTENSITY organization also achieve better results, although fewer improvements are significant. For SVMlight, the benefits of the INTENSITY organization are not as clear cut. Experiments using the INTENSITY organization all have higher accuracies, but their MSE is also worse. Furthermore, the differences are all significant, with the exception of the improvement in accuracy for experiment (3) level 3 and the increase in MSE for experiment (5) level 4. This makes it difficult to determine whether the INTENSITY organization is beneficial when performing support vector regression. For Ripper and BoosTexter, however, there is a clear benefit to using the INTENSITY organization for intensity classification.

9. RELATED WORK

To the best of our knowledge, this research is the first to automatically distinguish between not only subjective and objective (*neutral*) language, but among weak, medium, and strong subjectivity as well. The research most closely related is work by Yu and Hatzivassiloglou (2003) and our own earlier work (Wiebe, Bruce, and O'Hara 1999; Riloff et al. 2003; Riloff and Wiebe 2003) on classifying subjective and objective sentences. Yu and Hatzivassilogou use Naïve Bayes classifiers to classify sentences as subjective or objective. Included in the features they use are the words in each sentence, essentially BAG, bigrams, trigrams, and counts of positive and negative words. Their sets of positive and negative words were learned starting with positive and negative adjectives from Hatzivassiloglou and McKeown (1997), which are included in our lexicon of PREV clues. They also use clues that incorporate syntactic information, specifically clues that, for each sentence, encode the polarity of the head verb, main subject, and their modifiers, but these clues do not help with their classifier's performance.

Other researchers have worked to identify opinions below the sentence level (Morinaga et al. 2002; Kim and Hovy 2004; Dave et al. 2003; Nasukawa and Yi 2003; Yi et al. 2003; Hu and Liu 2004; Popescu and Etzioni 2005). Kim and Hovy (2004) identify sentences that mention particular topics, use a named entity tagger to identify the closest entity in the text, and then use the topic and entity phrases to define regions that are used for classifying sentiments. Morinaga et al. (2002), Dave et al. (2003), Nasukawa and Yi (2003), Yi et al. (2003), Hu and Liu (2004), and Popescu and Etzioni (2005) work on mining product reviews. In product review mining, the typical approach is to first identify references to particular products or product features of interest. Once these are identified, positive and negative opinions about the product are extracted. In contrast to the research above, the work in this paper seeks to classify the intensity of nested clauses in all sentences in the corpus.

10. CONCLUSIONS AND FUTURE WORK

This article presents promising results in identifying opinions in deeply nested clauses and classifying their intensities. We use a wide range of features, including new syntactic features. In 10-fold cross-validation experiments using boosting, we achieve improvements over baseline MSE ranging from 36% to 46% and improvements in accuracy ranging from 23% to 96%. Experiments using support vector regression show even stronger MSE results, with improvements ranging from 49% to 51% over baseline.

As tools become available for automatically recognizing different types of subjective expressions, one area of future research will be to investigate whether more complex models can be developed to improve classification. It is possible that different types of subjectivity may be better predicted by individual models, rather than using a single model for all types of subjectivity. These individual models could then be combined into an ensemble classifier with the potential for improved performance overall.[2]

ACKNOWLEDGMENTS

We would like to thank the reviewers for their helpful comments, Paul Hoffmann for programming support, and Adam Lopez for providing his dependency grammatical relationship labeling program. This work was supported in part by the National Science Foundation under grant IIS-0208798.

REFERENCES

ALM, C. O., D. ROTH, and R. SPROAT. 2005. Emotions from text: Machine learning for text-based emotion prediction. *In* Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 347–354, Vancouver, Canada.

BAKER, C., C. FILLMORE, and J. LOWE. 1998. The Berkeley FrameNet Project. *In* Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-98), pp. 86–90, Montreal, Canada.

BALLMER, T., and W. BRENNENSTUHL. 1981. Speech Act Classification: A Study in the Lexical Analysis of English Speech Activity Verbs. Springer-Verlag, Berlin and New York.

[2] We thank one of the anonymous reviewers for this suggestion.

BANFIELD, A. 1982. Unspeakable Sentences. Routledge and Kegan Paul, Boston.

BEINEKE, P., T. HASTIE, and S. VAITHYANATHAN. 2004. The sentimental factor: Improving review classification via human-provided information. *In* Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pp. 263–270, Barcelona, Spain.

