Lexicon-Based Methods for Sentiment Analysis

Maite Taboada*
Simon Fraser University

Julian Brooke**
University of Toronto

Milan Tofiloski†
Simon Fraser University

Kimberly Voll‡
University of British Columbia

Manfred Stede§
University of Potsdam

We present a lexicon-based approach to extracting sentiment from text. The Semantic Orientation CALculator (SO-CAL) uses dictionaries of words annotated with their semantic orientation (polarity and strength), and incorporates intensification and negation. SO-CAL is applied to the polarity classification task, the process of assigning a positive or negative label to a text that captures the text's opinion towards its main subject matter. We show that SO-CAL's performance is consistent across domains and on completely unseen data. Additionally, we describe the process of dictionary creation, and our use of Mechanical Turk to check dictionaries for consistency and reliability.

1. Introduction

Semantic orientation (SO) is a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency or strength (degree to which the word, phrase, sentence, or document in question is positive or negative)

* Corresponding author. Department of Linguistics, Simon Fraser University, 8888 University Dr., Burnaby, B.C. V5A 1S6 Canada. E-mail: mtaboada@sfu.ca.
** Department of Computer Science, University of Toronto, 10 King's College Road, Room 3302, Toronto, Ontario M5S 3G4 Canada. E-mail: jbrooke@cs.toronto.edu.
† School of Computing Science, Simon Fraser University, 8888 University Dr., Burnaby, B.C. V5A 1S6 Canada. E-mail: mta45@sfu.ca.
‡ Department of Computer Science, University of British Columbia, 201-2366 Main Mall, Vancouver, B.C. V6T 1Z4 Canada. E-mail: kvoll@cs.ubc.ca.
§ Department of Linguistics, University of Potsdam, Karl-Liebknecht-Str. 24-25. D-14476 Golm, Germany. E-mail: stede@uni-potsdam.de.

Submission received: 14 December 2009; revised submission received: 22 August 2010; accepted for publication: 28 September 2010.

towards a subject topic, person, or idea (Osgood, Suci, and Tannenbaum 1957). When used in the analysis of public opinion, such as the automated interpretation of on-line product reviews, semantic orientation can be extremely helpful in marketing, measures of popularity and success, and compiling reviews.

The analysis and automatic extraction of semantic orientation can be found under different umbrella terms: sentiment analysis (Pang and Lee 2008), subjectivity (Lyons 1981; Langacker 1985), opinion mining (Pang and Lee 2008), analysis of stance (Biber and Finegan 1988; Conrad and Biber 2000), appraisal (Martin and White 2005), point of view (Wiebe 1994; Scheibman 2002), evidentiality (Chafe and Nichols 1986), and a few others, without expanding into neighboring disciplines and the study of emotion (Ketal 1975; Ortony, Clore, and Collins 1988) and affect (Batson, Shaw, and Oleson 1992). In this article, **sentiment analysis** refers to the general method to extract subjectivity and polarity from text (potentially also speech), and **semantic orientation** refers to the polarity and strength of words, phrases, or texts. Our concern is primarily with the semantic orientation of texts, but we extract the sentiment of words and phrases towards that goal.

There exist two main approaches to the problem of extracting sentiment automatically.[1] The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document (Turney 2002). The text classification approach involves building classifiers from labeled instances of texts or sentences (Pang, Lee, and Vaithyanathan 2002), essentially a supervised classification task. The latter approach could also be described as a statistical or machine-learning approach. We follow the first method, in which we use dictionaries of words annotated with the word's semantic orientation, or polarity.

Dictionaries for lexicon-based approaches can be created manually, as we describe in this article (see also Stone et al. 1966; Tong 2001), or automatically, using seed words to expand the list of words (Hatzivassiloglou and McKeown 1997; Turney 2002; Turney and Littman 2003). Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of text (Hatzivassiloglou and McKeown 1997; Wiebe 2000; Hu and Liu 2004; Taboada, Anthony, and Voll 2006).[2] First, a list of adjectives and corresponding SO values is compiled into a dictionary. Then, for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text.

The majority of the statistical text classification research builds Support Vector Machine classifiers, trained on a particular data set using features such as unigrams or bigrams, and with or without part-of-speech labels, although the most successful features seem to be basic unigrams (Pang, Lee, and Vaithyanathan 2002; Salvetti, Reichenbach, and Lewis 2006). Classifiers built using supervised methods reach quite a high accuracy in detecting the polarity of a text (Chaovalit and Zhou 2005; Kennedy and Inkpen 2006; Boiy et al. 2007; Bartlett and Albright 2008). However, although such classifiers perform very well in the domain that they are trained on, their performance drops precipitously (almost to chance) when the same classifier is used in

1 Pang and Lee (2008) provide an excellent recent survey of the opinion mining or sentiment analysis problem and the approaches used to tackle it.

2 With some exceptions: Turney (2002) uses two-word phrases; Whitelaw, Garg, and Argamon (2005) adjective phrases; and Benamara et al. (2007) adjectives with adverbial modifiers. See also Section 2.1. We should also point out that Turney does not create a static dictionary, but rather scores two-word phrases on the fly.

a different domain (Aue and Gamon [2005]; see also the discussion about domain specificity in Pang and Lee [2008, section 4.4]).[3] Consider, for example, an experiment using the Polarity Dataset, a corpus containing 2,000 movie reviews, in which Brooke (2009) extracted the 100 most positive and negative unigram features from an SVM classifier that reached 85.1% accuracy. Many of these features were quite predictable: *worst*, *waste*, *unfortunately*, and *mess* are among the most negative, whereas *memorable*, *wonderful*, *laughs*, and *enjoyed* are all highly positive. Other features are domain-specific and somewhat inexplicable: If the writer, director, plot, or script are mentioned, the review is likely to be disfavorable towards the movie, whereas the mention of performances, the ending, or even flaws, indicates a good movie. Closed-class function words appear frequently; for instance, *as*, *yet*, *with*, and *both* are all extremely positive, whereas *since*, *have*, *though*, and *those* have negative weight. Names also figure prominently, a problem noted by other researchers (Finn and Kushmerick 2003; Kennedy and Inkpen 2006). Perhaps most telling is the inclusion of unigrams like *2*, *video*, *tv*, and *series* in the list of negative words. The polarity of these words actually makes some sense in context: Sequels and movies adapted from video games or TV series do tend to be less well-received than the average movie. However, these real-world facts are not the sort of knowledge a sentiment classifier ought to be learning; within the domain of movie reviews such facts are prejudicial, and in other domains (e.g., video games or TV shows) they are either irrelevant or a source of noise.

Another area where the lexicon-based model might be preferable to a classifier model is in simulating the effect of linguistic context. On reading any document, it becomes apparent that aspects of the local context of a word need to be taken into account in SO assessment, such as negation (e.g., *not good*) and intensification (e.g., *very good*), aspects that Polanyi and Zaenen (2006) named contextual valence shifters. Research by Kennedy and Inkpen (2006) concentrated on implementing those insights. They dealt with negation and intensification by creating separate features, namely, the appearance of *good* might be either *good* (no modification) *not_good* (negated *good*), *int_good* (intensified *good*), or *dim_good* (diminished *good*). The classifier, however, cannot determine that these four types of *good* are in any way related, and so in order to train accurately there must be enough examples of all four in the training corpus. Moreover, we show in Section 2.4 that expanding the scope to two-word phrases does not deal with negation adequately, as it is often a long-distance phenomenon. Recent work has begun to address this issue. For instance, Choi and Cardie (2008) present a classifier that treats negation from a compositional point of view by first calculating polarity of terms independently, and then applying inference rules to arrive at a combined polarity score. As we shall see in Section 2, our lexicon-based model handles negation and intensification in a way that generalizes to all words that have a semantic orientation value.

A middle ground exists, however, with semi-supervised approaches to the problem. Read and Carroll (2009), for instance, use semi-supervised methods to build domain-independent polarity classifiers. Read and Carroll built different classifiers and show that they are more robust across domains. Their classifiers are, in effect, dictionary-based, differing only in the methodology used to build the dictionary. Li et al. (2010) use co-training to incorporate labeled and unlabeled examples, also making use of

3 Blitzer, Dredze, and Pereira (2007) do show some success in transferring knowledge across domains, so that the classifier does not have to be re-built entirely from scratch.

a distinction between sentences with a first person subject and with other subjects. Other hybrid methods include those of Andreevskaia and Bergler (2008), Dang, Zhang, and Chen (2010), Dasgupta and Ng (2009), Goldberg and Zhu (2006), or Prabowo and Thelwall (2009). Wan (2009) uses co-training in a method that uses English labeled data and an English classifier to learn a classifier for Chinese.

In our approach, we seek methods that operate at a deep level of analysis, incorporating semantic orientation of individual words and contextual valence shifters, yet do not aim at a full linguistic analysis (one that involves analysis of word senses or argument structure), although further work in that direction is possible.

In this article, starting in Section 2, we describe the Semantic Orientation CALculator (SO-CAL) that we have developed over the last few years. We first extract sentiment-bearing words (including adjectives, verbs, nouns, and adverbs), and use them to calculate semantic orientation, taking into account valence shifters (intensifiers, downtoners, negation, and irrealis markers). We show that this lexicon-based method performs well, and that it is robust across domains and texts. One of the criticisms raised against lexicon-based methods is that the dictionaries are unreliable, as they are either built automatically or hand-ranked by humans (Andreevskaia and Bergler 2008). In Section 3, we present the results of several experiments that show that our dictionaries are robust and reliable, both against other existing dictionaries, and as compared to values assigned by humans (through the use of the Mechanical Turk interface). Section 4 provides comparisons to other work, and Section 5 conclusions.

2. SO-CAL, the Semantic Orientation CALculator

Following Osgood, Suci, and Tannenbaum (1957), the calculation of sentiment in SO-CAL begins with two assumptions: that individual words have what is referred to as **prior polarity**, that is, a semantic orientation that is independent of context; and that said semantic orientation can be expressed as a numerical value. Several lexicon-based approaches have adopted these assumptions (Bruce and Wiebe 2000; Hu and Liu 2004; Kim and Hovy 2004). In this section, we describe the different dictionaries used in SO-CAL, and the incorporation of valence shifters. We conclude the section with tests that show SO-CAL's performance on different data sets.

2.1 Adjectives

Much of the early research in sentiment focused on adjectives or adjective phrases as the primary source of subjective content in a document (Hatzivassiloglou and McKeown 1997; Hu and Liu 2004; Taboada, Anthony, and Voll 2006), albeit with some exceptions, especially more recently, which have also included the use of adverbs (Benamara et al. 2007); adjectives and verbs (Kim and Hovy 2004); adjective phrases (Whitelaw, Garg, and Argamon 2005); two-word phrases (Turney 2002; Turney and Littman 2003); adjectives, verbs, and adverbs (Subrahmanian and Reforgiato 2008); the exclusive use of verbs (Sokolova and Lapalme 2008); the use of *non-affective* adjectives and adverbs (Sokolova and Lapalme 2009a, 2009b); or rationales, words and phrases selected by human annotators (Zaidan and Eisner 2008). In general, the SO of an entire document is the combined effect of the adjectives or relevant words found within, based upon a dictionary of word rankings (scores). The dictionary can be created in different ways: manually, using existing dictionaries such as the General Inquirer (Stone et al. 1966), or semi-automatically, making use of resources like WordNet (Hu and Liu 2004; Kim

and Hovy 2004; Esuli and Sebastiani 2006). The dictionary may also be produced automatically via association, where the score for each new adjective is calculated using the frequency of the proximity of that adjective with respect to one or more seed words. **Seed words** are a small set of words with strong negative or positive associations, such as *excellent* or *abysmal*. In principle, a positive adjective should occur more frequently alongside the positive seed words, and thus will obtain a positive score, whereas negative adjectives will occur most often in the vicinity of negative seed words, thus obtaining a negative score. The association is usually calculated following Turney's method for computing mutual information (Turney 2002; Turney and Littman 2003), but see also Rao and Ravichandran (2009) and Velikovich et al. (2010) for other methods using seed words.

Previous versions of SO-CAL (Taboada and Grieve 2004; Taboada, Anthony, and Voll 2006) relied on an adjective dictionary to predict the overall SO of a document, using a simple aggregate-and-average method: The individual scores for each adjective in a document are added together and then divided by the total number of adjectives in that document.[4] As we describe subsequently, the current version of SO-CAL takes other parts of speech into account, and makes use of more sophisticated methods to determine the true contribution of each word.

It is important to note that how a dictionary is created affects the overall accuracy of subsequent results. In Taboada, Anthony, and Voll (2006) we report on experiments using different search engines and operators in trying to create dictionaries semi-automatically. We found that, although usable, dictionaries created using the Google search engine were unstable. When rerun, the results for each word were subject to change, sometimes by extreme amounts, something that Kilgarriff (2007) also notes, arguing against the use of Google for linguistic research of this type. An alternative would be to use a sufficiently large static corpus, as Turney (2006) does to measure relational similarity across word pairs.

Automatically or semi-automatically created dictionaries have some advantages. We found many novel words in our initial Google-generated dictionary. For instance, *unlistenable* was tagged accurately as highly negative, an advantage that Velikovich et al. (2010) point out. However, in light of the lack of stability for automatically generated dictionaries, we decided to create manual ones. These were produced by hand-tagging all adjectives found in our development corpus, a 400-text corpus of reviews (see the following) on a scale ranging from -5 for extremely negative to $+5$ for extremely positive, where 0 indicates a neutral word (excluded from our dictionaries). "Positive" and "negative" were decided on the basis of the word's prior polarity, that is, its meaning in most contexts. We do not deal with word sense disambiguation but suspect that using even a simple method to disambiguate would be beneficial. Some word sense ambiguities are addressed by taking part of speech into account. For instance, as we mention in Section 3.4, *plot* is only negative when it is a verb, but should not be so in a noun dictionary; *novel* is a positive adjective, but a neutral noun.

To build the system and run our experiments, we use the corpus described in Taboada and Grieve (2004) and Taboada, Anthony, and Voll (2006), which consists of a 400-text collection of Epinions reviews extracted from eight different categories: books, cars, computers, cookware, hotels, movies, music, and phones, a corpus we named "Epinions 1." Within each collection, the reviews were split into 25 positive and 25

4 To determine part of speech, we use the Brill tagger (Brill 1992).

negative reviews, for a total of 50 in each category, and a grand total of 400 reviews in the corpus (279,761 words). We determined whether a review was positive or negative through the "recommended" or "not recommended" feature provided by the review's author.

