Methods for Creating Semantic Orientation Dictionaries

Maite Taboada, Caroline Anthony and Kimberly Voll

Simon Fraser University

8888 University Dr., Burnaby, BC, V5A 1S6, Canada

E-mail: mtaboada@sfu.ca, canthony@sfu.ca, kvoll@sfu.ca

Abstract

We describe and compare different methods for creating a dictionary of words with their corresponding semantic orientation (SO). We tested how well different dictionaries helped determine the SO of entire texts. To extract SO for each individual word, we used a common method based on pointwise mutual information. Mutual information between a set of seed words and the target words was calculated using two different methods: a NEAR search on the search engine Altavista (since discontinued); an AND search on Google. These two dictionaries were tested against a manually annotated dictionary of positive and negative words. The results show that all three methods are quite close, and none of them performs particularly well. We discuss possible further avenues for research, and also point out some potential problems in calculating pointwise mutual information using Google.

1. Introduction

The problem of extracting the semantic orientation (SO) of a text (i.e., whether the text is positive or negative towards a particular subject matter) often takes as a starting point the problem of determining semantic orientation for individual words. The hypothesis is that, given the SO of relevant words in a text, we can determine the SO for the entire text. We will see later that this is not the whole or the only story. However, if we assume that SO for individual words is an important part of the problem, we need to consider what are the best methods to extract SO.

Turney (2002) proposed a method for automatically extracting SO using the NEAR operator available from Altavista. NEAR allowed a targeted search, finding two words in the vicinity of each other. The results of a NEAR-based search were then used to calculate SO. A word that is close to one or more seed words (positive or negative words) will likely share SO with those word(s). Since Altavista discontinued the NEAR operator, other options need to be researched. We have considered two alternatives: Google with the AND operator (which searches for two words anywhere in the same document), and a list of positive and negative words from the General Inquirer (Stone, 1997; Stone et al., 1966).

Turney & Littman (2002) performed an evaluation on the accuracy of NEAR versus AND in Altavista. Their evaluation was based on the accuracy of the SO for individual words, compared to a benchmark of words that had been manually annotated for positive and negative values (from the General Inquirer). Turney (2001) shows that NEAR provides better results than AND in a test of synonyms. Our evaluation of different methods examines not only the accuracy of individual words, but also their contribution to extracting SO for entire texts.

We extract the SO of 400 reviews from the website Epinions.com (about movies, music, books, hotels, cars, phones, computers, and cookware), using a weighted average of the adjectives in the texts. Our adjective dictionary contains 1,719 adjectives, whose SO was calculated using different methods: Altavista's NEAR (when it was available); Google's AND; and extracting a subset of the positive/negative values from the General Inquirer (a total of 521 adjectives). The results show that the difference between the original NEAR and the new AND dictionaries is not significant. However, further tests with extracting SO values using AND suggest that it is not a robust method, and that perhaps a static corpus might be better than Google, which indexes a dynamic corpus.

2. Background

The final goal of our project is to automatically extract the opinion expressed in a text. One of the methods proposed for such a task consists of extracting relevant words in the text, determining whether those words carry negative or positive meaning, and expressing such meaning in a numeric value. An aggregation of the positive/negative values of the words in the text produces the semantic orientation for the entire text. This approach entails determining which words or phrases are relevant (i.e., which words capture the SO for a sentence or text); which sentences are relevant (i.e., are some sentences or parts of a text more representative of its SO?); and how to aggregate the individual words or phrases extracted.

2.1 Which Words and Phrases

Most research in this area has focused on adjectives. Adjectives convey much of the subjective content in a text, and a great deal of effort has been devoted to extracting SO for adjectives. Hatzivassiloglou & McKeown (1997) pioneered the extraction of SO by association, using coordination: the phrase *excellent and X* predicts that *X* will be a positive adjective. Turney (2002), and Turney & Littman (2002; 2003) used a similar method, but this time using the Web as corpus. In their method, the adjective *X* is positive if it appears mostly in the vicinity of other positive adjectives, not only in a coordinated phrase. "Vicinity" was defined using the NEAR operator in the Altavista search engine, which by default looked for words within ten words of each other. The contribution of Turney & Littman was to find a way to not only extract the sign (positive or negative) for any given adjective, but also to extract the strength of the SO. They use Pointwise Mutual Information (PMI) for that purpose. PMI calculations do not have to be limited to adjectives. In fact, Turney (2002) used two-word combinations that included, mostly, Adjective+Noun, Adverb+Noun, and Adverb+Verb.