BRECK, E., and C. CARDIE. 2004. Playing the telephone game: Determining the hierarchical structure of perspective and speech expressions. *In* Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pp. 120–126, Geneva, Switzerland.

CHOI, Y., C. CARDIE, E. RILOFF, and S. PATWARDHAN. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. *In* Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 355–362, Vancouver, Canada.

COHEN, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, **20**: 37–46.

COHEN, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, **70**: 213–220.

COHEN, W. 1996. Learning trees and rules with set-valued features. In Proceedings of the 13th National Conference on Artificial Intelligence/8th Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI) Volume 1, pp. 709–716, Portland, Oregon.

COLLINS, M. 1997. Three generative, lexicalised models for statistical parsing. *In* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pp. 16–23, Madrid, Spain.

DAS, S. R., and M. Y. CHEN. 2001. Yahoo! for Amazon: Opinion extraction from small talk on the web. *In* Proceedings of the 8th Asia Pacific Finance Association Annual Conference, Bangkok, Thailand.

DAVE, K., S. LAWRENCE, and D. M. PENNOCK. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. *In* Proceedings of the 12th International World Wide Web Conference (WWW2003), Budapest, Hungary. Available at http://www2003.org.

ESULI, A., and F. SEBASTIANI. 2005. Determining the semantic orientation of terms through gloss analysis. *In* Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05), pp. 617–624, Bremen, Germany.

GORDON, A., A. KAZEMZADEH, A. NAIR, and M. PETROVA. 2003. Recognizing expressions of commonsense psychology in English text. *In* Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), pp. 208–215, Sapporo, Japan.

HATZIVASSILOGLOU, V., and K. MCKEOWN. 1997. Predicting the semantic orientation of adjectives. *In* Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL-97), pp. 174–181, Madrid, Spain.

HU, M., and B. LIU. 2004. Mining and summarizing customer reviews. *In* Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2004 (KDD 2004), pp. 168–177, Seattle, Washington.

HWA, R., and A. LOPEZ. 2004. On converting constituent parses to depedency parses. Technical Report TR-04-118, University of Pittsburgh.

JOACHIMS, T. 1999. Making large-scale SVM learning practical. *In* Advances in Kernel Methods–Support Vector Learning. *Edited by* B. Scholkopf, C. Burgess, and A. Smola. MIT-Press, Cambridge, Massachosetts.

KAMPS, J., and M. MARX. 2002. Words with attitude. *In* 1st International WordNet Conference, pp. 332–341, Mysore, India.

KIM, S.-M., and E. HOVY. 2004. Determining the sentiment of opinions. *In* Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pp. 1367–1373, Geneva, Switzerland.

KRIPPENDORFF, K. 1980. Content Analysis: An Introduction to Its Methodology. Sage Publications, Beverly, Hills, California.

KRIPPENDORFF, K. 2004. Content Analysis: An Introduction to Its Methodology, 2nd Edition. Sage Publications, Thousand Oaks, California.

KUDO, T., and Y. MATSUMOTO. 2004. A boosting algorithm for classification of semi-structured text. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp. 301–308, Barcelona, Spain.

LEVIN, B. 1993. English Verb Classes and Alternations: A Preliminary Investigation. University of Chicago Press, Chicago.

LIN, D. 1998. Automatic retrieval and clustering of similar words. *In* Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL-98), pp. 768–773, Montreal, Canada.

LIU, H., H. LIEBERMAN, and T. SELKER. 2003. A model of textual affect sensing using real-world knowledge. *In* Proceedings of the International Conference on Intelligent User Interfaces (IUI-2003), pp. 125–132, Miami, Florida.

MORINAGA, S., K. YAMANISHI, K. TATEISHI, and T. FUKUSHIMA. 2002. Mining product reputations on the web. *In* Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), pp. 341–349, Edmonton, Canada.

MULLEN, T., and N. COLLIER. 2004. Sentiment analysis using support vector machines with diverse information sources. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp. 412–418, Barcelona, Spain.

NASUKAWA, T., and J. YI. 2003. Sentiment analysis: Capturing favorability using natural language processing. *In* Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003), pp. 70–77, Sanibel lsland, Florida.

PANG, B., and L. LEE. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *In* Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05), pp. 115–124, Ann Arbor, Michigan.

