# Determining the Sentiment of Opinions

**Soo-Min Kim**
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
skim@isi.edu

**Eduard Hovy**
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Rey, CA 90292-6695
hovy@isi.edu

**Abstract**

Identifying sentiments (the affective parts of opinions) is a challenging problem. We present a system that, given a topic, automatically finds the people who hold opinions about that topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence. We experiment with various models of classifying and combining sentiment at word and sentence levels, with promising results.

## 1 Introduction

What is an opinion?

The many opinions on opinions are reflected in a considerable literature (Aristotle 1954; Perelman 1970; Toulmin et al. 1979; Wallace 1975; Toulmin 2003). Recent computational work either focuses on sentence 'subjectivity' (Wiebe et al. 2002; Riloff et al. 2003), concentrates just on explicit statements of evaluation, such as of films (Turney 2002; Pang et al. 2002), or focuses on just one aspect of opinion, e.g., (Hatzivassiloglou and McKeown 1997) on adjectives. We wish to study opinion in general; our work most closely resembles that of (Yu and Hatzivassiloglou 2003).

Since an analytic definition of opinion is probably impossible anyway, we will not summarize past discussion or try to define formally what is and what is not an opinion. For our purposes, we describe an opinion as a quadruple [Topic, Holder, Claim, Sentiment] in which the Holder believes a Claim about the Topic, and in many cases associates a Sentiment, such as *good* or *bad*, with the belief. For example, the following opinions contain Claims but no Sentiments:

"I believe the world is flat"

"The Gap is likely to go bankrupt"

"Bin Laden is hiding in Pakistan"

"Water always flushes anti-clockwise in the southern hemisphere"

Like Yu and Hatzivassiloglou (2003), we want to automatically identify Sentiments, which in this work we define as an explicit or implicit expression in text of the Holder's positive, negative, or neutral regard toward the Claim about the Topic. (Other sentiments we plan to study later.) Sentiments always involve the Holder's emotions or desires, and may be present explicitly or only implicitly:

"I think that attacking Iraq would put the US in a difficult position" (implicit)

"The US attack on Iraq is wrong" (explicit)

"I like Ike" (explicit)

"We should decrease our dependence on oil" (implicit)

"Reps. Tom Petri and William F. Goodling asserted that counting illegal aliens violates citizens' basic right to equal representation" (implicit)

In this paper we address the following challenge problem. Given a Topic (e.g., "Should abortion be banned?") and a set of texts about the topic, find the Sentiments expressed about (claims about) the Topic (but not its supporting subtopics) in each text, and identify the people who hold each sentiment. To avoid the problem of differentiating between shades of sentiments, we simplify the problem to: identify just expressions of positive, negative, or neutral sentiments, together with their holders. In addition, for sentences that do not express a sentiment but simply state that some sentiment(s) exist(s), return these sentences in a separate set. For example, given the topic "What should be done with Medicare?" the sentence "After years of empty promises, Congress has rolled out two Medicare prescription plans, one from House Republicans and the other from the Democratic

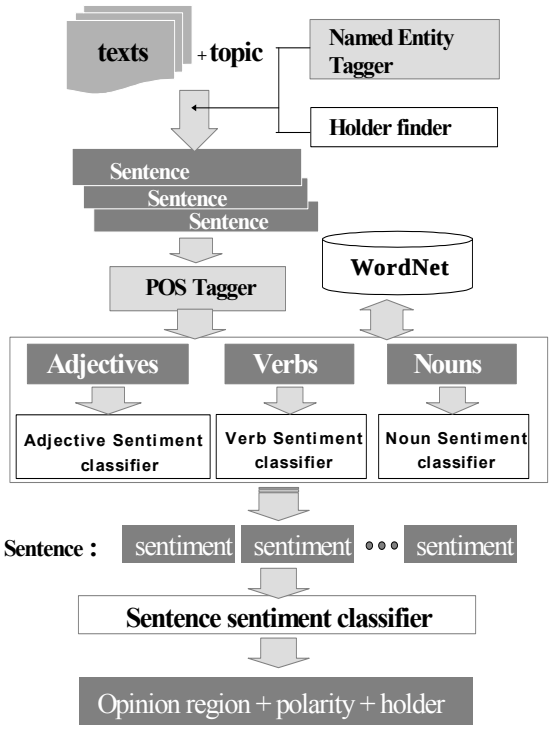


Figure 1: System architecture.

