# Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses

**Alina Andreevskaia and Sabine Bergler**
Concordia University
Montreal, Quebec, Canada
{andreev, bergler}@encs.concordia.ca

## Abstract

Many of the tasks required for semantic tagging of phrases and texts rely on a list of words annotated with some semantic features. We present a method for extracting sentiment-bearing adjectives from WordNet using the Sentiment Tag Extraction Program (STEP). We did 58 STEP runs on unique non-intersecting seed lists drawn from manually annotated list of positive and negative adjectives and evaluated the results against other manually annotated lists. The 58 runs were then collapsed into a single set of 7,813 unique words. For each word we computed a Net Overlap Score by subtracting the total number of runs assigning this word a negative sentiment from the total of the runs that consider it positive. We demonstrate that Net Overlap Score can be used as a measure of the words degree of membership in the fuzzy category of sentiment: the core adjectives, which had the highest Net Overlap scores, were identified most accurately both by STEP and by human annotators, while the words on the periphery of the category had the lowest scores and were associated with low rates of inter-annotator agreement.

## 1 Introduction

Many of the tasks required for effective semantic tagging of phrases and texts rely on a list of words annotated with some lexical semantic features. Traditional approaches to the development of such lists are based on the implicit assumption of classical truth-conditional theories of meaning representation, which regard all members of a category as equal: no element is more of a member than any other (Edmonds, 1999). In this paper, we challenge the applicability of this assumption to the semantic category of *sentiment*, which consists of positive, negative and neutral subcategories, and present a dictionary-based Sentiment Tag Extraction Program (STEP) that we use to generate a *fuzzy set* of English sentiment-bearing words for the use in sentiment tagging systems [1]. The proposed approach based on the fuzzy logic (Zadeh, 1987) is used here to assign fuzzy sentiment tags to all words in WordNet (Fellbaum, 1998), that is it assigns sentiment tags and a degree of centrality of the annotated words to the sentiment category. This assignment is based on WordNet glosses. The implications of this approach for NLP and linguistic research are discussed.

## 2 The Category of Sentiment as a Fuzzy Set

Some semantic categories have clear membership (e.g., lexical fields (Lehrer, 1974) of color, body parts or professions), while others are much more difficult to define. This prompted the development of approaches that regard the transition from membership to non-membership in a semantic category as gradual rather than abrupt (Zadeh, 1987; Rosch, 1978). In this paper we approach the category of *sentiment* as one of such fuzzy categories where some words — such as *good, bad* — are very central, prototypical members, while other, less central words may be interpreted differently by different people. Thus, as annotators proceed from the core of the category to its periphery, word mem-

[1] Sentiment tagging is defined here as assigning positive, negative and neutral labels to words according to the sentiment they express.

bership in this category becomes more ambiguous, and hence, lower inter-annotator agreement can be expected for more peripheral words. Under the classical truth-conditional approach the disagreement between annotators is invariably viewed as a sign of poor reliability of coding and is eliminated by 'training' annotators to code difficult and ambiguous cases in some standard way. While this procedure leads to high levels of inter-annotator agreement on a list created by a coordinated team of researchers, the naturally occurring differences in the interpretation of words located on the periphery of the category can clearly be seen when annotations by two independent teams are compared. The Table 1 presents the comparison of GI-H4 (General Inquirer Harvard IV-4 list, (Stone et al., 1966)) [2] and HM (from (Hatzivassiloglou and McKeown, 1997) study) lists of words manually annotated with sentiment tags by two different research teams.

| | GI-H4 | HM |
|---|---|---|
| List composition | nouns, verbs, adj., adv. | adj. only |
| Total list size | 8,211 | 1,336 |
| Total adjectives | 1,904 | 1,336 |
| Tags assigned | Positiv, Negativ or no tag | Positive or Negative |
| Adj. with non-neutral tags | 1,268 | 1,336 |
| Intersection (% intersection) | 774 (55% of GI-H4 adj) | 774 (58% of HM) |
| Agreement on tags | 78.7% | |

Table 1: Agreement between GI-H4 and HM annotations on sentiment tags.