2.2 Nouns, Verbs, and Adverbs

In the following example, adapted from Polanyi and Zaenen (2006), we see that lexical items other than adjectives can carry important semantic polarity information.

(1) a. The young man strolled+ purposefully+ through his neighborhood+.
b. The teenaged male strutted− cockily− through his turf−.

Although the sentences have comparable literal meanings, the plus-marked nouns, verbs, and adverbs in Example (1a) indicate the positive orientation of the speaker towards the situation, whereas the minus-marked words in Example (1b) have the opposite effect. It is the combination of these words in each of the sentences that conveys the semantic orientation for the entire sentence.[5]

In order to make use of this additional information, we created separate noun, verb, and adverb dictionaries, hand-ranked using the same +5 to −5 scale as our adjective dictionary. The enhanced dictionaries contain 2,252 adjective entries, 1,142 nouns, 903 verbs, and 745 adverbs.[6] The SO-carrying words in these dictionaries were taken from a variety of sources, the three largest being Epinions 1, the 400-text corpus described in the previous section; a 100-text subset of the 2,000 movie reviews in the Polarity Dataset (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2004, 2005);[7] and positive and negative words from the General Inquirer dictionary (Stone et al. 1966; Stone 1997).[8] The sources provide a fairly good range in terms of register: The Epinions and movie reviews represent informal language, with words such as *ass-kicking* and *nifty*; at the other end of the spectrum, the General Inquirer was clearly built from much more formal texts, and contributed words such as *adroit* and *jubilant*, which may be more useful in the processing of literary reviews (Taboada, Gillies, and McFetridge 2006; Taboada et al. 2008) or other more formal texts.

Each of the open-class words was assigned a hand-ranked SO value between 5 and −5 (neutral or zero-value words were excluded) by a native English speaker. The numerical values were chosen to reflect both the prior polarity and the strength of the word, averaged across likely interpretations. The dictionaries were later reviewed by a committee of three other researchers in order to minimize the subjectivity of ranking SO by hand. Examples are shown in Table 1.

One difficulty with nouns and verbs is that they often have both neutral and non-neutral connotations. In the case of *inspire* (or *determination*), there is a very positive meaning (Example (2)) as well as a rather neutral meaning (Example (3)).

(2) The teacher inspired her students to pursue their dreams.

(3) This movie was inspired by true events.

5 Something that Turney (2002) already partially addressed, by extracting two-word phrases.
6 Each dictionary also has associated with it a stop-word list. For instance, the adjective dictionary has a stop-word list that includes *more, much*, and *many*, which are tagged as adjectives by the Brill tagger.
7 Available from `www.cs.cornell.edu/People/pabo/movie-review-data/`.
8 Available from `www.wjh.harvard.edu/~inquirer/`.

Table 1
Examples of words in the noun and verb dictionaries.

Word	SO Value
monstrosity	−5
hate (noun and verb)	−4
disgust	−3
sham	−3
fabricate	−2
delay (noun and verb)	−1
determination	1
inspire	2
inspiration	2
endear	3
relish (verb)	4
masterpiece	5

Except when one sense was very uncommon, the value chosen reflected an averaging across possible interpretations. In some cases, the verb and related noun have a different SO value. For instance, *exaggerate* is −1, whereas *exaggeration* is −2, and the same values are applied to *complicate* and *complication*, respectively. We find that grammatical metaphor (Halliday 1985), that is, the use of a noun to refer to an action, adds a more negative connotation to negative words.

All nouns and verbs encountered in the text are lemmatized,[9] and the form (singular or plural, past tense or present tense) is not taken into account in the calculation of SO value. As with the adjectives, there are more negative nouns and verbs than positive ones.[10]

The adverb dictionary was built automatically using our adjective dictionary, by matching adverbs ending in *-ly* to their potentially corresponding adjective, except for a small selection of words that were added or modified by hand. When SO-CAL encountered a word tagged as an adverb that was not already in its dictionary, it would stem the word and try to match it to an adjective in the main dictionary. This worked quite well for most adverbs, resulting in semantic orientation values that seem appropriate (see examples in Table 2).

In other cases—for example, *essentially*—there is a mismatch between the meaning (or usage pattern) of the adverb when compared to the adjective it is based on, and the value was manually corrected.

Although the vast majority of the entries are single words, SO-CAL allows for multi-word entries written in a regular expression–like language; in particular, the verb dictionary contains 152 multi-word expressions (mostly phrasal verbs, e.g., *fall apart*), and the intensifier dictionary, described subsequently, contains 35 multi-word entries (e.g., *a little bit*). Multi-word expressions take precedence over single-word expressions; for instance, *funny* by itself is positive (+2), but if the phrase *act funny* appears, it is given a negative value (−1).

9 Lemmatization is a simple process of stripping any endings from words not in the dictionary, according to their part of speech. After stripping, we perform a new dictionary look-up.
10 The ratio for adjectives is 47:53 positive to negative, and for nouns it is 41:59.

Table 2
Examples from the adverb dictionary.

Word	SO Value
excruciatingly	−5
inexcusably	−3
foolishly	−2
satisfactorily	1
purposefully	2
hilariously	4

It is difficult to measure the coverage of our dictionaries, because there is no direct way to estimate the number of SO-carrying words and expressions in English (although it should probably be larger than 5,000, the rough total of our current dictionaries). Wilson, Wiebe, and Hoffmann (2005) provide a list of subjectivity cues with over 8,000 entries; there are many more neutral, repeated, and inflectionally related entries than in our dictionaries, however, as well as many more nouns, and far fewer adjectives. Automatically generated dictionaries are generally much larger: SentiWordNet (Baccianella, Esuli, and Sebastiani 2010) includes 38,182 non-neutral words (when the polarity of senses is averaged—see discussion in Section 3.4), and the Maryland dictionary (Mohammad, Dorr, and Dunne 2009) has 76,775 words and phrases tagged for polarity. We will see, in Section 3.4, that larger dictionaries are not necessarily better, in some cases because the information contained is not as detailed (the Maryland dictionary is not classified by part of speech), or because, in general, including more words may also lead to including more noise.

Independently of the difference between manually and automatically generated dictionaries, we have evidence that coverage is a double-edged sword: With an earlier version of SO-CAL we extracted 50 texts from the Polarity Dataset (texts from which we had not previously drawn words) and extracted all words judged to have sentiment that were not already in our dictionaries. We found 116 adjectives, 62 nouns, 43 verbs, and 7 adverbs, a small fraction (less than 3%) of the words in our present lexicon. When these words were added, we found that performance in that data set actually dropped (by 4%). We believe this effect is related to the large amounts of description in genres such as movie reviews (see Taboada, Brooke, and Stede [2009] for a way to address the problem of descriptive noise); basically, rarer vocabulary is likely to contain a strong descriptive (semantic) component, and thus be used in a way that is tangential to overall text sentiment. In any case, the best argument for good (and appropriate) coverage is acceptable performance for new texts in new domains, and indeed we will see in Sections 2.8 and 2.9 that there is little difference in performance between texts and domains which were used to build our dictionaries, and others which were not.

2.3 Intensification

Quirk et al. (1985) classify intensifiers into two major categories, depending on their polarity: Amplifiers (e.g., *very*) increase the semantic intensity of a neighboring lexical item, whereas downtoners (e.g., *slightly*) decrease it. Some researchers in sentiment analysis (Kennedy and Inkpen 2006; Polanyi and Zaenen 2006) have implemented

Table 3
Percentages for some intensifiers.

Intensifier	Modifier (%)
slightly	−50
somewhat	−30
pretty	−10
really	+15
very	+25
extraordinarily	+50
(the) most	+100

intensifiers using simple addition and subtraction—that is, if a positive adjective has an SO value of 2, an amplified adjective would have an SO value of 3, and a downtoned adjective an SO value of 1. One problem with this kind of approach is that it does not account for the wide range of intensifiers within the same subcategory. *Extraordinarily*, for instance, is a much stronger amplifier than *rather*. Another concern is that the amplification of already "loud" items should involve a greater overall increase in intensity when compared to more subdued counterparts (compare *truly fantastic* with *truly okay*); in short, intensification should also depend on the item being intensified.[11] In SO-CAL, intensification is modeled using modifiers, with each intensifying word having a percentage associated with it; amplifiers are positive, whereas downtoners are negative, as shown in Table 3.

For example, if *sleazy* has an SO value of −3, *somewhat sleazy* would have an SO value of: $-3 \times (100\% - 30\%) = -2.1$. If *excellent* has a SO value of 5, *most excellent* would have an SO value of: $5 \times (100\% + 100\%) = 10$. Intensifiers are applied recursively starting from the closest to the SO-valued word: If *good* has an SO value of 3, then *really very good* has an SO value of $(3 \times [100\% + 25\%]) \times (100\% + 15\%) = 4.3$.

Because our intensifiers are implemented using a percentage scale, they are able to fully capture the variety of intensifying words as well as the SO value of the item being modified. This scale can be applied to other parts of speech, given that adjectives, adverbs, and verbs use the same set of intensifiers, as seen in Example (4), where *really* modifies an adjective (*fantastic*), an adverb (*well*), and a verb (*enjoyed*).

(4) a. The performances were all really fantastic.
b. Zion and Planet Asai from the Cali Agents flow really well over this.
c. I really enjoyed most of this film.

Nouns, however, are modified exclusively by adjectives. We are able to take into account some kinds of modification using our main adjective dictionary; there is a small class of adjectives which would not necessarily amplify or downtone correctly if considered in isolation, however, as seen in the following (invented) examples. Here, adjectives such as *total* do not have a semantic orientation of their own, but, just like adverbial intensifiers, contribute to the interpretation of the word that follows them; *total failure* is presumably worse than just *failure*. Thus, we have a separate dictionary for adjectival intensifiers. When an intensifying adjective appears next to

11 Martin and White (2005, page 139) also suggest that the effect is different according to the polarity of the item being intensified. We have not explored that possibility.

an SO-valued noun, it is treated as an intensifier rather than as a separate SO-bearing unit.

(5) a. The plot had huge problems.
b. They have made clear progress.
c. This is a total failure.
d. It's an insignificant criticism.

Besides adverbs and adjectives, other intensifiers are quantifiers (*a great deal of*). We also included three other kinds of intensification that are common within our genre: the use of all capital letters, the use of exclamation marks, and the use of discourse connective *but* to indicate more salient information (e.g., *...but the movie was GREAT!*).[12] In all, our intensifier dictionary contains 177 entries, some of them multi-word expressions.

2.4 Negation

The obvious approach to negation is simply to reverse the polarity of the lexical item next to a negator, changing *good* (+3) into *not good* (−3). This we may refer to as **switch negation** (Saurí 2008). There are a number of subtleties related to negation that need to be taken into account, however. One is the fact that there are negators, including *not*, *none*, *nobody*, *never*, and *nothing*, and other words, such as *without* or *lack* (verb and noun), which have an equivalent effect, some of which might occur at a significant distance from the lexical item which they affect; a backwards search is required to find these negators, one that is tailored to the particular part of speech involved. We assume that for adjectives and adverbs the negation is fairly local, though it is sometimes necessary to look past determiners, copulas, and certain verbs, as we see in Example (6).

(6) a. Nobody gives a good performance in this movie. (*nobody* negates *good*)
b. Out of every one of the fourteen tracks, none of them approach being weak and are all stellar. (*none* negates *weak*)
c. Just a V-5 engine, nothing spectacular. (*nothing* negates *spectacular*)

Negation search in SO-CAL includes two options: Look backwards until a clause boundary marker is reached;[13] or look backwards as long as the words/tags found are in a backward search skip list, with a different list for each part of speech. The first approach is fairly liberal, and allows us to capture the true effects of negation raising (Horn 1989), where the negator for a verb moves up and attaches to the verb in the matrix clause. In the following examples the *don't* that negates the verb *think* is actually negating the embedded clause.

(7) I don't wish to reveal much else about the plot because I don't think it is worth mentioning.

12 The discourse connective *but* plays a role in linking clauses and sentences in a rhetorical relation (Mann and Thompson 1988). There are more sophisticated ways of making use of those relations, but we have not implemented them yet.

13 Clause boundary markers include punctuation and sentential connectives, including some ambiguous ones such as *and* and *but*.

(8) Based on other reviews, I don't think this will be a problem for a typical household environment.

The second approach is more conservative. In Example (7), the search would only go as far as *it*, because adjectives (*worth*), copulas, determiners, and certain basic verbs are on the list of words to be skipped (allowing negation of adjectives within VPs and NPs), but pronouns are not. Similarly, verbs allow negation on the other side of *to*, and nouns look past adjectives as well as determiners and copulas. This conservative approach seems to work better, and is what we use in all the experiments in this article.[14]

Another issue is whether a polarity flip (switch negation) is the best way to quantify negation. Though it seems to work well in certain cases (Choi and Cardie 2008), it fails miserably in others (Liu and Seneff 2009). Consider *excellent*, a +5 adjective: If we negate it, we get *not excellent*, which intuitively is a far cry from *atrocious*, a −5 adjective. In fact, *not excellent* seems more positive than *not good*, which would negate to a −3. In order to capture these pragmatic intuitions, we implemented another method of negation, a polarity shift (**shift negation**). Instead of changing the sign, the SO value is shifted toward the opposite polarity by a fixed amount (in our current implementation, 4). Thus a +2 adjective is negated to a −2, but the negation of a −3 adjective (for instance, *sleazy*) is only slightly positive, an effect we could call "damning with faint praise." Here are a few examples from our corpus.

(9) a. She's not terrific ($5 - 4 = 1$) but not terrible ($-5 + 4 = -1$) either.
b. Cruise is not great ($4 - 4 = 0$), but I have to admit he's not bad ($-3 + 4 = 1$) either.
c. This CD is not horrid ($-5 + 4 = -1$).

In each case, the negation of a strongly positive or negative value reflects a mixed opinion which is correctly captured in the shifted value. Further (invented) examples are presented in Example (10).

(10) a. Well, at least he's not sleazy. ($-3 \rightarrow 1$)
b. Well, it's not dreadful. ($-4 \rightarrow 0$)
c. It's just not acceptable. ($1 \rightarrow -3$)
d. It's not a spectacular film, but ... ($5 \rightarrow 1$)

As shown in the last example, it is very difficult to negate a strongly positive word without implying that a less positive one is to some extent true, and thus our negator becomes a downtoner.