Sens. Bob Graham of Florida and Zell Miller of Georgia" should be returned in the separate set.

We approach the problem in stages, starting with words and moving on to sentences. We take as unit sentiment carrier a single word, and first classify each adjective, verb, and noun by its sentiment. We experimented with several classifier models. But combining sentiments requires additional care, as Table 1 shows.

| |
|---|
| California Supreme Court ***agreed*** that the state's new term-limit law was ***constitutional***. |
| California Supreme Court ***disagreed*** that the state's new term-limit law was ***constitutional.*** |
| California Supreme Court ***agreed*** that the state's new term-limit law was ***unconstitutional.*** |
| California Supreme Court ***disagreed*** that the state's new term-limit law was ***unconstitutional.*** |

Table 1: Combining sentiments.

A sentence might even express opinions of different people. When combining word-level sentiments, we therefore first determine for each Holder a relevant region within the sentence and then experiment with various models for combining word sentiments.

We describe our models and algorithm in Section 2, system experiments and discussion in Section 3, and conclude in Section 4.

## 2 Algorithm

Given a topic and a set of texts, the system operates in four steps. First it selects sentences that contain both the topic phrase and holder candidates. Next, the holder-based regions of opinion are delimited. Then the sentence sentiment classifier calculates the polarity of all sentiment-bearing words individually. Finally, the system combines them to produce the holder's sentiment for the whole sentence. Figure 1 shows the overall system architecture. Section 2.1 describes the word sentiment classifier and Section 2.2 describes the sentence sentiment classifier.

### 2.1 Word Sentiment Classifier

#### 2.1.1 Word Classification Models

For word sentiment classification we developed two models. The basic approach is to assemble a small amount of seed words by hand, sorted by polarity into two lists—positive and negative—and then to grow this by adding words obtained from WordNet (Miller et al. 1993; Fellbaum et al. 1993). We assume synonyms of positive words are mostly positive and antonyms mostly negative, e.g., the positive word "good" has synonyms "virtuous, honorable, righteous" and antonyms "evil, disreputable, unrighteous". Antonyms of negative words are added to the positive list, and synonyms to the negative one.

To start the seed lists we selected verbs (23 positive and 21 negative) and adjectives (15 positive and 19 negative), adding nouns later.

Since adjectives and verbs are structured differently in WordNet, we obtained from it synonyms and antonyms for adjectives but only synonyms for verbs. For each seed word, we extracted from WordNet its expansions and added them back into the appropriate seed lists. Using these expanded lists, we extracted an additional cycle of words from WordNet, to obtain finally 5880 positive adjectives, 6233 negative adjectives, 2840 positive verbs, and 3239 negative verbs.

However, not all synonyms and antonyms could be used: some had opposite sentiment or were neutral. In addition, some common words such as "great", "strong", "take", and "get" occurred many times in both positive and negative categories. This indicated the need to develop a measure of strength of sentiment

polarity (the alternative was simply to discard such ambiguous words)—to determine how strongly a word is positive *and also* how strongly it is negative. This would enable us to discard sentiment-ambiguous words but retain those with strengths over some threshold.

Armed with such a measure, we can also assign strength of sentiment polarity to as yet unseen words. Given a new word, we use WordNet again to obtain a synonym set of the unseen word to determine how it interacts with our sentiment seed lists. That is, we compute

$$\begin{aligned} &\arg\max_c P(c \mid w) \\ &\cong \arg\max_c P(c \mid syn_1, syn_2 .... syn_n) \end{aligned} \quad (1)$$

where $c$ is a sentiment category (positive or negative), $w$ is the unseen word, and $syn_n$ are the WordNet synonyms of $w$. To compute Equation (1), we tried two different models:

$$\begin{aligned} &\arg\max_c P(c \mid w) = \arg\max_c P(c)P(w \mid c) \\ &= \arg\max_c P(c)P(syn_1 syn_2 syn_3 ... syn_n \mid c) \\ &= \arg\max_c P(c)\prod_{k=1}^{m} P(f_k \mid c)^{count(f_k, synset(w))} \end{aligned} \quad (2)$$

where $f_k$ is the $k^{th}$ feature (list word) of sentiment class $c$ which is also a member of the synonym set of $w$, and $count(f_k, synset(w))$ is the total number of occurrences of $f_k$ in the synonym set of $w$. $P(c)$ is the number of words in class $c$ divided by the total number of words considered. This model derives from document classification. We used the synonym and antonym lists obtained from Wordnet instead of learning word sets from a corpus, since the former is simpler and does not require manually annotated data for training.