A different strategy to find opinion words consists of finding synonyms and similar words in general. The synonyms are extracted using either PMI (Turney, 2001) or Latent Semantic Analysis (Landauer & Dumais, 1997). It is unclear which method provides the best results; published accounts vary (Rapp, 2004; Turney, 2001). Word similarity may be another way of building dictionaries, starting from words whose SO we already know. For this purpose, WordNet is a valuable resource, since synonymy relations are already defined (Kamps et al., 2004). Esuli and Sebastiani (2005) also use synonyms, but they exploit the glosses of synonym words to classify the terms defined by the glosses.

Pang et al. (2002) propose three different machine learning methods to extract the SO of adjectives. Their results are above a human-generated baseline, but the authors point out that discourse structure is necessary to detect and exploit the rhetorical devices used by the review authors. Machine Learning methods have also been applied to the whole problem, i.e., the classification of whole text as positive or negative, not just the classification of words (Bai et al., 2004; Gamon, 2004)

2.2 Relevant Sentences

It is obvious that not all parts of a text contribute equally to the possible overall opinion expressed therein. A movie review may contain sections relating to other movies by the same director, or with the same actors. Those sections have no or little bearing on the author's opinion towards the movie under discussion. A worse case involves texts where the author discusses a completely irrelevant topic (such as the restaurant they visited before the movie). In general, this is a topic-detection problem, to which solutions have been proposed (e.g., Yang, 1999 for statistical approaches).

A slightly different problem is that of a text that contains mostly relevant information, but where some information is more relevant than other. Less relevant aspects include background on the plot of the movie or book, or additional factual information on any aspect of the product. This problem has to do with distinguishing opinion from fact, or subjective from objective information. Janyce Wiebe and colleagues have annotated corpora with expressions of opinion (Wiebe et al., 2005), and have developed classifiers to distinguish objective from subjective sentences (Wiebe & Riloff, 2005).

Nigam and Hurst (2004) define the overall problem as one of recognizing topical sentences. Topical sentences that contain polar language (expressions of negative or positive sentiment) can then be used to capture the sentiment of the text.

Finally, another aspect of relevance is related to parts of the text that summarize or capture an overall opinion. Taboada & Grieve (2004) proposed that different weight be assigned to adjectives found in the first, second and third parts of the text, under the assumption that opinion summaries tend to appear towards the end of the text. They found a 14% improvement on the SO assigned to texts, in an evaluation that compared the results of their system to "thumbs up" or "thumbs down" evaluations given by the authors themselves. Note that this evaluation method is not foolproof: an author may assign a "recommended" or "not recommended" value that does not necessarily match what they say in the text. Also, star-based ratings (e.g., 3 out of 5 stars) are not consistent across reviewers. A reviewer's 2.5 may be more positive than another reviewer's 3 (see also the discussion in Pang & Lee, 2005).

2.3 Aggregation

Once we have extracted words from a text, with or without having used a pruning method for sentences, the next step is to aggregate the SO of those individual words. The most commonly used method for this purpose is to average the SO of the words found in the text (Turney, 2002). It has been pointed out that adjectives (if those are the primary words used) in different parts of the text may have different weights (Pang et al., 2002; Taboada & Grieve, 2004).

Aggregation methods should also exploit particular grammatical constructions and, of course, take negation into account. Polanyi and Zaenen (2004) describe negative items, intensifiers, connectors and presuppositional items as some of the items that affect the polarity of a word, phrase or sentence. Kennedy and Inkpen (2006) test this hypothesis, and show that including negation and intensifiers improves the accuracy of a classification system. Mulder et al. (2004) also discuss lexical and grammatical mechanisms that play a role in the formulation of SO.

3. Creating Dictionaries

By a dictionary (or a database) we mean a list of words annotated with their corresponding semantic orientation. For example, many researchers have taken the positive and negative words from the General Inquirer (Stone et al., 1966). The strength of the SO for those words is then extracted through different methods, as described in Section 2.1.

In order to create our own dictionaries, we first concentrate on adjectives. We aggregate the adjectives in a text to extract the opinion expressed by the text. Our initial task is to create a dictionary of adjectives with their SO. We tagged a set of 400 reviews from the Epinions website, from which we extracted a total of 1,719 adjectives. To test which SO method yields the best results, we created three different sets of SO values for the 1,719 adjectives in our dictionary.