A related problem for the polarity flip model, as noted by Kennedy and Inkpen (2006), is that negative polarity items interact with intensifiers in undesirable ways. *Not very good*, for instance, comes out more negative than *not good*. Another way to handle this problem while preserving the notion of a polarity flip is to allow the negative item to flip the polarity of both the adjective and the intensifier; in this way, an amplifier becomes a downtoner:

$$\text{Not good} = 3 \times -1 = -3$$
$$\text{Not very good} = 3 \times (100\% + 25\% \times -1) \times -1 = -2.25$$

14 Full parsing is also an option, but the speed of the parser could pose problems if the goal is to process text on-line. Parsing would still produce ambiguities, and may not be able to correctly interpret scope. Another option is to use parser results to learn the scope (Councill, McDonald, and Velikovich 2010).

Compare with the polarity shift version, which is only marginally negative:

$$\text{Not good} = 3 - 4 = -1$$
$$\text{Not very good} = 3 \times (100\% + 25\%) - 4 = -0.25$$

The problems with polarity shift could probably be resolved by fine-tuning SO values and modifiers; the polarity flip model seems fundamentally flawed, however. Polarity shifts seem to better reflect the pragmatic reality of negation, and is supported by Horn (1989), who suggests that affirmative and negative sentences are not symmetrical.

One other interesting aspect of the pragmatics of negation is that negative statements tend to be perceived as more marked than their affirmative counterparts, both pragmatically and psychologically (Osgood and Richards 1973; Horn 1989, chapter 3). This markedness is true in terms of linguistic form, with negative forms being marked across languages (Greenberg 1966), and it is also manifested as (token) frequency distribution, with negatives being less frequent (Boucher and Osgood 1969).[15] Negation tends to be expressed in euphemistic ways, which makes negative sentiment more difficult to identify in general.

In our treatment of negation, we consider mostly negators, but not negative polarity items (NPIs), such as *any*, *anything*, *ever*, or *at all*. In some cases, searching for an NPI would be more effective than searching for a negator. NPIs occur in negative sentences, but also in nonveridical contexts (Zwarts 1995; Giannakidou 1998), which also affect semantic orientation. For instance, *any* occurs in contexts other than negative sentences, as shown in Example (11), from Giannakidou (2001, page 99), where in all cases the presence of *any* is due to a nonveridical situation. Using NPIs would allow us to reduce semantic orientation values in such contexts. We address some of these issues through irrealis blocking, as we explain in the next section.

(11) a. Did you find any interesting books?
b. Pick any apple!
c. He might come any moment now.
d. I insist you allow anyone in.

Similarly, negation calculation does not include what Choi and Cardie (2008) term "content word negators," words such as *eliminate*. Most of those are included in the respective dictionaries (i.e., the verb dictionary for *eliminate*) with negative polarity. When they occur in a sentence, aggregation with other sentiment words in the sentence would probably yield a result similar to the compositional approach of Choi and Cardie or Moilanen and Pulman (2007).

2.5 Irrealis Blocking

There are a number of markers that indicate that the words appearing in a sentence might not be reliable for the purposes of sentiment analysis. We refer to these using the linguistic term *irrealis*, usually applied in non-factual contexts. English does not make extensive use of the subjunctive for this purpose, as opposed to other languages, such as Spanish, which tend to use the subjunctive mood to indicate that what is being

15 Some researchers argue that there is a negative bias in the human representation of experience (negative events are more salient), and the positive bias found by Boucher and Osgood is the result of euphemisms and political correctness in language (Jing-Schmidt 2007).

expressed is not factual. However, English has a few other mechanisms to convey irrealis. Word order, modal verbs, and private-state verbs related to expectation fulfill that function. The imperative mood also conveys that the action expressed has not occurred.

Irrealis markers can change the meaning of sentiment-bearing words in very subtle ways. In some cases, such as Example (12), the right approach is to reverse the SO of *good*, which is in the scope of the modal verb *would*. This interpretation is supported by the contrast in the *but* clause. In Example (13), on the other hand, the modal should not reverse the positive evaluation conveyed by *best*.

(12) I thought this movie would be as good as the Grinch, but unfortunately, it wasn't.

(13) But for adults, this movie could be one of the best of the holiday season.

The approach currently implemented consists of ignoring the semantic orientation of any word in the scope of an irrealis marker (i.e., within the same clause). In Example (14), the positive value of *great* is simply ignored.

(14) This should have been a great movie. (3 → 0)

Our list of irrealis markers includes modals, conditional markers (*if*), negative polarity items like *any* and *anything*, certain (mostly intensional) verbs (*expect*, *doubt*), questions, and words enclosed in quotes (which may be factual, but not necessarily reflective of the author's opinion).

There is good reason to include NPIs as irrealis blockers rather than as full-fledged negators: NPIs often appear in embedded alternatives which are not generally marked with question marks and where negation would not be appropriate. In the following example, *any* is part of the complement of *wonder*, which has an implicit alternative (*whether there are going to be any problems with that ... or not*).

(15) I wonder whether there are going to be any problems with that.

There is one case, at least, where it is clear that the SO value of a term should not be nullified by an irrealis blocker, as in Example (16), where the question mark currently blocks the negative orientation of *amateurish crap*. The question is rhetorical in this case, but we have no way of distinguishing it from a real question. Although not very common, this kind of off-hand opinion, buried in a question, imperative, or modal clause, is often quite strong and very reliable. SO-CAL looks for markers of definiteness within close proximity of SO-carrying words (within the NP, such as the determiner *this*), and ignores irrealis blocking if an irrealis marker is found.

(16) ... he can get away with marketing this amateurish crap and still stay on the bestseller list?

2.6 Text-Level Features

Lexicon-based sentiment classifiers generally show a positive bias (Kennedy and Inkpen 2006), likely the result of a universal human tendency to favor positive language (Boucher and Osgood 1969).[16] In order to overcome this problem, Voll and Taboada (2007) implemented normalization, shifting the numerical cut-off point between positive and negative reviews. In the current version of SO-CAL, we have used a somewhat

16 On average, there are almost twice as many positive as negative words in our texts.

different approach, instead supposing that negative expressions, being relatively rare, are given more cognitive weight when they do appear. Thus we increase the final SO of any negative expression (after other modifiers have applied) by a fixed amount (currently 50%). This seems to have essentially the same effect in our experiments, and is more theoretically satisfying.

Pang, Lee, and Vaithyanathan (2002) found that their machine-learning classifier performed better when a binary feature was used indicating the presence of a unigram in the text, instead of a numerical feature indicating the number of appearances. Counting each word only once does not seem to work equally well for word-counting models. We have, however, improved overall performance by decreasing the weight of words that appear more often in the text: The nth appearance of a word in the text will have only $1/n$ of its full SO value.[17] Consider the following invented example.

(17) Overall, the film was excellent. The acting was excellent, the plot was excellent, and the direction was just plain excellent.

Pragmatically, the repetition of *excellent* suggests that the writer lacks additional substantive commentary, and is simply using a generic positive word. We could also impose an upper limit on the distance of repetitions, and decrease the weight only when they appear close to each other. Repetitive weighting does not apply to words that have been intensified, the rationale being that the purpose of the intensifier is to draw special attention to them.

Another reason to tone down words that appear often in a text is that a word that appears regularly is more likely to have a neutral sense. This is particularly true of nouns. In one example from our corpus, the words *death*, *turmoil*, and *war* each appear twice. A single use of any of these words might indicate a comment (e.g., *I was bored to death*), but repeated use suggests a descriptive narrative.

2.7 Other Features of SO-CAL

Two other features merit discussion: weighting and multiple cut-offs. First of all, SO-CAL incorporates an option to assign different weights to sentences or portions of a text. Taboada and Grieve (2004) improved performance of an earlier version of the SO calculator by assigning the most weight at the two-thirds mark of a text, and significantly less at the beginning. The current version has a user-configurable form of this weighting system, allowing any span of the text (with the end points represented by fractions of the entire text) to be given a certain weight. An even more flexible and powerful system is provided by the XML weighting option. When this option is enabled, XML tag pairs in the text (e.g., <topic>, </topic>) can be used as a signal to the calculator that any words appearing between these tags should be multiplied by a certain given weight. This gives SO-CAL an interface to outside modules. For example, one module could pre-process the text and tag spans that are believed to be topic sentences, another module could provide discourse information such as rhetorical relations (Mann and Thompson 1988), and a third module could label the sentences that seem to be subjective. Armed with this information, SO-CAL can disregard or de-emphasize parts of the text that are less relevant to sentiment analysis. This weighting feature is used in Taboada, Brooke, and

17 One of the reviewers points out that this is similar to the use of term frequency (tf-idf) in information retrieval (Salton and McGill 1983). See also Paltoglou and Thelwall (2010) for a use of information retrieval techniques in sentiment analysis.

Stede (2009) to lower the weight of descriptive paragraphs, as opposed to paragraphs that contain mostly commentary.

Secondly, SO-CAL allows for multiple cut-offs. Most work in sentiment analysis has focused on binary positive/negative classification. Notable exceptions include Koppel and Schler (2005) and Pang and Lee (2005), who each adapted relevant SVM machine-learning algorithms to sentiment classification with a three- and four-class system, respectively. Because SO-CAL outputs a numerical value that reflects both the polarity and strength of words appearing in the text, it is fairly straightforward to extend the function to any level of granularity required; in particular, the SO-CAL grouping script takes a list of n cut-off values, and classifies texts into $n + 1$ classes based on text values. The evaluative output gives information about exact matches and also near-misses (when a text is incorrectly classified into a neighboring class). This allows SO-CAL to identify, for instance, the star rating that would be assigned to a consumer review, as shown in Brooke (2009).

2.8 Evaluation of Features

To test the performance of all of SO-CAL's features, we used the following data sets:

Epinions 1:
Our original collection of 400 review texts (Taboada and Grieve 2004), used in various phases of development. The collection consists of 50 reviews each of: books, cars, computers, cookware, hotels, movies, music, and phones.

Epinions 2:
A new collection of 400 texts from the epinions.com site, with the same composition as Epinions 1.

Movie:
1,900 texts from the Polarity Dataset (Pang and Lee 2004).[18]

Camera:
A 2,400-text corpus of camera, printer, and stroller reviews, taken from a larger set of Epinions reviews; also used in Bloom, Garg, and Argamon (2007).

All corpora contain equal numbers of positive and negative texts, and the last three were completely new test corpora.[19] We include our development Epinions 1 corpus in part to show how similar the performance on "familiar" text is to that for unseen texts. In corpus-based or statistical methods, it is essential that training and testing be carried out on separate corpora, because a classifier will often learn to classify its training set too well, using features that are irrelevant to the actual task; for this reason, testing data is usually set aside, or cross-validation used. This is, however, not such an important concern for a lexicon-based model, provided that the dictionary values are assigned to words based on their real-world prior polarity, and not the polarity of the text in which they appear (which is how an SVM classifier would learn its weights). There is, of course, the danger that the words chosen for the dictionaries are reflective of the domain in question. Our dictionaries do contain words that are more frequent or mostly present in review domains, and Epinions 1 influenced our choice of words to include

18 One hundred reviews of the Polarity Dataset were used for development, and thus those are excluded from our testing. The performance difference between using the full 2,000 texts or 1,900 is negligible.

19 The development corpus (Epinions 1) and two annotated versions of it, for rhetorical relations and Appraisal, are available from the project's Web site: `www.sfu.ca/~mtaboada/research/nserc-project.html`.

in the dictionaries. The performance is constant across review domains, however, and remains very good in completely new domains, which shows that there was no overfitting of the original set.

Table 4 shows a comparison using the current version of SO-CAL with various dictionary alternatives. These results were obtained by comparing the output of SO-CAL to the "recommended" or "not recommended" field of the reviews. An output above zero is considered positive (recommended), and negative if below zero.

The Simple dictionary is a version of our main dictionary that has been simplified to $2/-2$ values, switch negation, and $1/-1$ intensification, following Polanyi and Zaenen (2006). Only-Adj excludes dictionaries other than our main adjective dictionary, and the One-Word dictionary uses all the dictionaries, but disregards multi-word expressions. Asterisks indicate a statistically-significant difference using chi-square tests, with respect to the full version of SO-CAL, with all features enabled and at default settings.

These results indicate a clear benefit to creating hand-ranked, fine-grained, multiple-part-of-speech dictionaries for lexicon-based sentiment analysis; the full dictionary outperforms all but the One-Word dictionary to a significant degree ($p < 0.05$) in the corpora as a whole. It is important to note that some of the parameters and features that we have described so far (the fixed number for negative shifting, percentages for intensifiers, negative weighting, etc.) were fine-tuned in the process of creating the software, mostly by developing and testing on Epinions 1. Once we were theoretically and experimentally satisfied that the features were reasonable, we tested the final set of parameters on the other corpora.

Table 5 shows the performance of SO-CAL with a number of different options, on all corpora (recall that all but Epinions 1 are completely unseen data). "Neg w" and "rep w" refer to the use of negative weighting (the SO of negative terms is increased by 50%) and repetition weighting (the nth appearance of a word in the text has $1/n$ of its full SO value). Space considerations preclude a full discussion of the contribution of each part of speech and sub-feature, but see Brooke (2009) for a full range of tests using these data. Here the asterisks indicate a statistically-significant difference compared to the preceding set of options.

As we can see in the table, the separate features contribute to performance. Negation and intensification together increase performance significantly. One of the most important gains comes from negative weighting, with repetition weighting also contributing in some, but not all, of the corpora. Although the difference is small, we see here that shifted polarity negation does, on average, perform better than switched polarity negation. We have not presented all the combinations of features, but we know from other experiments, that, for instance, basic negation is more important than basic

Table 4
Comparison of performance using different dictionaries.

Dictionary	Percent correct by corpus				
	Epinions 1	**Epinions 2**	**Movie**	**Camera**	**Overall**
Simple	76.75	76.50	69.79*	78.71	75.11*
Only-Adj	72.25*	74.50	76.63	71.98*	73.93*
One-Word	80.75	80.00	75.68	79.54	78.23
Full	80.25	80.00	76.37	80.16	78.74

*Statistically significant using the chi-square test, $p < 0.05$.

Table 5
Performance of SO-CAL using various options.