Equation (3) shows the second model for a word sentiment classifier.

$$\begin{aligned} &\arg\max_c P(c \mid w) = \arg\max_c P(c)P(w \mid c) \\ &= \arg\max_c P(c)\frac{\sum_{i=1}^{n} count(syn_i, c)}{count(c)} \end{aligned} \quad (3)$$

To compute the probability $P(w|c)$ of word $w$ given a sentiment class $c$, we count the occurrence of $w$'s synonyms in the list of $c$. The intuition is that the more synonyms occuring in $c$, the more likely the word belongs.

We computed both positive and negative sentiment strengths for each word and compared their relative magnitudes. Table 2 shows several examples of the system output, computed with Equation (2), in which "+" represents positive category strength and "-" negative. The word "amusing", for example, was classified as carrying primarily positive sentiment, and "blame" as primarily negative. The absolute value of each category represents the strength of its sentiment polarity. For instance, "afraid" with strength -0.99 represents strong negavitity while "abysmal" with strength -0.61 represents weaker negativity.

abysmal : NEGATIVE
[+ : 0.3811][- : 0.6188]

adequate : POSITIVE
[+ : 0.9999][- : 0.0484e-11]

afraid : NEGATIVE
[+ : 0.0212e-04][- : 0.9999]

ailing : NEGATIVE
[+ : 0.0467e-8][- : 0.9999]

amusing : POSITIVE
[+ : 0.9999][- : 0.0593e-07]

answerable : POSITIVE
[+ : 0.8655][- : 0.1344]

apprehensible: POSITIVE
[+ : 0.9999][- : 0.0227e-07]

averse : NEGATIVE
[+ : 0.0454e-05][- : 0.9999]

blame : NEGATIVE
[+ : 0.2530][- : 0.7469]

Table 2: Sample output of word sentiment classifier.

## 2.2 Sentence Sentiment Classifier

As shows in Table 1, combining sentiments in a sentence can be tricky. We are interested in the sentiments of the Holder about the Claim. Manual analysis showed that such sentiments can be found most reliably close to the Holder; without either Holder or Topic/Claim nearby as anchor points, even humans sometimes have trouble reliably determining the source of a sentiment. We therefore included in the algorithm steps to identify the Topic (through direct matching, since we took it as given) and any likely opinion Holders (see Section 2.2.1). Near each Holder we then identified a region in which sentiments would be considered; any sentiments outside such a region we take to be of undetermined origin and ignore (Section 2.2.2). We then defined several models for combining the sentiments expressed within a region (Section 2.2.3).

### 2.2.1 Holder Identification

We used BBN's named entity tagger IdentiFinder to identify potential holders of an opinion. We considered PERSON and ORGANIZATION as the only possible opinion holders. For sentences with more than one Holder, we chose the one closest to the Topic phrase, for simplicity. This is a very crude step. A more sophisticated approach would employ a parser to identify syntactic relationships between each Holder and all dependent expressions of sentiment.

### 2.2.2 Sentiment Region

Lacking a parse of the sentence, we were faced with a dilemma: How large should a region be? We therefore defined the sentiment region in various ways (see Table 3) and experimented with their effectiveness, as reported in Section 3.

| |
|---|
| **Window1**: full sentence |
| **Window2**: words between Holder and Topic |
| **Window3**: *window2* $\pm$ 2 words |
| **Window4**: *window2* to the end of sentence |

Table 3: Four variations of region size.

### 2.2.3 Classification Models

We built three models to assign a sentiment category to a given sentence, each combining the individual sentiments of sentiment-bearing words, as described above, in a different way.

Model 0 simply considers the polarities of the sentiments, not the strengths:

**Model 0**: $\prod$ (signs in region)

The intuition here is something like "negatives cancel one another out". Here the system assigns the same sentiment to both "the California Supreme Court ***agreed*** that the state's new term-limit law was ***constitutional***" and "the California Supreme Court ***disagreed*** that the state's new term-limit law was ***unconstitutional***". For this model, we also included negation words such as *not* and *never* to reverse the sentiment polarity.