The first set of SO values was calculated using Turney's

PMI method. The assumption is that a negative word will be found in the neighbourhood of other negative words, and conversely for positive words. Turney (2002) used Altavista, querying the search engine with the target word and NEAR positive or negative words (set to ten words in the vicinity of the word in question), and then calculating PMI using the hits. The result is a positive or negative number (SO-PMI), the degree of which determines the SO of the word or phrase in question. Our NEAR dictionary was compiled at a time when NEAR was still available (Taboada & Grieve, 2004).

Given that NEAR is no longer available, we need a new method to extract SO beyond the original set of adjectives. We decided to test the AND operator with the Google search engine. AND is likely to be less precise, since it finds two words that are located anywhere in a document that could be several pages long, as pointed out by Turney & Littman (2003). For our test, we took the original set of adjectives, and calculated SO-PMI using AND searches through Google.

Finally, we created a dictionary based on the values assigned to words in the General Inquirer, a large list of words annotated with several values [1]. We extracted only those words that overlapped with our set of adjectives (which resulted in 521 adjectives), and assigned them a 1 or -1 value, based on the positive/negative labels of the General Inquirer.

4. Tests

The tests are based on three different dictionaries that cover most of the adjectives found in a set of 400 reviews from Epinions. For each text, we produce a number that captures the opinion in the text, and is easily comparable to the opinion expressed by the reviewer, in terms of "recommended" or "not recommended". Turney & Littman (2003) compared the goodness of AND and NEAR using the General Inquirer as a benchmark. They decided that their method was performing well when the sign of the resulting SO matched the positive/negative value of a GI word. Since our goal is to use SO in determining the orientation of entire texts, we also tested which dictionary produced the highest percentage of agreement with the reviewers' own rating.

4.1 Comparisons to a Benchmark

We performed the same benchmark comparison found in Turney & Littman (2003): the number of adjectives that agree with the sign of GI words. We extracted the adjectives from GI that coincided with those in our dictionary (a total of 521 adjectives). We then compared them based on sign: if in GI the word is positive, and our sign was positive, the test was successful.

In Table 1 we show the results of that test, where we can see that AND does not perform as well as NEAR. However, even NEAR is inadequate: only about 68% of the words coincided in polarity with those in GI. These results are lower than those reported by Turney & Littman. They use, however, a much larger dictionary.

	Correct polarity	Percentage
NEAR	354	67.95%
AND	243	46.64%
n (Adjs.compared)	521	

Table 1: Comparisons to the GI dictionary

4.2 Results for Entire Texts

A better test of a dictionary consists of determining its contribution to extracting SO for entire texts. We use the three dictionaries (the full NEAR and AND, and a reduced set of GI words) to calculate semantic orientation for 400 texts from Epinions. These include reviews for: movies, books, cars, cookware, phones, hotels, music, and computers.

A crude approach to the aggregation problem is to average all the adjectives in the text. We have already shown, however, that weighting the adjectives by position in the text produces better results (Taboada & Grieve, 2004). Our results are based on weighted averages, calculated according to weights described in our earlier work. In Figure 1, we show the weights assigned to adjectives, depending on where they appear in the text.

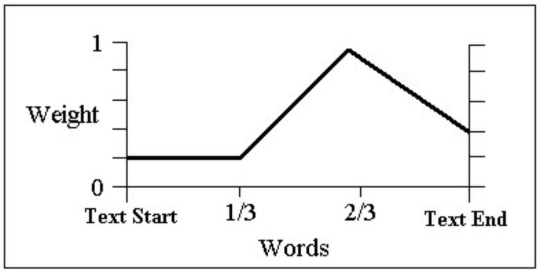

Figure 1: Prominence schema

A second approach to the calculation is to manipulate the dictionary. We assume that the dictionary contains a normal distribution of positive and negative adjectives, but the resulting values do not necessarily correspond to a scale with zero as the middle point. To make the SO values of adjectives more transparent, we calculated the median value of the dictionary, and shifted the entire list by the median value. This was done for both the AND and NEAR dictionaries. The GI dictionary is already encoded in 1 and -1 values, with zero as the middle point.

Values for each text are calculated as follows: Adjectives are extracted. If the adjective is within the scope of a negating word (e.g., *not, no, nor, neither*), its sign is flipped. Negating words are considered within scope if they are found up to five words to the left of the adjective. Then all the adjectives are averaged, using weights according to their position in the text.