SO-CAL options	Percent correct by corpus				
	Epinions 1	**Epinions 2**	**Movie**	**Camera**	**Overall**
All words (nouns, verbs, adjectives, adverbs)	65.50	65.25	68.05	64.70	66.04
All words + negation (shift)	67.75	67.25	70.10	67.25*	68.35*
All words + neg (shift) + intensification	69.25	71.50	73.47*	70.00*	71.35*
All words + neg (shift) + int + irrealis blocking	71.00	71.25	74.95	71.37	72.66*
All words + neg (shift) + int + irrealis + neg w	**81.50***	78.25*	75.08	78.24*	77.32*
All words + neg (shift) + int + modals + neg w + rep w	80.25	**80.00**	**76.37**	**80.16***	**78.74***
All words + neg (switch) + int + modals + neg w + rep w	80.00	**80.00**	75.57	80.04	78.37

*Statistically significant compared to the preceding set of options (Table 4) $p < 0.05$.

intensification, as also shown by Kennedy and Inkpen (2006). In summary, it is the combination of carefully crafted dictionaries and features inspired by linguistic insights that we believe makes SO-CAL a robust method for sentiment analysis.

Table 6 shows the accuracy of the SO-CAL system across different review types in Epinions 1 and Epinions 2 (the latter unseen data), including the F-measure for classification of positive and negative texts. The table shows that SO-CAL's performance on positive reviews is generally better than on negative reviews (run with all options and shifted negation). This is despite the fact that all of our dictionaries contain far more negative words than positive ones. As noted already by Boucher and Osgood (1969), there is a strong preference for avoiding negation and negative terms even when expressing negative opinions, making the detection of text sentiment difficult for systems which depend solely on these indicators (also see results in Dave, Lawrence, and Pennock 2003, Turney 2002). The exception are books and movies, where performance is more even across positive and negative, or often better in negative reviews. We hypothesize this is because consumer product reviews contain more factual information that the reader is required to interpret as positive or negative (for instance, the range for a cordless phone or the leg room in the back seat of a car). Some of that factual information may be implicitly negative, and therefore missed by SO-CAL.

The breakdown for the Polarity Dataset (Movies in Table 5) is 89.37% precision for negative reviews and 63.2% for positive ones, with an overall accuracy of 76.37%. A number of other papers have used the Polarity Dataset created by Bo Pang, most of them following statistical methods. Pang and Lee's own results show an overall accuracy of 87.15% for polarity classification of whole reviews (Pang and Lee 2004). Fletcher and Patrick (2005) used bags-of-words that included Appraisal features, and obtained 83.7% accuracy in that same data set, whereas Whitelaw, Garg, and Argamon (2005), using bags-of-words combined with Appraisal groups achieved 90.2%. In all cases, the accuracy reflects a single domain and data set. Andreevskaia and Bergler (2008) show, however, that cross-domain performance drops significantly. They used a hybrid

Table 6
Performance across review types and on positive and negative reviews.

Subcorpus	Epinions 1			Epinions 2		
	Pos-F	Neg-F	Accuracy	Pos-F	Neg-F	Accuracy
Books	0.69	0.74	0.72	0.69	0.77	0.74
Cars	0.90	0.89	0.90	0.80	0.75	0.78
Computers	0.94	0.94	0.94	0.90	0.89	0.90
Cookware	0.74	0.58	0.68	0.79	0.76	0.78
Hotels	0.76	0.67	0.72	0.80	0.70	0.76
Movies	0.84	0.84	0.84	0.76	0.79	0.78
Music	0.82	0.82	0.82	0.83	0.81	0.82
Phones	0.81	0.78	0.80	0.85	0.83	0.84
Total	0.81	0.79	0.80	0.81	0.79	0.80

method, combining statistical and lexicon-based approaches, on reviews (the Polarity Dataset and product reviews), news, and blogs, with an average accuracy across all domains of 71.1% (on sentences, not full texts).

We will, in Section 3.4, provide a more detailed comparison of SO-CAL's dictionaries to other dictionaries. Although the results presented in this section are below those of some statistical methods, we argue that our system performs more evenly across domains, and can be enhanced with contextual sources of information. We show performance in non-review domains in the next section.

We defined, in the introduction, sentiment as polarity plus strength, although the results presented here evaluate only the polarity accuracy. Space precludes a full discussion of SO-CAL's measure of strength, but Brooke (2009) shows that SO-CAL's output correlates well with star ratings in reviews.

2.9 Evaluation of SO-CAL in Other Domains

Reference to domain portability, in this article and in other work, is usually limited to portability across different types of reviews (Aue and Gamon 2005; Blitzer, Dredze, and Pereira 2007; Andreevskaia and Bergler 2008). This section shows that SO-CAL's performance is maintained across domains other than reviews, and across different types of text structures. Even though SO-CAL was primarily designed to determine the sentiment of texts roughly a paragraph or longer, the evaluations reported in this section demonstrate comparable performance when applied to shorter texts such as headlines and sentences extracted from news and blogs.

We tested SO-CAL with four different data sets: the Multi-Perspective Question Answering (MPQA) corpus, version 2.0 (Wiebe, Wilson, and Cardie 2005); a collection of MySpace.com comments from Mike Thelwall (Prabowo and Thelwall 2009); a set of news and blog posts from Alina Andreevskaia (Andreevskaia and Bergler 2008); and a set of headlines from Rada Mihalcea and Carlo Strappavara (Strappavara and Mihalcea 2007).[20]

The first set of data is the MPQA corpus (version 2.0), a collection of news articles and other documents (texts from the American National Corpus and other sources) annotated for opinions and other private states (beliefs, emotions, speculation, etc.). We

20 We thank them all for sharing their data with us.

extracted all the sentences that contained subjective positive and negative expressions, in all levels of intensity (low, medium, high, and extreme). The extracted set contains 663 positive and 1,211 negative sentences.

The data from Mike Thelwall consists of comments posted on MySpace.com. The annotation is done on a 1 to 5 scale, where 1 indicates "no emotion." As a consequence, we focused on the comments with scores of 4 and 5. Because each comment had both a positive and negative label, we labeled "positive" those with a higher positive score and vice versa for negative labels, and excluded comments with the same score for both (i.e., neutral). This yielded a total of 83 comments (59 positive, 24 negative).

The data from Alina Andreevskaia consist of individual sentences from both news and blogs, annotated according to whether they are negative, positive, or neutral. We used only the negative and positive sentences (788 from news, and 802 from blogs, equally divided between positive and negative).

The Affective Text data from Rada Mihalcea and Carlo Strappavara was used in the 2007 SemEval task. It contains 1,000 news headlines annotated with a range between −100 (very negative) and 100 (very positive). We excluded six headlines that had been labeled as 0 (therefore neutral), yielding 468 positive and 526 negative headlines. In addition to the full evaluation, Strappavara and Mihalcea (2007) also propose a coarse evaluation, where headlines with scores −100 to −50 are classified as negative, and those 50 to 100 as positive. Excluding the headlines in the middle gives us 155 positive headlines and 255 negative ones.

Table 7 shows the results of the evaluation. Included in the table is a baseline for each data set, assigning polarity to the most frequent class for the data. These data sets include much smaller spans of text than are found in consumer reviews, with some sentences or headlines not containing any words from the SO-CAL dictionaries. This ranged from about 21% of the total in the MySpace comments to 54% in the headlines.[21] Two approaches were used in this cross-domain evaluation when SO-CAL encountered texts for which it found no words in its dictionaries (SO-empty texts). First, the back-off method involves using the most frequent polarity for the corpus (or positive, when they are equal), and assigning that polarity to all SO-empty texts. This method provides results that can be directly compared to other results on these data sets, although, like the baseline, it assumes some knowledge about the polarity balance of the corpus. The figures in the first section of Table 7 suggest robust performance as compared to a most-frequent-class baseline, including modest improvement over the relevant cross-domain results of Andreevskaia and Bergler (2008).[22] Moilanen, Pulman, and Zhang (2010) also use the headlines data, and obtain a polarity classification accuracy of 77.94% below our results excluding empty.[23]

21 By default, SO-CAL assigns a zero to such texts, which is usually interpreted to mean that the text is neither positive nor negative. However, in a task where we know a priori that all texts are either positive or negative, this can be a poor strategy, because we will get all of these empty texts wrong: When there are a significant number of empty texts, performance can be worse than guessing. Note that the problem of how to interpret empty texts is not a major issue for the full text reviews where we typically apply SO-CAL, because there are very few of them; for instance, out of the 2,400 texts in the Camera corpus, only 4 were assigned a zero by SO-CAL. Guessing the polarity or removing those four texts entirely has no effect on the accuracy reported in Table 5.

22 Their ensemble classifier had 73.3% accuracy in news, but only 70.9% in blogs, and their performance in the Polarity Dataset was 62.1%, or over 14% lower than ours.

23 Our results are not comparable to those of Thelwall et al. (2010) on the MySpace comments, as they classify the comments on a 1–5 scale (obtaining average accuracy of 60.6% and 72.8% in positive and negative comments, respectively), whereas we have a much simpler two-point scale (positive or negative).

Table 7
Performance of SO-CAL in other domains.

SO-CAL options	Percent correct by corpus					
	MPQA	**MySpace**	**News**	**Blogs**	**Headlines**	**Headlines (coarse)**
SO-CAL (back-off)	73.64	81.93	71.57	75.31	62.17	74.63
Baseline (most frequent class)	64.64	71.08	50.00	50.00	52.92	62.20
SO-CAL (excluding empty)	79.38	78.69	77.76	82.33	79.83	88.98
Baseline (most frequent class, excluding empty)	66.94	69.93	51.10	50.00	59.87	67.37
% SO-empty	28.61	21.12	25.25	21.70	54.00	43.98

The second method used in evaluating SO-CAL on SO-empty texts is to only classify texts for which it has direct evidence to make a judgment. Thus, we exclude such SO-empty texts from the evaluation. The second part of Table 7 shows the results of this evaluation. The results are strikingly similar to the performance we saw on full review texts, with most attaining a minimum of 75–80% accuracy. Although missing vocabulary (domain-specific or otherwise) undoubtedly plays a role, the results provide strong evidence that relative text size is the primary cause of SO-empty texts in these data sets. When the SO-empty texts are removed, the results are entirely comparable to those that we saw in the previous section. Although sentence-level polarity detection is a more difficult task, and not one that SO-CAL was specifically designed for, the system has performed well on this task, here, and in related work (Murray et al. 2008; Brooke and Hurst 2009).

3. Validating the Dictionaries

To a certain degree, acceptable performance across a variety of data sets, and, in particular, improved performance when the full granularity of the dictionaries is used (see Table 5), provides evidence for the validity of SO-CAL's dictionary rankings. Recall also that the individual word ratings provided by a single researcher were reviewed by a larger committee, mitigating some of the subjectivity involved. Nevertheless, some independent measure of how well the dictionary rankings correspond to the intuitions of English speakers would be valuable, particularly if we wish to compare our dictionaries with automatically generated ones.

The most straightforward way to investigate this problem would be to ask one or more annotators to re-rank our dictionaries, and compute the inter-annotator agreement. However, besides the difficulty and time-consuming nature of the task, any simple metric derived from such a process would provide information that was useful only in the context of the absolute values of our -5 to $+5$ scale. For instance, if our re-ranker is often more conservative than our original rankings (ranking most SO 5 words as 4, SO 4 words as 3, etc.), the absolute agreement might approach zero, but we would like to be able to claim that the rankings actually show a great deal of consistency given

their relative distribution. To accomplish this, we focus on relative comparisons rather than absolute ranking.

In retrospect, the use of a 10-point scale is somewhat arbitrary, and so another goal of our explorations is to test whether we can validate our choice of granularity of scale. Generally, our concerns are general patterns, not individual words, and to what extent those patterns coincide with the discrete, linear model that we have assumed; however, we also hope to obtain information that might show some inconsistencies for particular, commonly used words.

In this section, we perform two types of validation. First, we compare the dictionary scores to scores provided by human annotators, recruited through Amazon's Mechanical Turk service. Second, we compare our dictionaries and their performance to a few other available dictionaries. In both cases, the results show that our dictionaries are robust.

3.1 Data Collection

To collect data on the validity of our dictionary, we made use of Amazon's Mechanical Turk service,[24] which provides access to a pool of workers who have signed up to perform small-scale tasks that require human intelligence. Mechanical Turk is quickly becoming a popular resource among computational linguists, and has been used in sentiment, emotion, and subjectivity tasks (Akkaya et al. 2010; Mellebeek et al. 2010; Mohammad and Turney 2010; Yano, Resnik, and Smith 2010), although there are mixed reports on its reliability (Snow et al. 2008; Callison-Burch 2009; Zaenen to appear).

In short, the Mechanical Turk service acts as a marketplace that connects Requesters, people who have tasks that require human intelligence, and Workers, or "Turkers," people who will perform such tasks for a small payment. Typical tasks include tagging images, transcribing spoken data, or finding information on the Web. Payments for each individual task may be as low as $0.01. Quality is controlled through several measures, which include the request of Turkers with specific qualifications, or with high approval ratings for previously completed tasks. Requesters may also reject a Turker's work if it has been incorrectly completed, in which case they are not paid, also bringing their approval rating down.

Our early testing suggested that if we restricted our task to those Turkers who had an extremely high approval rating (99%), the results were mostly acceptable, with a certain amount of noise inherent in the use of anonymous Turkers. Under the circumstances, keeping the task simple was important. Note that, although we have six sets of results for each data set, this is not analogous to having six annotators, since each set included work by multiple Turkers. We can, however, be sure that no Turker provided an answer to the same question twice, and for the purposes of this study a more diverse set of responses may actually be viewed as a positive.

Carrying out an evaluation on all five dictionaries (nouns, verbs, adjectives, adverbs, and intensifiers) would have been somewhat redundant, because most verbs have a (synonymous) noun counterpart, and most adverbs are derived from an adjective. Here, we focus mostly on the adjective dictionary, which is the primary dictionary in terms of overall size as well as frequency of occurrence. We began with our full adjective dictionary (minus multi-word expressions), but filtered using word counts from our two large corpora (the 2,000 text Polarity Dataset and 3,000 text Camera

24 www.mturk.com/.

corpus), including a word in the analysis only if it appeared at least five times in both corpora. That gave us a collection of 483 commonly occurring adjectives; these, as well as our intensifier dictionary, were the focus of our evaluation. We also investigated a set of nouns chosen using the same rubric. In most cases, the results were comparable to those for adjectives; however, with a smaller test set (only 184 words), the data were generally messier, and so we omit the details here.