The approach to *sentiment* as a category with fuzzy boundaries suggests that the 21.3% disagreement between the two manually annotated lists reflects a natural variability in human annotators' judgment and that this variability is related to the degree of centrality and/or relative importance of certain words to the category of sentiment. The attempts to address this difference in importance of various sentiment markers have crystallized in two main approaches: automatic assignment of weights based on some statistical criterion ((Hatzivassiloglou and McKeown, 1997; Turney and Littman, 2002; Kim and Hovy, 2004), and others) or manual annotation (Subasic and Huettner, 2001). The statistical approaches usually employ some quantitative criterion (e.g., magnitude of pointwise mutual information in (Turney and Littman, 2002), "goodness-for-fit" measure in (Hatzivassiloglou and McKeown, 1997), probability of word's sentiment given the sentiment if its synonyms in (Kim and Hovy, 2004), etc.) to define the strength of the sentiment expressed by a word or to establish a threshold for the membership in the crisp sets [3] of positive, negative and neutral words. Both approaches have their limitations: the first approach produces coarse results and requires large amounts of data to be reliable, while the second approach is prohibitively expensive in terms of annotator time and runs the risk of introducing a substantial subjective bias in annotations.

In this paper we seek to develop an approach for semantic annotation of a fuzzy lexical category and apply it to sentiment annotation of all WordNet words. The sections that follow (1) describe the proposed approach used to extract sentiment information from WordNet entries using STEP (Semantic Tag Extraction Program) algorithm, (2) discuss the overall performance of STEP on WordNet glosses, (3) outline the method for defining centrality of a word to the sentiment category, and (4) compare the results of both automatic (STEP) and manual (HM) sentiment annotations to the manually-annotated GI-H4 list, which was used as a gold standard in this experiment. The comparisons are performed separately for each of the subsets of GI-H4 that are characterized by a different distance from the core of the lexical category of sentiment.

## 3 Sentiment Tag Extraction from WordNet Entries

Word lists for sentiment tagging applications can be compiled using different methods. Automatic methods of sentiment annotation at the word level can be grouped into two major categories: (1) corpus-based approaches and (2) dictionary-based

[2]The General Inquirer (GI) list used in this study was manually cleaned to remove duplicate entries for words with same part of speech and sentiment. Only the Harvard IV-4 list component of the whole GI was used in this study, since other lists included in GI lack the sentiment annotation. Unless otherwise specified, we used the full GI-H4 list including the Neutral words that were not assigned Positiv or Negativ annotations.

[3]We use the term *crisp set* to refer to traditional, non-fuzzy sets

approaches. The first group includes methods that rely on syntactic or co-occurrence patterns of words in large texts to determine their sentiment (e.g., (Turney and Littman, 2002; Hatzivassiloglou and McKeown, 1997; Yu and Hatzivassiloglou, 2003; Grefenstette et al., 2004) and others). The majority of dictionary-based approaches use WordNet information, especially, synsets and hierarchies, to acquire sentiment-marked words (Hu and Liu, 2004; Valitutti et al., 2004; Kim and Hovy, 2004) or to measure the similarity between candidate words and sentiment-bearing words such as *good* and *bad* (Kamps et al., 2004).

In this paper, we propose an approach to sentiment annotation of WordNet entries that was implemented and tested in the Semantic Tag Extraction Program (STEP). This approach relies both on lexical relations (synonymy, antonymy and hyponymy) provided in WordNet and on the WordNet glosses. It builds upon the properties of dictionary entries as a special kind of structured text: such lexicographical texts are built to establish semantic equivalence between the left-hand and the right-hand parts of the dictionary entry, and therefore are designed to match as close as possible the components of meaning of the word. They have relatively standard style, grammar and syntactic structures, which removes a substantial source of noise common to other types of text, and finally, they have extensive coverage spanning the entire lexicon of a natural language.