Model 1 is the harmonic mean (average) of the sentiment strengths in the region:

**Model 1**:
$$P(c \mid s) = \frac{1}{n(c)} \sum_{i=1}^{n} p(c \mid w_i),$$
$$\text{if } \operatorname*{argmax}_{j} p(c_j \mid w_i) = c$$

Here $n(c)$ is the number of words in the region whose sentiment category is $c$. If a region contains more and stronger positive than negative words, the sentiment will be positive.

Model 2 is the geometric mean:

**Model 2**:
$$P(c \mid s) = 10^{n(c)-1} \times \prod_{i=1}^{n} p(c \mid w_i),$$
$$\textit{if } \operatorname*{argmax}_{j} p(c_j \mid w_i) = c$$

### 2.2.4 Examples

The following are two example outputs.

| |
|---|
| Public officials throughout California have condemned a ***U.S. Senate*** vote Thursday to exclude ***illegal aliens*** from the 1990 census, saying the action will shortchange California in Congress and possibly deprive the state of millions of dollars of federal aid for medical emergency services and other programs for poor people. |
| TOPIC : illegal alien |
| HOLDER : U.S. Senate |
| OPINION REGION: vote/NN Thursday/NNP to/TO exclude/VB illegal/JJ aliens/NNS from/IN the/DT 1990/CD census,/NN |
| SENTIMENT_POLARITY: negative |
| For that reason and others, the Constitutional Convention unanimously rejected term limits and the ***First Congress*** soundly defeated two subsequent ***term-limit*** proposals. |
| TOPIC : term limit |
| HOLDER : First Congress |
| OPINION REGION: soundly/RB defeated/VBD two/CD subsequent/JJ term-limit/JJ proposals./NN |
| SENTIMENT_POLARITY: negative |

## 3 Experiments

The first experiment examines the two word sentiment classifier models and the second the three sentence sentiment classifier models.

### 3.1 Word Sentiment Classifier

For test material, we asked three humans to classify data. We started with a basic English word list for foreign students preparing for the TOEFL test and intersected it with an adjective list containing 19748 English adjectives and a verb list of 8011 verbs to obtain common

adjectives and verbs. From this we randomly selected 462 adjectives and 502 verbs for human classification. Human1 and human2 each classified 462 adjectives, and human2 and human3 502 verbs.

The classification task is defined as assigning each word to one of three categories: positive, negative, and neutral.

#### 3.1.1 Human—Human Agreement

| | Adjectives | Verbs |
|---|---|---|
| | Human1 : Human2 | Human1 : Human3 |
| Strict | 76.19% | 62.35% |
| Lenient | 88.96% | 85.06% |

Table 4: Inter-human classification agreement.

Table 4 shows inter-human agreement. The strict measure is defined over all three categories, whereas the lenient measure is taken over only two categories, where positive and neutral have been merged, should we choose to focus only on differentiating words of negative sentiment.

#### 3.1.2 Human—Machine Agreement

Table 5 shows results, using Equation (2) of Section 2.1.1, compared against a baseline that randomly assigns a sentiment category to each word (averaged over 10 iterations). The system achieves lower agreement than humans but higher than the random process.

Of the test data, the algorithm classified 93.07% of adjectives and 83.27% of verbs as either positive and negative. The remainder of adjectives and verbs failed to be classified, since they did not overlap with the synonym set of adjectives and verbs.

In Table 5, the seed list included just a few manually selected seed words (23 positive and 21 negative verbs and 15 and 19 adjectives, repectively). We decided to investigate the effect of more seed words. After collecting the annotated data, we added half of it (231 adjectives and 251 verbs) to the training set, retaining the other half for the test. As Table 6 shows, agreement of both adjectives and verbs with humans improves. Recall is also improved.

| Adjective (Train: 231 Test : 231) | | | Verb (Train: 251 Test : 251) | | |
|---|---|---|---|---|---|
| Lenient agreement | | recall | Lenient agreement | | recall |
| H1:M | H2:M | | H1:M | H3:M | |
| 75.66% | 77.88% | 97.84% | 81.20% | 79.06% | 93.23% |

Table 6: Results including manual data.

### 3.2 Sentence Sentiment Classifier

#### 3.2.1 Data

100 sentences were selected from the DUC 2001 corpus with the topics “illegal alien”, “term limits”, “gun control”, and “NAFTA”. Two humans annotated the 100 sentences with three categories (positive, negative, and N/A). To measure the agreement between humans, we used the Kappa statistic (Siegel and Castellan Jr. 1988). The Kappa value for the annotation task of 100 sentences was 0.91, which is considered to be reliable.