The basic premise behind our evaluation technique can be described as follows: The distributional spread of answers in a simple, three-way decision task should directly reflect the relative distance of words on an SO (semantic orientation) scale. In particular, we can validate our dictionaries without forcing Turkers to use our 11-point scale (including zero), making their task significantly less onerous as well as less subject to individual bias. We chose to derive the data in two ways: one task where the goal is to decide whether a word is positive, negative, or neutral; and another where two polar words are presented for comparison (Is one stronger, or are they the same strength?). In the former task, we would expect to see more "errors," that is, cases where a polar term in our dictionary is labeled neutral, in words that are in fact more neutral (SO value 1 versus SO value 5). Similarly, in the second task we would expect to see more "equal" judgments of words which are only 1 SO unit apart than those which are 3 SO units apart. More formally, we predict that a good ranking of words subjected to this testing should have the following characteristics:

- The percentage of "equal/neutral" judgments should be maximized at exactly the point we would expect, given the SO value assigned to the word(s) in question. The number of polar responses should be relatively balanced.
- As we move away from this point (SO = 0) in either direction, we would expect a linear increase in the appropriate polar response, and a corresponding drop in the other responses.
- The percentage of polar responses may hit a maximum, after which we would expect a consistent distribution (within a small margin of error; this maximum might not be 100%, due to noise in the data).

For the first task, that is, the neutral/negative/positive single word decision task, we included neutral words from our Epinions corpus which had originally been excluded from our dictionaries. Purely random sampling would have resulted in very little data from the high SO end of the spectrum, so we randomly selected first by SO (neutral SO = 0) and then by word, re-sampling from the first step if we had randomly selected a word that had been used before. In the end, we tested 400 adjectives chosen using this method. For each word, we solicited six judgments through Mechanical Turk.

Preparing data for the word comparison task was slightly more involved, because we did not want to remove words from consideration just because they had been used once. Note that we first segregated the words by polarity: We compared positive words with positive words and negative words with negative words. This first test yielded a nice wide range of comparisons by randomly selecting first by SO and then by word, as well as by lowering the probability of picking high (absolute) SO words, and discounting words which had been used recently (in early attempts we saw high absolute SO words like *great* and *terrible* appearing over and over again, sometimes in consecutive queries). Though the odds of this occurring were low, we explicitly disallowed duplicate

pairs. Once we settled on a method, we created 500 pairs of positive adjectives and 500 pairs of negative adjectives. Again, for each pair we solicited six judgments.

In addition to our "standard" pair comparison data sets, we also created, using the same method, four data sets which compared negated words to words of opposite polarity (i.e., *not bad* and *good*). The primary goal here was not to evaluate the ranking of the words, but rather to see how well our two models of negation (switch and shift) correspond to human intuition across a wide variety of cases. To do so, we assume that our dictionary is generally correct, and then use the SO values after negation as input.

Finally, we wanted to evaluate the intensifier dictionary, again using pair-wise comparisons of strength. To this end, we selected 500 pairs of adverbial modifiers (e.g., *very*). Similar to the main dictionary pairs, we randomly selected first by modifier value, and then by word, discounting a pair if one of the words had appeared in the 10 most recently selected pairs. These words were presented with a uniform adjective pairing (*likeable*), to assist the Turkers in interpreting them.

3.2 Evaluation

Figure 1 shows the distribution of responses by SO value in the single-word identification task. The graph is very close to what we predicted. Neutral judgments peak at 0 SO, but are also present for those SO values in the neighborhood of 0, decreasing as we increase our SO distance from the original. The effect is not quite linear, which might reflect either on our scale (it is not as linear as we presume) or, more likely, the fact that the distance between 0 and 1/−1 is simply a much more relevant distance for the purposes of the task; unlike words that differ in strength, the difference between 0 and 1 is theoretically a difference in kind, between a word that has a positive or negative connotation, and one that is purely descriptive. Another limitation of this method, with respect to confirming our dictionary rankings, is the fact that it does not illuminate the edges of the spectrum, as the distributions hit their maximums before the 5/−5 extreme.

Because we asked six Turkers to provide responses for each word, we can also calculate average percentage of pairwise agreement (the number of pairs of Turkers who agreed, divided by the total number of possible pairings), which for this task was 67.7%, well above chance but also far from perfect agreement. Note that we are not trying to establish reliability in the traditional sense. Our method depends on a certain

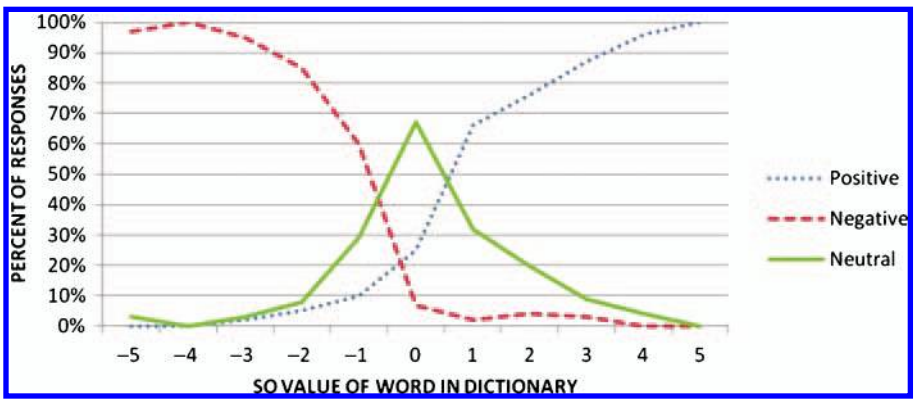

Figure 1
Distribution of responses by SO value, single-word task.

amount of disagreement; if it were simply the case that some −1 words were judged by all rankers as neutral (or positive), the best explanation for that fact would be errors in individual SO values. If Turkers, however, generally agree about SO 5/−5 words, but generally disagree about SO 1/−1 words, this "unreliability" actually reflects the SO scale. This is indeed the pattern we see in the data: Average pairwise agreement is 60.1% for 1/−1 words, but 98.2% for 5/−5 (see Andreevskaia and Bergler [2006] for similar results in a study of inter-annotator agreement in adjective dictionaries).[25]

Interestingly, although we expected relatively equal numbers of positive and negative judgments at SO = 0, that was not the result. Instead, words with SO = 0 were sometimes interpreted as positive, but almost never interpreted as negative. This is mostly likely attributable to the default status of positive, and the marked character of negative expression (Boucher and Osgood 1969; Horn 1989; Jing-Schmidt 2007); neutral description might be taken as being vaguely positive, but it would not be mistaken for negative expression.[26]

For the word-pair task, we categorize the distribution data by the difference in SO between the two words, putting negative and positive words in separate tables. For instance, a −4 difference with negative adjectives means the result of comparing a −1 word to a −5 word, and a +2 difference corresponds to comparing a −5 with a negative −3, a −4 with a −2, or a −3 with a −1. (We always compare words with the same sign, i.e., negative to negative.) In early testing, we found that Turkers almost completely ignored the *same* category, and so we took steps (changing the instructions and the order of presentation) to try to counteract this effect. Still, the *same* designation was underused. There are a number of possible explanations for this, all of which probably have some merit. One is that our scale is far too coarse-grained, that we are collapsing distinguishable words into a single classification. The trade-off here is with ease of ranking; if we provided, for instance, 20 SO values instead of 10, it would be more difficult to provide confident rankings, and would probably yield little in the way of tangible benefits.[27] Another potential confounding factor is that words within the same SO category often vary considerably on some other dimension, and it is not natural to think of them as being equivalent. For instance, we judged *savory*, *lush*, and *jolly* to be equally positive, but they are applied to very different kinds of things, and so are not easily compared. And, of course, even assuming our 10-point scale, there are words in our dictionary that do not belong together in the same category; our focus here is on the big picture, but we can use this data to identify words which are problematic and improve the dictionary in the next iteration.

The results in Figures 2 and 3 for the adjective word pair task are otherwise very encouraging. Unlike the single word task, we see a clear linear pattern that covers the entire SO spectrum (though, again, there is noise). At SO value difference = 0, *same* reaches a maximum, and positive and negative judgments are almost evenly distributed. The average pairwise agreement on this task was somewhat lower, 60.0%

25 Note also that another drawback with pairwise agreement is that agreement does not change linearly with respect to the number of dissenters. For example, in the six-rater task, a single disagreement drops agreement from 100% to 66.7%; a second disagreement drops the score to 40% if different than the first agreement, or 46.7% if the same.

26 This is the opposite result from the impressions reported by Cabral and Hortaçsu (2010), where, in an evaluation of comments for eBay sellers, neutral comments were perceived as close to negative.

27 Other evidence that suggests making our scale more fine-grained is unlikely to help: When two words were difficult to distinguish, we often saw three different answers across the six Turkers. For example, for −3 SO words *fat* and *ignorant*, three Turkers judged them the same, two judged *ignorant* as stronger, and one judged *fat* as stronger.

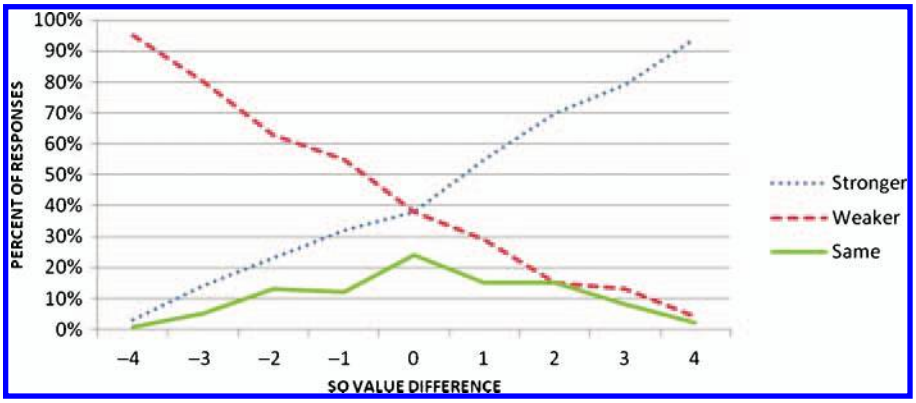

Figure 2
Distribution of responses by SO difference for positive adjectives, word-pair task.

and 63.7% for Figure 2 and Figure 3, respectively. This is not surprising, because the vast majority of the data come from the difficult-to-judge range between +2 and −2. Outside of this range, agreement was much higher: Both experiments showed roughly 50% agreement when the SO value of the two words was the same, and an increase of approximately 10% for each point difference in SO value.

Figure 4 shows the results for adverbial intensifiers. Pairwise agreement here was higher, at 68.4%. The basic trends are visible; there is, however a lot of noise throughout. This is the drawback of having a relatively fine-grained scale, and in retrospect we perhaps should have followed the model for adjectives and split our words further into downplayers and amplifiers. The other reason for fluctuations, particularly at the extremes, was our inclusion of comparative and pragmatic intensifiers like *more*, *less*, *the most*, *barely*, *hardly*, *almost*, and *not only*, which, unlike regular scalar intensifiers (*very*, *immensely*), are very difficult to interpret outside of a discourse context, and are not easily compared.

For the negation evaluation task, we can state directly the result of comparing a positive word with a negated negative word: Outside of colloquial expressions such as *not bad*, it is nearly impossible to express any positive force by negating a negative; the percentage of negated negatives that were ranked higher than positives was about 5%

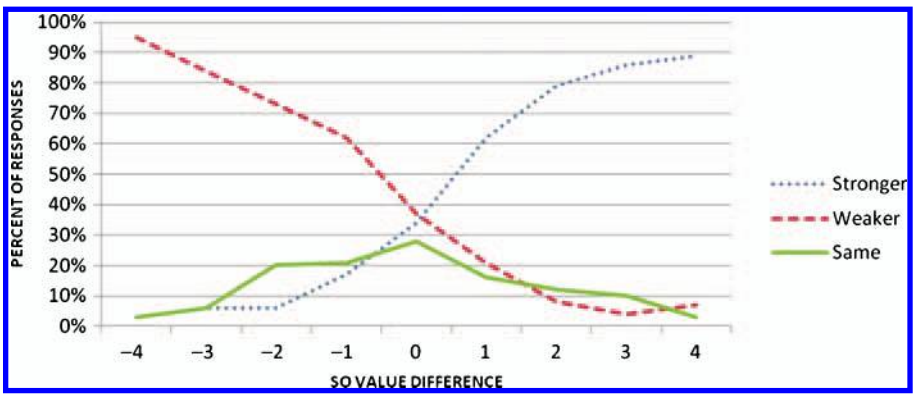

Figure 3
Distribution of responses by SO difference for negative adjectives, word-pair task.

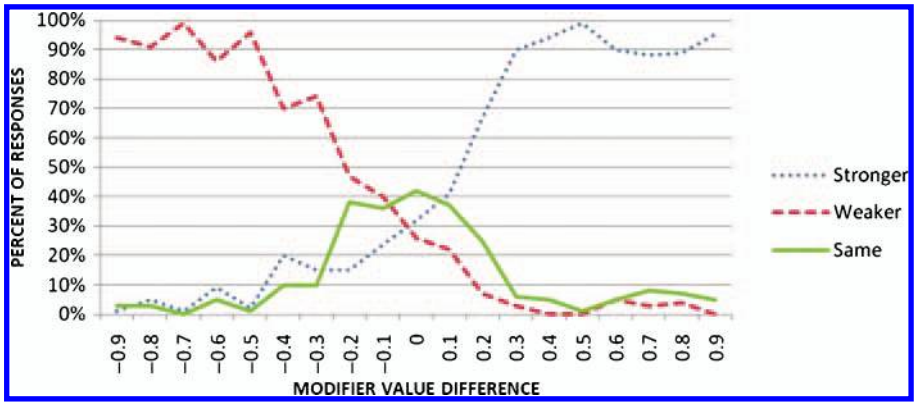

Figure 4
Distribution of responses by modifier value. Difference for adverbial intensifiers.

Table 8
Distribution percentages for the negative/negated positive SO comparison task.

Word	−1			−2			−3			−4			−5		
SO	pos	neg	neu	pos	neg	neu	pos	neg	neu	pos	neg	neu	pos	neg	neu
1	50	19	32	16	54	30	8	78	14	4	82	14	0	95	5
2	47	32	21	11	66	23	7	67	27	8	72	20	4	86	11
3	39	26	35	17	67	17	10	84	5	3	95	3	0	9	10
4	36	36	29	31	45	23	20	68	12	12	82	5	0	93	7
5	25	45	30	25	58	17	17	61	22	0	95	5	6	72	22

(a result that was duplicated for noun negation), concentrated mostly on SO 1 positive words (which, as we have seen, are sometimes viewed as neutral). This result is not predicted by either of our models of negation (switch and shift), but it may be somewhat irrelevant because negated negatives, being essentially a double negative, are fairly rare. The main use of negation, we have found, is to negate a positive word.