The STEP algorithm starts with a small set of seed words of known sentiment value (positive or negative). This list is augmented during the first pass by adding synonyms, antonyms and hyponyms of the seed words supplied in WordNet. This step brings on average a 5-fold increase in the size of the original list with the accuracy of the resulting list comparable to manual annotations (78%, similar to HM vs. GI-H4 accuracy). At the second pass, the system goes through all WordNet glosses and identifies the entries that contain in their definitions the sentiment-bearing words from the extended seed list and adds these head words (or rather, lexemes) to the corresponding category — positive, negative or neutral (the remainder). A third, clean-up pass is then performed to partially disambiguate the identified WordNet glosses with Brill's part-of-speech tagger (Brill, 1995), which performs with up to 95% accuracy, and eliminates errors introduced into the list by part-of-speech ambiguity of some words acquired in pass 1 and from the seed list. At this step, we also filter out all those words that have been assigned contradicting, positive and negative, sentiment values within the same run.

The performance of STEP was evaluated using GI-H4 as a gold standard, while the HM list was used as a source of seed words fed into the system. We evaluated the performance of our system against the complete list of 1904 adjectives in GI-H4 that included not only the words that were marked as *Positiv, Negativ*, but also those that were not considered sentiment-laden by GI-H4 annotators, and hence were by default considered neutral in our evaluation. For the purposes of the evaluation we have partitioned the entire HM list into 58 non-intersecting seed lists of adjectives. The results of the 58 runs on these non-intersecting seed lists are presented in Table 2. The Table 2 shows that the performance of the system exhibits substantial variability depending on the composition of the seed list, with accuracy ranging from 47.6% to 87.5% percent (Mean = 71.2%, Standard Deviation (St.Dev) = 11.0%).

| | Average run size | | Average % correct | |
|---|---|---|---|---|
| | # of adj | StDev | % | StDev |
| PASS 1 (WN Relations) | 103 | 29 | 78.0% | 10.5% |
| PASS 2 (WN Glosses) | 630 | 377 | 64.5% | 10.8% |
| PASS 3 (POS clean-up) | 435 | 291 | 71.2% | 11.0% |

Table 2: Performance statistics on STEP runs.

The significant variability in accuracy of the runs (Standard Deviation over 10%) is attributable to the variability in the properties of the seed list words in these runs. The HM list includes some sentiment-marked words where not all meanings are laden with sentiment, but also the words where some meanings are neutral and even the words where such neutral meanings are much more frequent than the sentiment-laden ones. The runs where seed lists included such ambiguous adjectives were labeling a lot of neutral words as sentiment marked since such seed words were more likely to be found in the WordNet glosses in their more frequent neutral meaning. For example, run # 53 had in its seed list two ambiguous adjectives

*dim* and *plush*, which are neutral in most of the contexts. This resulted in only 52.6% accuracy (18.6% below the average). Run # 48, on the other hand, by a sheer chance, had only unambiguous sentiment-bearing words in its seed list, and, thus, performed with a fairly high accuracy (87.5%, 16.3% above the average).

In order to generate a comprehensive list covering the entire set of WordNet adjectives, the 58 runs were then collapsed into a single set of unique words. Since many of the clearly sentiment-laden adjectives that form the core of the category of sentiment were identified by STEP in multiple runs and had, therefore, multiple duplicates in the list that were counted as one entry in the combined list, the collapsing procedure resulted in a lower-accuracy (66.5% - when GI-H4 neutrals were included) but much larger list of English adjectives marked as positive ($n = 3,908$) or negative ($n = 3,905$). The remainder of WordNet's $22,141$ adjectives was not found in any STEP run and hence was deemed neutral ($n = 14,328$).

Overall, the system's 66.5% accuracy on the collapsed runs is comparable to the accuracy reported in the literature for other systems run on large corpora (Turney and Littman, 2002; Hatzivassiloglou and McKeown, 1997). In order to make a meaningful comparison with the results reported in (Turney and Littman, 2002), we also did an evaluation of STEP results on positives and negatives only (i.e., the neutral adjectives from GI-H4 list were excluded) and compared our labels to the remaining 1266 GI-H4 adjectives. The accuracy on this subset was 73.4%, which is comparable to the numbers reported by Turney and Littman (2002) for experimental runs on $3,596$ sentiment-marked GI words from different parts of speech using a $2x10^9$ corpus to compute point-wise mutual information between the GI words and 14 manually selected positive and negative paradigm words (76.06%).