The following table displays the results of our experiment. For each method, we compared our results to what reviewers said. If our result was greater than or equal to 0, and the reviewer said "recommended", then our system is correct. If the result is below 0, and the reviewer said "not

[1] http://www.wjh.harvard.edu/~inquirer/

recommended”, then the system is correct too. The table displays the number of texts where the system was correct (out of 400), and the percentage.

Dictionary	Correct texts (*n*=400)	Percentage
NEAR	211	52.75%
AND	198	49.50%
GI	201	50.25%

Table 2: Results using three different dictionaries

As we expected, the NEAR dictionary produces the best results, but the AND dictionary is not far behind. A surprising result is that the General Inquirer dictionary, a mere 521 adjectives with only polarity (no strength) performs above the AND dictionary. Upon close examination, we observed that the GI dictionary yields a large number of texts with “0” as output value (a total of 83.25% of the 400 texts). We considered a value of 0 as positive: below 0 was negative, equal to or above 0 meant a positive text. In all three cases, we are barely at a guessing baseline, which makes it obvious that mere aggregation of adjectives is not sufficient. In the next section, we show results of tests with fewer adjectives, pruned according to the strength of their SO.

4.3 Results by Confidence

We decided to perform the same tests with a smaller subset of the AND and NEAR dictionaries, based on the strength of the SO (Turney & Littman, 2003). We sorted both dictionaries according to the strength of the SO, regardless of its sign, and calculated SO values for entire texts just as described in the previous section, with the top 75%, 50%, and 25% adjectives (a total of 1,289, 859, and 430 adjectives, respectively). The hypothesis was that using only words with a strong SO would help identify the adjectives in texts that best capture their overall SO. Table 3 shows those results.

Dictionary	Accuracy		
	Top 75%	Top 50%	Top 25%
NEAR	52.75%	53.25%	48.00%
AND	50.00%	49.75%	46.25%

Table 3: Performance of pruned dictionaries

Table 3 shows that performance fluctuates when we use the top 75% and 50% of the dictionaries, as compared with the full set (see Table 2). However, performance does seem to decline if the set is too small, at 25% of the words. NEAR still outperforms AND in all cases, but not by much. As with the GI dictionary, the lower the number of adjectives in the dictionary, the higher the number of “0” output values, which are classified as positive. The number of texts with 0 was as high as 22% for AND and the top 25% adjective list, but only 0.75% for both AND and NEAR lists with the top 75% of adjectives.

4.4 Other Dictionaries

The third type of test we performed was using two existing dictionaries. They were both made available by Peter Turney[2]. The first was a list of General Inquirer words (a total of 3,596), whose SO was calculated using Altavista and the NEAR operator. The second set is the list of 1,336 adjectives from Hatzivassiloglou & McKeown (1997), with SO values calculated the same way (methods described in Turney & Littman, 2002). We used these two dictionaries to calculate SO for entire texts, as described in Section 4.2 above.

Dictionary	Correct texts (*n*=400)	Percentage
T&L GI	256	64%
T&L H&M	248	62%

Table 4: Results using external dictionaries

The conclusion seems to be that dictionaries do matter, to some extent. The GI dictionary contains words which were manually added because of their perceived sentiment value. The Hatzivassiloglou & McKeown list was built using coordination, and thus also probably includes words that more reliably indicate sentiment. Our list includes all the adjectives present in the texts, some of which may not carry any subjective meaning.

5. Google Searches

Before a final discussion of the results and of future avenues for research, we would like to draw attention to our experience using Google to calculate SO. We observed some inconsistency in the results obtained for the same word on multiple runs through Google. Since Google indexes a dynamic set of pages, results may vary, depending on which pages are available at any given time. We performed a couple of small tests, to determine whether the difference in results was significant.

The first test involved a small subset of adverbs (a new class of words beyond our initial adjective list). The adverbs were run through the Google API a total of eight times over three consecutive days. In Table 5 we show the highest and lowest values of the eight, the average value, and the standard deviation for each adverb. Although a few show standard deviation values around 1, three were above 2. It is difficult to interpret the values themselves, since the adverbs do not all have evaluative meaning

Adverb	Highest value	Lowest value	Average value	Standard deviation
away	-1.7719	-6.1984	-3.5277	1.5241
equally	-2.9020	-5.7973	-3.6547	1.0045
eventually	-0.9755	-8.3008	-4.0837	2.1861
hastily	-1.9759	-8.6323	-4.5750	2.1765
heavily	-0.9695	-9.2568	-3.8807	2.7425
madly	-6.8169	-11.5520	-8.3208	1.8720
masterfully	-2.4512	-7.8460	-3.3973	1.9736

Table 5: Adverb SO values on different days

For the second test, we extracted twenty adjectives from our list of 1,719, selected at random. We performed this comparison for eight days, once a day and at different times each day. In Table 6, we show only the highest and the lowest value obtained for each, and the average. The standard deviation was calculated over the eight values.