Table 8 shows the distribution percentages for the negative/negated positive SO comparison task. Here, *pos* refers to the percentage of people who rated the negated positive word as being stronger, *neg* refers to the percentage of people who rated the negative word as being stronger, and *neu* refers to a judgment of *same*. Pairwise agreement across raters on this task was only 51.2%, suggesting that the comparisons involving negatives are the most difficult of our tasks.[28]

As the SO value of the negative word increases, we of course expect that it is judged stronger, a pattern visible from left to right in Table 8. The more interesting direction is from top to bottom: If the switch model is correct, we expect increasing judgments in favor of the (negated) positive word, but if the shift model is correct, we would see the opposite. The results in Table 8 are not conclusive. There are aspects of the

28 We in fact saw even lower agreement than this after our initial data collection. We investigated the low agreement, and attributed it to a single Turker (identified by his/her ID) who exhibited below-chance agreement with other Turkers. A visual inspection of the data also indicated that this Turker, who provided more responses than any other, either did not understand the task, or was deliberately sabotaging our results. We removed this Turker's data, and solicited a new set of responses.

table consistent with our shift model, for instance a general decrease in *pos* judgments between *pos* SO 3 and 5 for lower *neg* SO. However, there are a number of discrepancies. For instance, the shift model would predict that a negated +1 word is stronger than a −2 word (+1 becomes −3 under negation), which is often not the case. Note also that the shift trend seems somewhat reversed for higher negative SOs. In general, the shift model accounts for 45.2% of Mechanical Turk (MT) judgments (see our definition of MT correspondence, in the following section), whereas the switch model accounts for 33.4%. Negation is clearly a more complicated phenomenon than either of these simple models can entirely represent, although shifting does a somewhat better job of capturing the general trends.

3.3 Dictionary Comparisons

We now turn to using the data that we have collected to evaluate other dictionaries and scales. We use the same Mechanical Turk judgments as in the previous section, with six Turkers per word or pair. For simplicity, we look only at the single-word task and pairwise comparison of negative adjectives. We chose negative words because they are better distinguished by our automatic classifiers. Note that our definition of negative adjective is tied to our original SO ratings, and has been integrated into the selection of pairs for Mechanical Turk. At this stage, we hope to have shown that the adjective dictionary does a sufficiently accurate job of distinguishing positive, negative, and neutral words, and provides a good range of words within those categories, with which other dictionaries can be tested. Here, we will use the term **Mechanical Turk (MT) correspondence** as follows:

$$\textit{MT correspondence} = \frac{\text{Mech. Turk judgments predicted by dictionary}}{\text{Total Mech. Turk judgments}}$$

For example, if one rater thought A was more positive than B, and the other thought they were of the same strength, then an SO dictionary which predicts either of these results would have 50% MT correspondence (on this word), whereas a dictionary where the SO value of B is greater than A would have 0% MT correspondence. As an absolute measure, correspondence is somewhat misleading: Because there are disagreements among Turkers, it is impossible for a dictionary to reach 100% MT correspondence. For instance, the highest possible MT correspondence in the single-word task is 79%, and the highest possible MT correspondence for the negative adjective task is 76.8%. MT correspondence is useful as a relative measure, however, to compare how well the dictionaries predict MT judgments.

Our first comparison is with the dictionary of adjectives that was derived using the SO-PMI method (Turney 2002), using Google hit counts (Taboada, Anthony, and Voll 2006). The SO values for the words tested here vary from 8.2 to −5.74. We have already noted in Section 2 that using this dictionary instead of our manually-ranked one has a strongly negative effect on performance. Because the Google dictionary is continuous, we place the individual SO values into evenly spaced buckets so that we can graph their distribution. For easy comparison with our dictionary, we present the results when buckets equivalent to our 11-point SO scale are used. The results for the single-word task are given in Figure 5.[29]

29 When bucketing for the single word task, we used a neutral (zero) bucket that was twice the size of the other buckets, reflecting the fact that zero is a more significant point on the scale in this context.

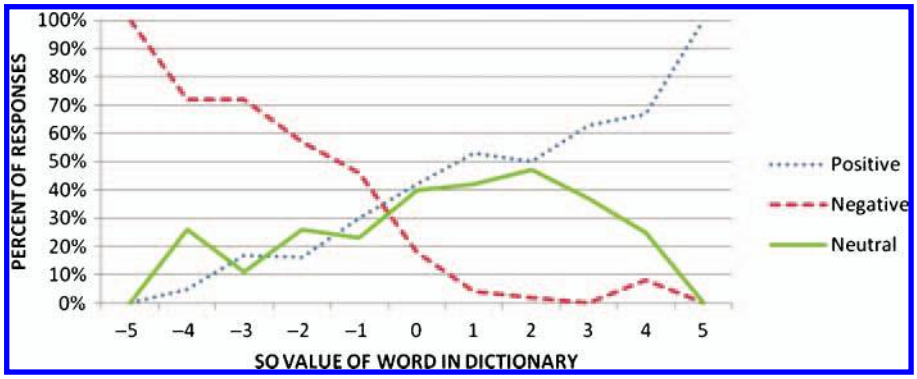

Figure 5
Distribution of responses by adjective SO value for Google PMI dictionary, single-word task.

Figure 6
Distribution of responses by adjective SO value for Google PMI dictionary, negative word-pair task.

As compared to the manually ranked SO dictionary, the Google PMI dictionary does not maximize as quickly, suggesting significant error at even fairly high SO values. Interestingly, the graph shows a striking similarity with the manually ranked dictionary in terms of the asymmetry between positive and negative words; negative words are almost never ranked as positive, although the reverse is not true. The neutral curve peaks well into the positive SO range, indicating that neutral and positive words are not well distinguished by the dictionary.[30] Overall, the SO-PMI dictionary correctly predicts 48.5% of the Mechanical Turk rankings in this task, which places it well below the manually ranked adjective dictionary (73.7%).

Figure 6 shows the results for the negative adjective comparison task using the Google PMI dictionary. Here, the Google PMI dictionary performs fairly well, comparable to the manual rankings, though the overall MT correspondence is somewhat lower, 47% to 64%. This is partially due to bunching in the middle of the scale. Recall that the highest possible MT correspondence for this task is 76.8%. MT correspondence of

30 Distinguishing neutral and polar terms, sentences, or texts is, in general, a hard problem (Wilson, Wiebe, and Hwa 2004; Pang and Lee 2005).

nearly 55% is possible if the number of buckets is increased significantly, an effect which is due at least partially to the fact that the *same* designation is so underused that it is generally preferable to always guess that one of the adjectives is stronger than the other. Along with the results in the previous figure, this suggests that this method actually performs fairly well at distinguishing the strength of negative adjectives; the problem with automated methods in general seems to be that they have difficulty properly distinguishing neutral and positive terms.

Our next comparison is with the Subjectivity dictionary of Wilson, Wiebe, and Hoffmann (2005). Words are rated for polarity (positive or negative) and strength (weak or strong), meaning that their scale is much more coarse-grained than ours. The dictionary is derived from both manual and automatic sources. It is fairly comprehensive (over 8,000 entries), so we assume that any word not mentioned in the dictionary is neutral. Figure 7 shows the result for the single word task.

The curves are comparable to those in Figure 1; the neutral peak is significantly lower, however, and the positive and negative curves do not reach their maximum. This is exactly what we would expect if words of varying strength are being collapsed into a single category. The overall MT Correspondence, however, is comparable (71.8%).

The negative adjective pair comparison task (shown in Figure 8) provides further evidence for this (Strong/Weak means a weak negative word compared with a strong negative word).

The MT correspondence is only 48.7% in this task. There is a clear preference for the predicted judgment in weak/strong comparisons, although the distinction is far from unequivocal, and the overall change in neutrality across the options is minimal. This may be partially attributed to the fact that the strong/weak designation for this dictionary is defined in terms of whether the word strongly or weakly indicates subjectivity, not whether the term itself is strong or weak (a subtle distinction). However, the results suggest that the scale is too coarse to capture the full range of semantic orientation.

Another publicly available corpus is SentiWordNet (Esuli and Sebastiani 2006; Baccianella, Esuli, and Sebastiani 2010), an extension of WordNet (Fellbaum 1998) where each synset is annotated with labels indicating how objective, positive, and negative the terms in the synset are. We use the average across senses for each word given in version 3.0 (see discussion in the next section). Figure 9 gives the result for the

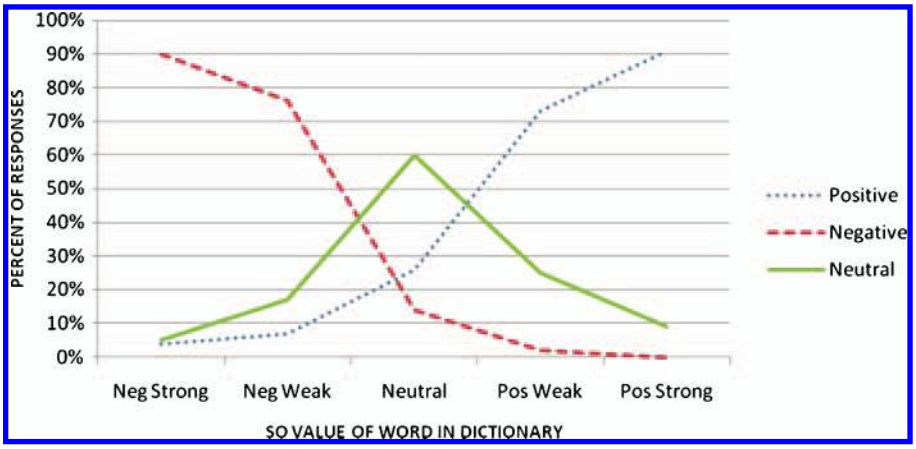

Figure 7
Distribution of responses by adjective SO value for Subjectivity dictionary, single-word task.

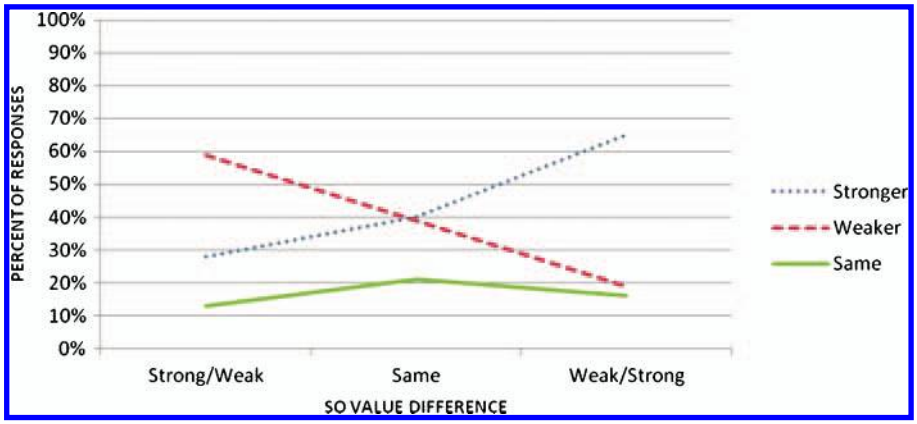

Figure 8
Distribution of responses by adjective SO value for Subjectivity dictionary, negative word-pair task.

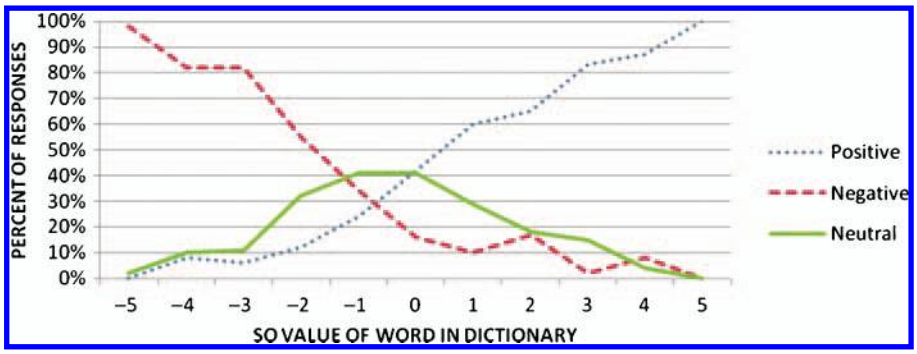

Figure 9
Distribution of responses by adjective SO value for SentiWordNet, single-word task.

single-word task. Like the Subjectivity dictionary, this dictionary is an improvement over the fully automatic Google dictionary,[31] with overall MT correspondence of 57.8%, although still well below the manual dictionaries.

Figure 10 shows the curve for the negative adjective comparison. The general trends are visible, but there is significant error at the edges of the spectrum, and a much less pronounced neutral curve. The MT correspondence with 5 SO buckets is 48.3%, although this can be boosted to around 52% by drastically increasing the number of buckets, which overall suggests it is roughly equivalent or slightly worse than the Google dictionary with respect to correctly identifying relative strength.

In summary, we have shown that our dictionaries are robust in comparison to scores provided by human raters, and that they show higher agreement with human raters than other publicly available dictionaries. More words may be added, and the scores may be changed, but the dictionary comparisons show that creating dictionaries

31 Although the values in SentiWordNet itself are calculated automatically, they are based on the knowledge from WordNet. This is why we believe it is not a fully automatic dictionary. In addition, version 3.0 includes the possibility of user feedback.

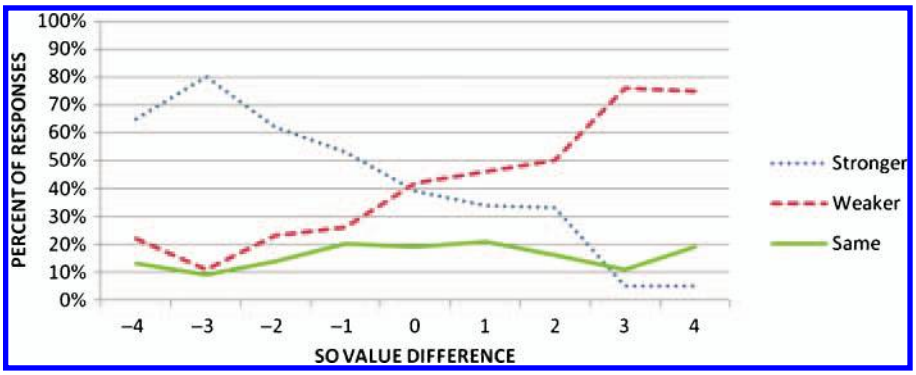

Figure 10
Distribution of responses by adjective SO value for SentiWordNet, negative word-pair task.

by hand is not necessarily a subjective task. Reflecting on our experience, we can say that the manually created dictionaries are superior for two main reasons. First of all, we tended to exclude words with ambiguous meaning, or that convey sentiment only in some occasions, but not in most. Secondly, judicious restraint is necessary when expanding the dictionaries. We found that adding more words to the dictionaries did not always help performance, because new words added noise. The type of noise that we refer to is that deriving from problems with word sense disambiguation, part-of-speech tagging, or simply strength of the word itself. Words with stronger positive or negative connotations tend to be more informative. We made a decision to exclude, of course, all neutral words (those that would have a value of 0), but also words with only a mild sentiment orientation, although there are some 1 and −1 words in the dictionaries.