The analysis of STEP system performance vs. GI-H4 and of the disagreements between manually annotated HM and GI-H4 showed that the greatest challenge with sentiment tagging of words lies at the boundary between sentiment-marked (positive or negative) and sentiment-neutral words. The 7% performance gain (from 66.5% to 73.4%) associated with the removal of neutrals from the evaluation set emphasizes the importance of neutral words as a major source of sentiment extraction system errors [4]. Moreover, the boundary between sentiment-bearing (positive or negative) and neutral words in GI-H4 accounts for 93% of disagreements between the labels assigned to adjectives in GI-H4 and HM by two independent teams of human annotators. The view taken here is that the vast majority of such inter-annotator disagreements are not really errors but a reflection of the natural ambiguity of the words that are located on the periphery of the sentiment category.

## 4 Establishing the degree of word's centrality to the semantic category

The approach to sentiment category as a fuzzy set ascribes the category of sentiment some specific structural properties. First, as opposed to the words located on the periphery, more central elements of the set usually have stronger and more numerous semantic relations with other category members [5]. Second, the membership of these central words in the category is less ambiguous than the membership of more peripheral words. Thus, we can estimate the centrality of a word in a given category in two ways:

1. Through the density of the word's relationships with other words — by enumerating its semantic ties to other words within the field, and calculating membership scores based on the number of these ties; and

2. Through the degree of word membership ambiguity — by assessing the inter-annotator agreement on the word membership in this category.

Lexicographical entries in the dictionaries, such as WordNet, seek to establish semantic equivalence between the word and its definition and provide a rich source of human-annotated relationships between the words. By using a bootstrapping system, such as STEP, that follows the links between the words in WordNet to find similar words, we can identify the paths connecting members of a given semantic category in the dictionary. With multiple bootstrapping runs on different seed

[4] It is consistent with the observation by Kim and Hovy (2004) who noticed that, when positives and neutrals were collapsed into the same category opposed to negatives, the agreement between human annotators rose by 12%.

[5] The operationalizations of centrality derived from the number of connections between elements can be found in social network theory (Burt, 1980)

lists, we can then produce a measure of the density of such ties. The ambiguity measure derived from inter-annotator disagreement can then be used to validate the results obtained from the density-based method of determining centrality.

In order to produce a centrality measure, we conducted multiple runs with non-intersecting seed lists drawn from HM. The lists of words fetched by STEP on different runs partially overlapped, suggesting that the words identified by the system many times as bearing positive or negative sentiment are more central to the respective categories. The number of times the word has been fetched by STEP runs is reflected in the *Gross Overlap* Measure produced by the system. In some cases, there was a disagreement between different runs on the sentiment assigned to the word. Such disagreements were addressed by computing the *Net Overlap* Scores for each of the found words: the total number of runs assigning the word a negative sentiment was subtracted from the total of the runs that consider it positive. Thus, the greater the number of runs fetching the word (i.e., Gross Overlap) and the greater the agreement between these runs on the assigned sentiment, the higher the Net Overlap Score of this word.

The Net Overlap scores obtained for each identified word were then used to stratify these words into groups that reflect positive or negative distance of these words from the zero score. The zero score was assigned to (a) the WordNet adjectives that were not identified by STEP as bearing positive or negative sentiment [6] and to (b) the words with equal number of positive and negative hits on several STEP runs. The performance measures for each of the groups were then computed to allow the comparison of STEP and human annotator performance on the words from the core and from the periphery of the sentiment category. Thus, for each of the Net Overlap Score groups, both automatic (STEP) and manual (HM) sentiment annotations were compared to human-annotated GI-H4, which was used as a gold standard in this experiment.