Although standard deviations are lower than for the adverb test, we still observe fluctuations in SO. Some of

[2] Through the SentimentAI group (http://groups.yahoo.com/group/SentimentAI/)

those may be due to whatever is on the news on a particular day, or what pages are down at any given time, but some of them are harder to explain. *Asian*, for instance, has one of the lowest standard deviations, although it could well be influenced by particular news about anything Asian (markets, governments, etc.). On the other hand, *flimsy* seems like a good candidate for a stable meaning across. However, it has the highest standard deviation of the group (albeit always negative in sign).

Adjective	Highest value	Lowest value	Average value	Standard deviation
adequate	-0.5314	-2.7420	-2.0360	0.7008
aerobic	1.1149	-2.4038	-0.9958	1.2411
Asian	-1.9146	-4.3546	-3.3133	0.7784
auxiliary	2.5832	-1.3392	0.9166	1.1620
big-breasted	-3.5606	-7.1181	-6.0920	1.1486
bored	-2.3589	-7.9899	-5.3013	1.5711
catchy	-0.2863	-4.7181	-3.3404	1.4688
emotional	-1.996	-4.9403	-3.5404	1.0929
fantastic	-0.2370	-3.4672	-1.6005	1.0152
flimsy	-2.4754	-9.4459	-5.8369	2.0491
incoming	1.1198	-1.1358	0.1001	0.7725
punk	-2.9369	-6.7838	-5.3662	1.2399
random	-2.9965	-4.3448	-3.7544	0.4632
shameless	-4.3964	-8.4925	-6.5824	1.3732
slender	-3.2786	-5.9899	-4.4707	0.9034
solid	0.7516	-2.7486	-1.4001	1.2965
stuffy	-2.7448	-7.5396	-4.9043	1.4987
sudden	-3.3002	-6.9039	-5.2828	1.1051
supernatural	-2.8298	-7.6004	-5.9618	1.5067
surreal	-1.4339	-5.3066	-4.0518	1.2250

Table 6: Adjective SO values on different days

It seems that Google is not completely reliable, and static data may be best for extracting SO. Although Turney & Littman (2003) suggest that a smaller (static) corpus yields lower accuracy, the variability of Google also poses a problem. Kennedy and Inkpen (2006) used a static corpus, with encouraging results.

6. Discussion

A remedy to the low overall accuracy that one could propose would be to increase the number of adjectives in the dictionary. We did exactly that: we improved our part-of-speech tagging methods (to take into account idiosyncratic punctuation and other aspects of on-line posts), which resulted in tagging more adjectives in the 400 texts, to a total of 3,231. We calculated SO for those using AND on Google. The overall accuracy using the larger dictionary is 56.75%. It is interesting to compare this figure with the AND results for the smaller dictionary (49.50% accuracy) and to the much smaller GI dictionary (52.75%). The accuracy does improve with a larger dictionary, but not as much as with a purely subjective dictionary. The list of words that Turney and Littman extracted from the General Inquirer results in 64% accuracy. Of course, this increase could be due to either the words themselves, or to the better calculation of SO using NEAR. We believe that using adjectives that are known to convey sentiment contributes more to accuracy than using all adjectives found in a text. The problem becomes one of determining which adjectives (or words) convey sentiment. The GI list is a good starting point, but it will be impossible to create a list of all words that convey SO. For instance, the adjective *big-breasted* (in our list in Table 6) seems to convey SO, but is unlikely to be found in any standard dictionary or thesaurus. This is an even more acute problem in on-line posts, where words are often invented, but take currency quickly.

We would also like to point out that results vary across review types. We examined eight different types of reviews: books, cars, cookware, hotels, movies, music and phones. Although the variance is not high, movies tend to get the worst results, regardless of the dictionary used. Phone and computer reviews, on the other hand, tend to have higher accuracy. It has often been pointed out that movie review writers use complex rhetorical devices. Movies are also more difficult to classify, because they may contain information about the movie, the director, or the actors, which has no direct bearing on the writer's opinion on the movie.