#### 3.2.2 Test on Human Annotated Data

We experimented on Section 2.2.3’s 3 models of sentiment classifiers, using the 4 different window definitions and 4 variations of word-level classifiers (the two word sentiment equations introduced in Section 2.1.1, first with and then without normalization, to compare performance).

Since Model 0 considers not probabilities of words but only their polarities, the two word-level classifier equations yield the same results. Consequently, Model 0 has 8 combinations and Models 1 and 2 have 16 each.

To test the identification of opinion Holder, we first ran models with holders that were annotated by humans then ran the same models with the automatic holder finding strategies. The results appear in Figures 2 and 3. The models are numbered as follows: m0 through m4 represent 4 sentence classifier models,

| | Adjective (test: 231 adjectives) | | | Verb (test : 251 verbs) | | |
|---|---|---|---|---|---|---|
| | Lenient agreement | | recall | Lenient agreement | | recall |
| | H1:M | H2:M | | H1:M | H3:M | |
| Random selection (average of 10 iterations) | 59.35% | 57.81% | 100% | 59.02% | 56.59% | 100% |
| Basic method | 68.37% | 68.60% | 93.07% | 75.84% | 72.72% | 83.27% |

Table 5. Agreement between humans and system.

p1/p2 and p3/p4 represent the word classifier models in Equation (2) and Equation (3) with normalization and without normalization respectively.

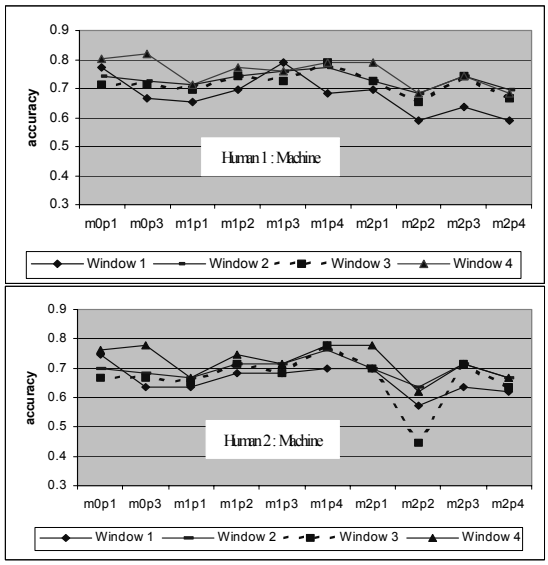


Figure 2: Results with manually annotated Holder.

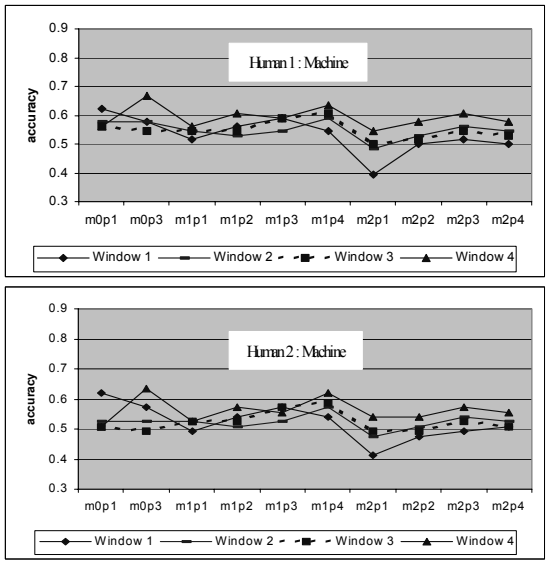


Figure 3: Results with automatic Holder detection.

Correctness of an opinion is determined when the system finds both a correct holder and the appropriate sentiment within the sentence. Since human1 classified 33 sentences positive and 33 negative, random classification gives 33 out of 66 sentences. Similarly, since human2 classified 29 positive and 34 negative, random classification gives 34 out of 63 when the system blindly marks all sentences as negative and 29 out of 63 when it marks all as positive. The system's best model performed at 81% accuracy with the manually provided holder and at 67% accuracy with automatic holder detection.