3.4 SO-CAL with Other Dictionaries

The previous section provided comparisons of our dictionary to existing dictionaries. In this section, we use those lexicons and others to carry out a full comparison using SO-CAL. For each of the dictionaries discussed below, we used it instead of our set of manually ranked dictionaries as a source of SO values, and tested accuracy across different corpora. Accuracy in this case is calculated for the polarity identification task, that is, deciding whether a text is negative or positive, using the author's own ranking ("recommended" or "not recommended," or number of stars).

In our comparisons, we tested two options: *Full* uses all the default SO-CAL features described in Section 2, including intensification and negation.[32] *Basic*, on the other hand, is just a sum of the SO value of words in relevant texts, with none of the SO-CAL features enabled.

The first dictionary that we incorporated into SO-CAL was the Google-generated PMI-based dictionary described in Taboada, Anthony, and Voll (2006), and mentioned earlier in this article.

32 Except for tests with the Maryland dictionary, where we disabled negative weighting because with weighting the performance was close to chance (all negative correct, all positive wrong). We believe that this is because the dictionary contains a disproportionately large number of negative words (likely the result of expanding existing dictionaries, which also tend to include more negative than positive words).

The "Maryland" dictionary (Mohammad, Dorr, and Dunne 2009) is a very large collection of words and phrases (around 70,000) extracted from the Macquarie Thesaurus. The dictionary is not classified according to part of speech, and only contains information on whether the word is positive or negative. To integrate it into our system, we assigned all positive words an SO value of 3, and all negative words a value of -3.[33]

We used the same type of quantification for the General Inquirer (GI; Stone et al. 1966), which also has only positive and negative tags; a word was included in the dictionary if any of the senses listed in the GI were polar.

The Subjectivity dictionary is the collection of subjective expressions compiled by Wilson, Wiebe, and Hoffmann (2005), also used in our Mechanical Turk experiments in the previous section. The Subjectivity dictionary only contains a distinction between weak and strong opinion words. For our tests, weak words were assigned 2 or -2 values, depending on whether they were positive or negative, and strong words were assigned 4 or -4.

The SentiWordNet dictionary (Esuli and Sebastiani 2006; Baccianella, Esuli, and Sebastiani 2010), also used in the previous section, was built using WordNet (Fellbaum 1998), and retains its synset structure. There are two main versions of SentiWordNet available, 1.0 and 3.0, and two straightforward methods to calculate an SO value:[34] Use the first sense SO, or average the SO across senses. For the first sense method, we calculate the SO value of a word w (of a given POS) based on its first sense f as follows:

$$\mathrm{SO}(w) = 5 \times (\mathrm{Pos}(f) - \mathrm{Neg}(f))$$

For the averaging across senses method, SO is calculated as

$$\mathrm{SO}(w) = \frac{5}{|\mathrm{senses}|} \sum_{x \in \mathrm{senses}} (\mathrm{Pos}(x) - \mathrm{Neg}(x))$$

that is, the difference between the positive and negative scores provided by SentiWordNet (each in the 0–1 range), averaged across all word senses for the desired part of speech, and multiplied by 5 to provide SO values in the -5 to 5 range. Table 9 contains a comparison of performance (using simple word averaging, no SO-CAL features) in our various corpora for each version of SentiWordNet using each method. What is surprising is that the best dictionary using just basic word counts (1.0, first sense) is actually the worst dictionary when using SO-CAL, and the best dictionary using all SO-CAL features (3.0, average across senses) is the worst dictionary when features are disabled. We believe this effect is due almost entirely to the degree of positive bias in the various dictionaries. The 3.0 average dictionary is the most positively biased, which results in degraded basic performance (in the Camera corpus, only 20.5% of negative reviews are correctly identified). When negative weighting is applied, however, it reaches an almost perfect balance between positive and negative accuracy (70.7% to 71.3% in the Camera corpus), which optimizes overall performance. We cannot therefore conclude definitively that any of the SentiWordNet dictionaries is superior for our task; in fact,

33 We chose 3 as a value because it is the middle value between 2 and 4 that we assigned to strong and weak words in the Subjectivity Dictionary, as explained subsequently.

34 A third alternative would be to calculate a weighted average using sense frequency information; SentiWordNet does not include such information, however. Integrating this information from other sources, though certainly possible, would take us well beyond "off-the-shelf" usage, and, we believe, would provide only marginal benefit.

Table 9
Comparison of performance of different dictionaries derived from SentiWordNet.

SWN Dictionary			Percent correct by corpus				
Ver.	**Method**	**Test**	**Epinions 1**	**Epinions 2**	**Movie**	**Camera**	**Overall**
1.0	Average	Basic	59.25	62.50	62.89	59.92	61.18
	Average	Full	66.50	66.50	61.89	67.00	65.02
	First	Basic	60.25	62.75	62.00	60.79	**61.35**
	First	Full	65.00	64.50	62.89	66.67	64.96
3.0	Average	Basic	56.75	60.25	60.10	58.37	59.03
	Average	Full	67.50	71.50	66.21	71.00	**68.98**
	First	Basic	61.50	60.75	59.58	61.42	60.69
	First	Full	64.50	69.25	65.73	67.37	66.58

it is likely that they are roughly equivalent. We use the top-performing 3.0 average dictionary here and elsewhere.

Table 10 shows the performance of the various dictionaries when run within SO-CAL. For all dictionaries and corpora, the performance of the original SO-CAL dictionary is significantly better ($p < 0.05$). We have already discussed the Google dictionary, which contains only adjectives, and whose results are not reliable (see also Taboada, Anthony, and Voll 2006). The Maryland dictionary suffers from too much coverage: Most words in a text are identified by this dictionary as containing some form of subjectivity or opinion, but a cursory examination of the texts reveals that this is not the case. In some cases, the problem is part-of-speech assignment (the Maryland dictionary is not classified according to part of speech). For example, the noun *plot* was classified as negative when referring to a movie's plot. We imagine this is negative in the dictionary because of the negative meaning of the verb *plot*. Similarly, *novel* as a noun is classified as positive, although we believe this ought to be the case in the adjective use only. More problematic is the presence of words such as *book*, *cotton*, *here*, *legal*, *reading*, *saying*, or *year*.

Table 10
Comparison of performance using different dictionaries with SO-CAL.

Dictionary	Percent correct by corpus				
	Epinions 1	**Epinions 2**	**Movie**	**Camera**	**Overall**
Google-Full	62.00	58.50	66.31	61.25	62.98
Google-Basic	53.25	53.50	67.42	51.40	59.25
Maryland-Full-NoW	58.00	63.75	67.42	59.46	62.65
Maryland-Basic	56.50	56.00	62.26	53.79	58.16
GI-Full	68.00	70.50	64.21	72.33	68.02
GI-Basic	62.50	59.00	65.68	63.87	64.23
SentiWordNet-Full	66.50	66.50	61.89	67.00	65.02
SentiWordNet-Basic	59.25	62.50	62.89	59.92	61.47
Subjectivity-Full	72.75	71.75	65.42	77.21	72.04
Subjectivity-Basic	64.75	63.50	68.63	64.83	66.51
SO-CAL-Full	80.25	80.00	76.37	80.16	**78.74**
SO-CAL-Basic	65.50	65.25	68.05	64.70	66.04

SentiWordNet performs better than either the Google or Maryland dictionaries, but it is still somewhat low; again, we believe it suffers from the same problem of too much coverage: Potentially, every word in WordNet will receive a score, and many of those are not sentiment-bearing words. The General Inquirer lexicon (Stone et al. 1966), the only other fully manually dictionary considered here, does comparably quite well despite being relatively small. Finally, the Subjectivity dictionary, with the added strong/weak distinctions, is the closest in performance to our dictionary, though significantly worse when all features are enabled. The comparison is not completely fair to the Subjectivity dictionary, as it was built to recognize subjectivity, not polarity.

We must note that the comparison is different for the Maryland dictionary, where we turned off negative weighting. This resulted in anomalously high performance on the Movies corpus, despite poor performance elsewhere. In general, there is significantly less positive bias in the movie review domain, most likely due to the use of negative terms in plot and character description (Taboada, Brooke, and Stede 2009), thus the negative weighting that is appropriate for other domains is often excessive for movie reviews.

Comparing the performance of various dictionaries with or without SO-CAL features, two facts are apparent: First, SO-CAL features are generally beneficial no matter what dictionary is used (in fact, all Overall improvements from Basic to Full in Table 10 are statistically significant); the only exceptions are due to negative weighting in the movie domain, which for most of the dictionaries causes a drop in performance.[35] Second, the benefit provided by SO-CAL seems to be somewhat dependent on the reliability of the dictionary; in general, automatically derived SO dictionaries derive less benefit from the use of linguistic features, and the effects are, on the whole, much less consistent; this is in fact the same conclusion we reached in other work where we compared automatically translated dictionaries to manually built ones for Spanish (Brooke, Tofiloski, and Taboada 2009). Interestingly, the Subjectivity dictionary performs slightly above the SO-CAL dictionary in some data sets when no features are enabled (which we might attribute to a mixture of basic reliability with respect to polarity and an appropriate level of coverage), but its lack of granularity seems to blunt the benefit of SO-CAL features, which were designed to take advantage of a finer-grained SO scale, an effect which is even more pronounced in binary dictionaries like the GI. We can summarize this result as follows: When using lexical methods, the effectiveness of any linguistic enhancements will to some extent depend on the characteristics of the underlying lexicon and, as such, the two cannot be considered in isolation.

4. Other Related Work

The SO-CAL improvements described in this article have been directly inspired by the work of Polanyi and Zaenen (2006), who proposed that "valence shifters" change the base value of a word. We have implemented their idea in the form of intensifiers and downtoners, adding a treatment of negation that does not involve switching polarity, but instead shifting the value of a word when in the scope of a negator.

The bulk of the work in sentiment analysis has focused on classification at either the sentence level, for example, the subjectivity/polarity detection of Wiebe and Riloff (2005), or alternatively at the level of the entire text. With regard to the latter, two major

35 When negative weighting is excluded (for example, the results for the Maryland dictionary in Table 10), SO-CAL features have a positive effect on performance in the movie domain.

approaches have emerged: the use of machine-learning classifiers trained on *n*-grams or similar features (Pang, Lee, and Vaithyanathan 2002), and the use of sentiment dictionaries (Esuli and Sebastiani 2006; Taboada, Anthony, and Voll 2006). Support Vector Machine (SVM) classifiers have been shown to outperform lexicon-based models within a single domain (Kennedy and Inkpen 2006); they have trouble with cross-domain tasks (Aue and Gamon 2005), however, and some researchers have argued for hybrid classifiers (Andreevskaia and Bergler 2008). Although some of the machine-learning-based work makes use of linguistic features for training (Riloff and Wiebe 2003; Mullen and Collier 2004; Wiebe et al. 2004; Kennedy and Inkpen 2006; Ng, Dasgupta, and Niaz Arifin 2006; Sokolova and Lapalme 2008), it nonetheless still suffers from lack of cross-domain portability. The results presented here suggest that a lexicon-based system could outperform pure or hybrid machine-learning methods in cross-domain situations, though further research would be necessary to establish this point conclusively.

Ours is not the only method that uses linguistic information or dictionaries. Many other systems make use of either the Subjectivity dictionary of Wiebe and colleagues, or of SentiWordNet (Devitt and Ahmad 2007; Thet et al. 2009), and some work relies on Appraisal Theory (Whitelaw, Garg, and Argamon 2005; Bloom, Garg, and Argamon 2007), a theory developed by Martin and White (2005). We also discuss, in Section 2.4, work on incorporating linguistic insights for the treatment of negation (Moilanen and Pulman 2007; Choi and Cardie 2008).

5. Conclusions and Future Research

We have presented a word-based method for extracting sentiment from texts. Building on previous research that made use of adjectives, we extend the Semantic Orientation CALculator (SO-CAL) to other parts of speech. We also introduce intensifiers, and refine our approach to negation. The current results represent a statistically-significant improvement over previous instantiations of the SO-CAL system.

Additionally, we show that a manually built dictionary provides a solid foundation for a lexicon-based approach, one that is necessary to get full benefit from a system like SO-CAL. We compare our dictionaries to other, manual or automatic, dictionaries, and show that they are generally superior in terms of performance. This we attribute to our criteria for selecting and ranking words, which include excluding ambiguous words and including fewer rather than more words. Furthermore, we show that the dictionary rankings are in agreement with human judgments collected through the use of the Mechanical Turk interface. More importantly, we show that SO-CAL has robust performance across different types of reviews, a form of domain-independence that is difficult to achieve with text classification methods.

The existing SO-CAL can be enhanced with many other sources of information. In Taboada, Brooke, and Stede (2009) we built classifiers to distinguish among paragraphs that contained mostly description, mostly comment, a combination of the two, or meta-information (such as titles, authors, review ratings, or movie ratings). Weighting paragraphs according to this classification, with lower weights assigned to description, results in a statistically-significant improvement in the polarity classification task.

The classification of paragraphs into comment and description is but one of the many ways in which contextual information can be incorporated into a robust approach to sentiment extraction. In previous work (Voll and Taboada 2007), we showed a prototype for extracting topic sentences, and performing sentiment analysis on those only. We also showed how a sentence-level discourse parser, developed by Soricut and Marcu (2003), could be used to differentiate between main and secondary parts of the text. At

the sentence level, exploring the types of syntactic patterns that indicate subjectivity and sentiment is also a possibility (Greene and Resnik 2009). Syntactic patterns can also be used to distinguish different types of opinion and appraisal (Bednarek 2009).