PANG, B., L. LEE, and S. VAITHYANATHAN. 2002. Thumbs up? Sentiment classification using machine learning techniques. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), pp. 79–86, Philadelphia, Pennsylvania.

POPESCU, A.-M., and O. ETZIONI. 2005. Extracting product features and opinions from reviews. *In* Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 339–346, Vancouver, Canada.

QUIRK, R., S. GREENBAUM, G. LEECH, and J. SVARTVIK. 1985. A Comprehensive Grammar of the English Language. Longman, New York.

RILOFF, E., and J. ROSIE. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. *In* Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-1999), pp. 474–479, Orlando, Florida.

RILOFF, E., and J. WIEBE. 2003. Learning extraction patterns for subjective expressions. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003), pp. 105–112, Sapporo, Japan.

RILOFF, E., J. WIEBE, and T. WILSON. 2003. Learning subjective nouns using extraction pattern bootstrapping. *In* Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003), pp. 25–32, Edmonton, Canada.

RILOFF, E., J. WIEBE, and W. PHILLIPS. 2005. Exploiting subjectivity classification to improve information extraction. *In* Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-2005), pp. 1106–1111, Pittsburgh, Pennsylvania.

SCHAPIRE, R. E., and Y. SINGER. 2000. BoosTexter: A boosting-based system for text categorization. Machine Learning, **39**(2/3):135–168.

SPERTUS, E. 1997. Smokey: Automatic recognition of hostile messages. *In* Proceedings of the Eighth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-97), pp. 1058–1065, Providence, Rhode Island.

STOYANOV, V., C. CARDIE, and J. WIEBE. 2005. Multi-perspective question answering using the opqa corpus. *In* Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 923–930, Vancouver, Canada.

THELEN, M., and E. RILOFF. 2002. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. *In* Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002), pp. 214–221, Philadelphia, Pennsylvania.

TONG, R. 2001. An operational system for detecting and tracking opinions in on-line discussions. *In* Working Notes of the SIGIR Workshop on Operational Text Classification, pp. 1–6, New Orleans, Louisiana.

TURNEY, P. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *In* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), pp. 417–424, Philadelphia, Pennsylvania.

WHITELAW, C., N. GARG, and S. ARGAMON. 2005. Using appraisal groups for sentiment analysis. *In* Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM-05), pp. 625–631, Bremen, Germany.

WIEBE, J. 1990. Recognizing subjective sentences: A computational investigation of narrative text. Ph.D. Thesis, State University of New York at Buffalo.

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

WIEBE, J. 2000. Learning subjective adjectives from corpora. *In* Proceedings of the17th National Conference on Artificial Intelligence (AAAI-2000), pp. 735–740, Austin, Texas.

WIEBE, J., K. MCKEEVER, and R. BRUCE. 1998. Mapping collocational properties into machine learning features. *In* Proceedings of the 6th Workshop on Very Large Corpora (WVLC-6), pp. 225–233, Montreal, Canada.

WIEBE, J., R. BRUCE, and T. O'HARA. 1999. Development and use of a gold standard data set for subjectivity classifications. *In* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99), pp. 246–253, College Park, Maryland.

WIEBE, J., T. WILSON, and M. BELL. 2001. Identifying collocations for recognizing opinions. *In* Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, pp. 24–31, Toulouse, France.

WIEBE, J., T. WILSON, and C. CARDIE. 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation (formerly Computers and the Humanities), **39**(2/3):165–210.

WIEBE, J., T. WILSON, R. BRUCE, M. BELL, and M. MARTIN. 2004. Learning subjective language. Computational Linguistics, **30**(3):277–308.

WIEBE, J., E. BRECK, C. BUCKLEY, C. CARDIE, P. DAVIS, B. FRASER, D. LITMAN, D. PIERCE, E. RILOFF, T. WILSON, D. DAY, and M. 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, pp. 12–19, Palo Alto, California.

WILSON, T., J. WIEBE, and P. HOFFMAN. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *In* Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 347–354, Vancouver, Canada.

YI, J., T. NASUKAWA, R. BUNESCU, and W. NIBLACK. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. *In* Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003), pp. 427–434, Melbourne, Florida.

YU, H., and V. HATZIVASSILOGLOU. 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 (EMNLP-2003), pp. 129–136, Sapporo, Japan.