It is also interesting to compare the performance across reviews that are labelled by authors as positive or negative. We found that methods tended to perform well on one type, but not so well on the other. As Table 7 shows, most dictionaries tend to do better on positive reviews than on the negative ones. It has often been pointed out that reviewers do not always use negative adjectives in negative reviews, whereas they tend to use positive adjectives in positive reviews. However, the AND dictionary was the opposite of the other ones: its accuracy is the poorest of all of the dictionaries in the positive reviews, but very good on negative ones. We want to explore these differences in future work, but we can say for now that the output values for entire texts are always quite low, rarely above 1 or below -1, and most commonly around 0. Therefore, even a small change in the dictionary often results in a change of sign for the output value. Obviously, AND also has a negative bias.

Dictionary	Positive reviews correct (*n*=200)	Negative reviews correct (*n*=200)
NEAR	98.5%	7%
AND	1.5%	97.5%
GI	97%	3.5%
T&L GI	93.5%	34.5%
T&L H&M	93.5%	30.5%

Table 7: Results according to review polarity

7. Conclusions

We first conclude that there is a considerable amount of work remaining, since our results are barely above a guessing baseline. More importantly, we conclude that compiling a dictionary using the AND operator will provide results that are close to those found for NEAR in Altavista. However, Google does not seem to be a reliable search engine for this purpose, and static copies of large corpora may be more reliable. Future work aims at tagging texts more accurately, using phrases instead of just adjectives, and incorporating discourse information, in the form of rhetorical relations (Mann & Thompson, 1988).

8. Acknowledgments

This project was funded through a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada, and through NSERC's Undergraduate Student Research Award Program.

9. References

Bai, X., Padman, R. & Airoldi, E. (2004). Sentiment Extraction from Unstructured Text Using Tabu Search-Enhanced Markov Blanket (Technical Report CMU-ISRI-04-127). Pittsburgh: Carnegie Mellon University.

Esuli, A. & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss classification. Proceedings of the 14th ACM International Conference on Information and Knowledge Management. Bremen, Germany.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of COLING 2004 (pp. 841-847). Geneva, Switzerland.

Hatzivassiloglou, V. & McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of 35th Meeting of the Association for Computational Linguistics (pp. 174-181). Madrid, Spain.

Kamps, J., Marx, M., Mokken, R.J. & de Rijke, M. (2004). Using WordNet to measure semantic orientation of adjectives. Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004) (pp. 1115-1118). Lisbon, Portugal.

Kennedy, A. & Inkpen, D. (2006). Sentiment classification of movie and product reviews using contextual valence shifters. Computational Intelligence, to appear.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological Review, 104(2), 211-240.

Mann, W.C. & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. Text, 8(3), 243-281.

Mulder, M., Nijholt, A., den Uyl, M. & Terpstra, P. (2004). A lexical grammatical implementation of affect. In P. Sojka, I. Kopecek & K. Pala (Eds.), Proceedings of the 7th International Conference on Text, Speech & Dialogue (TSD 2004) (pp. 171-178). Berlin: Springer.

Nigam, K. & Hurst, M. (2004). Towards a robust metric of opinion. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (pp. 98-105). Stanford University, California.

Pang, B. & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. Proceedings of 43rd ACL. Ann Arbor, MI.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using Machine Learning techniques. Proceedings of Conference on Empirical Methods in NLP (pp. 79-86).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon, Portugal.

Stone, P.J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), Text Analysis for the Social Sciences. Mahwah, NJ: Lawrence Erlbaum.

Stone, P.J., Dunphy, D.C., Smith, M.S. & Ogilvie, D.M. (1966). The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA: MIT Press.

Taboada, M. & Grieve, J. (2004). Analyzing appraisal automatically. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (pp. 158-161). Stanford University, CA.

Turney, P. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. Proceedings of the 12th European Conference on Machine Learning (ECML-2001). Freiburg, Germany.

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of 40th ACL (pp. 417-424).

Turney, P. & Littman, M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus (No. ERB-1094, NRC #44929): National Research Council of Canada.

Turney, P. & Littman, M. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4), 315-346.

Wiebe, J. & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. Proceedings of Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005). Mexico City, Mexico.

Wiebe, J., Wilson, T. & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 1(2).

Yang, Y. (1999). An evaluation of statistical approaches to text classification. Journal of Information Retrieval, 1(1-2), 67-88.

Zaenen, A. & Polanyi, L. (2004). Contextual valence shifters. Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (pp. 106-111). Stanford University, CA.