Our current work focuses on developing discourse parsing methods, both general and specific to the review genre. At the same time, we will investigate different aggregation strategies for the different types of relations in the text (see also Asher, Benamara, and Mathieu [2008, 2009] for preliminary work in this area), and build on existing discourse parsing systems and proposals (Schilder 2002; Soricut and Marcu 2003; Subba and Di Eugenio 2009).

The main conclusion of our work is that lexicon-based methods for sentiment analysis are robust, result in good cross-domain performance, and can be easily enhanced with multiple sources of knowledge (Taboada, Brooke, and Stede 2009). SO-CAL has performed well on blog postings (Murray et al. 2008) and video game reviews (Brooke and Hurst 2009), without any need for further development or training.

In related work, we have also shown that creating a new version of SO-CAL for a new language, Spanish, is as fast as building text classifiers for the new language, and results in better performance (Brooke 2009; Brooke, Tofiloski, and Taboada 2009). SO-CAL has also been successfully deployed for the detection of sentence-level polarity (Brooke and Hurst 2009).

Acknowledgments

This work was supported by grants to Maite Taboada from the Natural Sciences and Engineering Research Council of Canada (Discovery Grant 261104-2008 and a University Faculty Award), and from the Social Sciences and Humanities Research Council of Canada (410-2006-1009). We thank members of the Sentiment Research Group at SFU for their feedback, and in particular Vita Markman and Ping Yang for their help in ranking our dictionaries. Thanks to Janyce Wiebe and to Rada Mihalcea and Carlo Strappavara for making their data public. Mike Thelwall and Alina Andreevskaia generously shared their data with us. The three anonymous reviewers and Robert Dale provided detailed comments and suggestions. Finally, our appreciation to colloquia audiences in Hamburg and Saarbrücken.

References

Akkaya, Cem, Alexander Conrad, Janyce Wiebe, and Rada Mihalcea. 2010. Amazon Mechanical Turk for subjectivity word sense disambiguation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 195–203, Los Angeles, CA.

Andreevskaia, Alina and Sabine Bergler. 2006. Mining WordNet for fuzzy sentiment: Sentiment tag extraction from WordNet glosses. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, EACL-2006*, pages 209–216, Trento.

Andreevskaia, Alina and Sabine Bergler. 2008. When specialists and generalists work together: Domain dependence in sentiment tagging. In *Proceedings of 46th Annual Meeting of the Association for Computational Linguistics*, pages 290–298, Columbus, OH.

Asher, Nicholas, Farah Benamara, and Yvette Yannick Mathieu. 2008. Distilling opinion in discourse: A preliminary study. In *Proceedings of COLING*, pages 7–10, Manchester.

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

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

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 Seventh Conference on International Language Resources and Evaluation (LREC'10)*, pages 2200–2204, Valletta.

Bartlett, Jake and Russ Albright. 2008. Coming to a theater near you! Sentiment classification techniques using SAS Text Miner. In *SAS Global Forum 2008*, San Antonio, TX.

Batson, C. Daniel, Laura L. Shaw, and Kathryn C. Oleson. 1992. Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions. In Margaret S. Clark, editor, *Emotion. Review of Personality and Social Psychology*. Sage, Newbury Park, CA, pages 294–326.

Bednarek, Monika. 2009. Language patterns and attitude. *Functions of Language*, 16(2):165–192.

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

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

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague.

Bloom, Kenneth, Navendu Garg, and Shlomo Argamon. 2007. Extracting appraisal expressions. In *Proceedings of HLT/NAACL*, pages 308–315, Rochester, NY.

Boiy, Erik, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. 2007. Automatic sentiment analysis of on-line text. In *Proceedings of the 11th International Conference on Electronic Publishing*, pages 349–360, Vienna.

Boucher, Jerry D. and Charles E. Osgood. 1969. The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behaviour*, 8:1–8.

Brill, Eric. 1992. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento.

Brooke, Julian. 2009. *A Semantic Approach to Automatic Text Sentiment Analysis*. M.A. thesis, Simon Fraser University, Burnaby, B.C., Canada.

Brooke, Julian and Matthew Hurst. 2009. Patterns in the stream: Exploring the interaction of polarity, topic, and discourse in a large opinion corpus. In *Proceedings of 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pages 1–8, Hong Kong.

Brooke, Julian, Milan Tofiloski, and Maite Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pages 50–54, Borovets.

Bruce, Rebecca F. and Janyce M. Wiebe. 2000. Recognizing subjectivity: A case study of manual tagging. *Natural Language Engineering*, 5(2):187–205.

Cabral, Luís and Ali Hortaçsu. 2010. The dynamics of seller reputation: Evidence from eBay. *The Journal of Industrial Economics*, 58(1):54–78.

Callison-Burch, Chris. 2009. Fast, cheap and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore.

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

Chaovalit, Pinvadee and Lina Zhou. 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. In *Proceedings of the 38th Hawaii International Conference on System Sciences*, Hawaii.

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

Conrad, Susan and Douglas Biber. 2000. Adverbial marking of stance in speech and writing. In Geoff Thompson, editor, *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.

Dang, Yan, Yulei Zhang, and Hsinchun Chen. 2010. A lexicon enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53.

Dasgupta, Sajib and Vincent Ng. 2009. Mine the easy, classify the hard: A semi-supervised approach to automatic sentiment classification. In *Proceedings of the 47th Annual Meeting of the Association for*

Computational Linguistics, pages 701–709, Singapore.
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 Twelfth International World Wide Web Conference (WWW 2003)*, pages 519–528, Budapest.
Devitt, Ann and Khurshid Ahmad. 2007. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 984–991, Prague.
Esuli, Andrea and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 417–422, Genoa.
Fellbaum, Christiane, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.
Finn, Aidan and Nicholas Kushmerick. 2003. Learning to classify documents according to genre. In *Proceedings of IJCAI Workshop on Computational Approaches to Text Style and Synthesis*, Acapulco.
Fletcher, Jeremy and Jon Patrick. 2005. Evaluating the utility of appraisal hierarchies as a method for sentiment classification. In *Proceedings of the Australasian Language Technology Workshop*, pages 134–142, Sydney.
Giannakidou, Anastasia. 1998. *Polarity Sensitivity as (Non)Veridical Dependency*. John Benjamins, Amsterdam and Philadelphia.
Giannakidou, Anastasia. 2001. Varieties of polarity items and the (non)veridicality hypothesis. In Ton van der Wouden, editor, *Perspectives on Negation and Polarity Items*. John Benjamins, Amsterdam and Philadelphia, pages 99–127.
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 HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing*, pages 45–52, New York.
Greenberg, Joseph H. 1966. *Language Universals, with Special Reference to Feature Hierarchies*. Mouton, The Hague.
Greene, Stephan and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 503–511, Boulder, CO.
Halliday, Michael A. K. 1985. *An Introduction to Functional Grammar*. Arnold, London, 1st edition.
Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of 35th Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid.
Horn, Laurence R. 1989. *A Natural History of Negation*. University of Chicago Press, Chicago, IL.
Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, pages 168–177, Seattle, WA.
Jing-Schmidt, Zhuo. 2007. Negativity bias in language: A cognitive-affective model of emotive intensifiers. *Cognitive Linguistics*, 18(3):417–443.
Kennedy, Alistair and Diana Inkpen. 2006. Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125.
Ketal, R. 1975. Affect, mood, emotion, and feeling: Semantic considerations. *American Journal of Psychiatry*, 132:1215–1217.
Kilgarriff, Adam. 2007. Googleology is bad science. *Computational Linguistics*, 33(1):147–151.
Kim, Soo-Min and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of COLING 2004*, pages 1367–1373, Geneva.
Koppel, Moshe and Jonathan Schler. 2005. Using neutral examples for learning polarity. In *Proceedings of IJCAI 2005*, pages 1616–1617, Edinburgh.
Langacker, Ronald W. 1985. Observations and speculations on subjectivity. In John Haiman, editor, *Iconicity in Syntax*. John Benjamins, Amsterdam and Philadelphia, pages 109–150.
Li, Shoushan, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 414–423, Uppsala.
Liu, Jingjing and Stephanie Seneff. 2009. Review sentiment scoring via a

parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 161–169, Singapore.
Lyons, John. 1981. *Language, Meaning and Context*. Fontana, London.
Mann, William C. and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.
Martin, James R. and Peter R. R. White. 2005. *The Language of Evaluation*. Palgrave, New York.
Mellebeek, Bart, Francesc Benavent, Jens Grivolla, Joan Codina, Marta R. Costa-jussà, and Rafael Banchs. 2010. Opinion mining of Spanish customer comments with non-expert annotations on Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 114–121, Los Angeles, CA.
Mohammad, Saif, Bonnie Dorr, and Cody Dunne. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pages 599–608, Singapore.
Mohammad, Saif and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA.
Moilanen, Karo and Stephen Pulman. 2007. Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing*, pages 27–29, Borovets.
Moilanen, Karo, Stephen Pulman, and Yue Zhang. 2010. Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010)*, pages 36–43, Lisbon.
Mullen, Tony and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 412–418, Barcelona.
Murray, Gabriel, Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2008. The University of British Columbia at TAC 2008. In *Proceedings of TAC 2008*, Gaithersburg, MD.
Ng, Vincent, Sajib Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the identification and classification of reviews. In *Proceedings of the COLING/ ACL 2006 Main Conference Poster Sessions*, pages 611–618, Sydney.
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., George Suci, and Percy Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press, Urbana, IL.
Paltoglou, Georgios and Mike Thelwall. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1386–1395, Uppsala.
Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd 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 ACL 2005*, pages 115–124, Ann Arbor, MI.
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 NLP*, pages 79–86, Philadelphia, PA.
Polanyi, Livia and Annie Zaenen. 2006. Contextual valence shifters. In Janyce Wiebe, editor, *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Dordrecht, pages 1–10.
Prabowo, Rudy and Mike Thelwall. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(1):143–157.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985.

A Comprehensive Grammar of the English Language. Longman, London.
Rao, Delip and Deepak Ravichandran. 2009. Semi-supervised polarity lexicon induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 675–682, Athens.
Read, Jonathon and John Carroll. 2009. Weakly supervised techniques for domain-independent sentiment classification. In *Proceedings of the First International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pages 45–52, Hong Kong.
Riloff, Ellen and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 25–32, Sapporo.
Salton, Gerard and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.
Salvetti, Franco, Christoph Reichenbach, and Stephen Lewis. 2006. Opinion polarity identification of movie reviews. In Janyce Wiebe, editor, *Computing Attitude and Affect in Text: Theory and Applications*. Springer, Dordrecht, pages 303–316.
Saurí, Roser. 2008. *A Factuality Profiler for Eventualities in Text*. Ph.D. dissertation, Brandeis University, Waltham, MA.
Scheibman, Joanne. 2002. *Point of View and Grammar: Structural Patterns of Subjectivity in American English*. John Benjamins, Amsterdam and Philadelphia.
Schilder, Frank. 2002. Robust discourse parsing via discourse markers, topicality, and position. *Natural Language Engineering*, 8(2/3):235–255.
Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in NLP (EMNLP)*, pages 254–263, Waikiki, HI.
Sokolova, Marina and Guy Lapalme. 2008. Verbs speak loud: Verb categories in learning polarity and strength of opinions. In Sabine Bergler, editor, *Canadian AI 2008*. Springer, Berlin, pages 320–331.
Sokolova, Marina and Guy Lapalme. 2009a. Classification of opinions with non-affective adverbs and adjectives. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pages 416–420, Borovets.
Sokolova, Marina and Guy Lapalme. 2009b. Opinion learning without emotional words. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence*, pages 253–256, Kelowna.
Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT-NAACL'03)*, pages 149–156, Edmonton.
Stone, Philip J. 1997. Thematic text analysis: New agendas for analyzing text content. In Carl Roberts, editor, *Text Analysis for the Social Sciences*. Lawrence Erlbaum, Mahwah, NJ, pages 35–54.
Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.
Strappavara, Carlo and Rada Mihalcea. 2007. SemEval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on the Semantic Evaluations (SemEval 2007)*, Prague.
Subba, Rajen and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of HLT-ACL 2009*, pages 566–574, Boulder, CO.
Subrahmanian, V. S. and Diego Reforgiato. 2008. Ava: Adjective-verb-adverb combinations for sentiment analysis. *Intelligent Systems*, 23(4):43–50.
Taboada, Maite, Caroline Anthony, and Kimberly Voll. 2006. Creating semantic orientation dictionaries. In *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*, pages 427–432, Genoa.
Taboada, Maite, Julian Brooke, and Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. In *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 62–70, London.
Taboada, Maite, Mary Ann Gillies, and Paul McFetridge. 2006. Sentiment classification techniques for tracking literary reputation. In *Proceedings of the LREC Workshop, "Towards Computational Models of Literary Analysis,"* pages 36–43, Genoa.
Taboada, Maite, Mary Ann Gillies, Paul McFetridge, and Robert Outtrim. 2008. Tracking literary reputation with text analysis tools. In *Meeting of the Society for Digital Humanities*, Vancouver.

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

Thelwall, Mike, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558.

Thet, Tun Thura, Jin-Cheon Na, Christopher S. G. Khoo, and Subbaraj Shakthikumar. 2009. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *Proceedings of the First International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pages 81–84, Hong Kong.

Tong, Richard M. 2001. An operational system for detecting and tracking opinions in on-line discussions. In *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*, pages 1–6, New York, NY.

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

Turney, Peter. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, Peter and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.

Velikovich, Leonid, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald. 2010. The viability of Web-derived polarity lexicons. In *Proceedings of the 11th Conference of the North American Association for Computational Linguistics*, pages 777–785, Los Angeles, CA.

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 Artificial Intelligence*, pages 337–346, Gold Coast.

Wan, Xiaojun. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the 47th Annual Meeting of the ACL*, pages 235–243, Singapore.

Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005)*, pages 625–631, Bremen.

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

Wiebe, Janyce. 2000. Learning subjective adjectives from corpora. In *Proceedings of 17th National Conference on Artificial Intelligence (AAAI)*, pages 735–740, Austin, TX.

Wiebe, Janyce and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, Mexico City.

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.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the 2005 Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-05)*, pages 347–354, Vancouver.

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *AAAI 2004*, pages 761–767, San Jose, CA.

Yano, Tae, Philip Resnik, and Noah A. Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158, Los Angeles, CA.

Zaenen, Annie. to appear. Do give a penny for their thoughts. *Natural Language Engineering*.

Zaidan, Omar F. and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, HI.

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