### 3.3 Problems

#### 3.3.1 Word Sentiment Classification

As mentioned, some words have both strong positive and negative sentiment. For these words, it is difficult to pick one sentiment category without considering context. Second, a unigram model is not sufficient: common words without much sentiment alone can combine to produce reliable sentiment. For example, in "'Term limits really hit at democracy,' says Prof. Fenno", the common and multi-meaning word "hit" was used to express a negative point of view about term limits. If such combinations occur adjacently, we can use bigrams or trigrams in the seed word list. When they occur at a distance, however, it is more difficult to identify the sentiment correctly, especially if one of the words falls outside the sentiment region.

#### 3.3.2 Sentence Sentiment Classification

Even in a single sentence, a holder might express two different opinions. Our system only detects the closest one.

Another difficult problem is that the models cannot infer sentiments from facts in a sentence. "She thinks term limits will give women more opportunities in politics" expresses a positive opinion about term limits but the absence of adjective, verb, and noun sentiment-words prevents a classification.

Although relatively easy task for people, detecting an opinion holder is not simple either. As a result, our system sometimes picks a wrong holder when there are multiple plausible opinion holder candidates present. Employing a parser to delimit opinion regions and more accurately associate them with potential holders should help.

### 3.4 Discussion

**Which combination of models is best?**

The best overall performance is provided by Model 0. Apparently, the mere presence of negative words is more important than sentiment strength. For manually tagged holder and topic, Model 0 has the highest single performance, though Model 1 averages best.

**Which is better, a sentence or a region?**

With manually identified topic and holder, the region window4 (from Holder to sentence end) performs better than other regions.

**How do scores differ from manual to automatic holder identification?**

Table 7 compares the average results with automatic holder identification to manually annotated holders in 40 different models. Around 7 more sentences (around 11%) were misclassified by the automatic detection method.

| | positive | negative | total |
|---|---|---|---|
| Human1 | 5.394 | 1.667 | 7.060 |
| Human2 | 4.984 | 1.714 | 6.698 |

Table 7: Average difference between manual and automatic holder detection.

**How does adding the neutral sentiment as a separate category affect the score?**

It is very confusing even for humans to distinguish between a neutral opinion and non-opinion bearing sentences. In previous research, we built a sentence subjectivity classifier. Unfortunately, in most cases it classifies neutral and weak sentiment sentences as non-opinion bearing sentences.

## 4 Conclusion

Sentiment recognition is a challenging and difficult part of understanding opinions. We plan to extend our work to more difficult cases such as sentences with weak-opinion-bearing words or sentences with multiple opinions about a topic. To improve identification of the Holder, we plan to use a parser to associate regions more reliably with holders. We plan to explore other learning techniques, such as decision lists or SVMs.

Nonetheless, as the experiments show, encouraging results can be obtained even with relatively simple models and only a small amount of manual seeding effort.

## References


Aristotle. *The Rhetorics and Poetics* (trans. W. Rhys Roberts), Modern Library, 1954.

Fellbaum, C., D. Gross, and K. Miller. 1993. Adjectives in WordNet. http://www.cosgi. princeton.edu/~wn.

Hatzivassiloglou, V. and K. McKeown 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of the 35th ACL conference,* 174–181.

Miller, G.A., R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An On-Line Lexical Database. http://www.cosgi.princeton.edu/~wn.

Pang, B. L. Lee, and S. Vaithyanathan, 2002. Thumbs up? Sentiment classification using Machine Learning Techniques. *Proceedings of the EMNLP conference.*

Perelman, C. 1970. The New Rhetoric: A Theory of Practical Reasoning. In *The Great Ideas Today*. Chicago: Encyclopedia Britannica.

Riloff, E., J. Wiebe, and T. Wilson 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *Proceedings of the CoNLL-03 conference.*

Siegel, S. and N.J. Castellan Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Toulmin, S.E., R. Rieke, and A. Janik. 1979. *An Introduction to Reasoning*. Macmillan, New York.

Toulmin, S.E. 2003. *The Uses of Argument*. Cambridge University Press.

Turney, P. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, 417–424.

Wallace, K. 1975. *Topoi* and the Problem of Invention. In W. Ross Winterowd (ed), *Contemporary Rhetoric*. Harcourt Brace Jovanovich.

Wiebe, J. et al. 2002. NRRC summer study Jan Wiebe and group (University of Pittsburgh) on 'subjective' statements.

Yu, H. and V. Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *Proceedings of the EMNLP conference*.