On 58 runs, the system has identified 3,908 English adjectives as positive, 3,905 as negative, while the remainder (14,428) of WordNet's 22,141 adjectives was deemed neutral. Of these 14,328 adjectives that STEP runs deemed neutral,

[6]The seed lists fed into STEP contained positive or negative, but no neutral words, since HM, which was used as a source for these seed lists, does not include any neutrals.

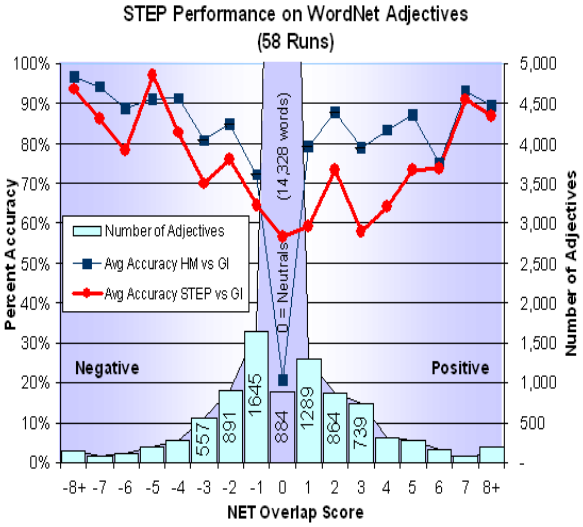


Figure 1: Accuracy of word sentiment tagging.

884 were also found in GI-H4 and/or HM lists, which allowed us to evaluate STEP performance and HM-GI agreement on the subset of neutrals as well. The graph in Figure 1 shows the distribution of adjectives by Net Overlap scores and the average accuracy/agreement rate for each group.

Figure 1 shows that the greater the Net Overlap Score, and hence, the greater the distance of the word from the neutral subcategory (i.e., from zero), the more accurate are STEP results and the greater is the agreement between two teams of human annotators (HM and GI-H4). On average, for all categories, including neutrals, the accuracy of STEP vs. GI-H4 was 66.5%, human-annotated HM had 78.7% accuracy vs. GI-H4. For the words with Net Overlap of $\pm 7$ and greater, both STEP and HM had accuracy around 90%. The accuracy declined dramatically as Net Overlap scores approached zero (= Neutrals). In this category, human-annotated HM showed only 20% agreement with GI-H4, while STEP, which deemed these words neutral, rather than positive or negative, performed with 57% accuracy.

These results suggest that the two measures of word centrality, Net Overlap Score based on multiple STEP runs and the inter-annotator agreement (HM vs. GI-H4), are directly related [7]. Thus, the Net Overlap Score can serve as a useful tool in the identification of core and peripheral members of a fuzzy lexical category, as well as in predic-

[7]In our sample, the coefficient of correlation between the two was 0.68. The *Absolute* Net Overlap Score on the subgroups 0 to 10 was used in calculation of the coefficient of correlation.

tion of inter-annotator agreement and system performance on a subgroup of words characterized by a given Net Overlap Score value.

In order to make the Net Overlap Score measure usable in sentiment tagging of texts and phrases, the absolute values of this score should be normalized and mapped onto a standard $[0, 1]$ interval. Since the values of the Net Overlap Score may vary depending on the number of runs used in the experiment, such mapping eliminates the variability in the score values introduced with changes in the number of runs performed. In order to accomplish this normalization, we used the value of the Net Overlap Score as a parameter in the standard fuzzy membership S-function (Zadeh, 1975; Zadeh, 1987). This function maps the absolute values of the Net Overlap Score onto the interval from 0 to 1, where 0 corresponds to the absence of membership in the category of sentiment (in our case, these will be the neutral words) and 1 reflects the highest degree of membership in this category. The function can be defined as follows:

$$S(u;\alpha,\beta,\gamma) = \begin{cases} 0 \text{ for } u \leq \alpha \\ 2(\frac{u-\alpha}{\gamma-\alpha})^2 \text{ for } \alpha \leq u \leq \beta \\ 1 - 2(\frac{u-\alpha}{\gamma-\alpha})^2 \text{ for } \beta \leq u \leq \gamma \\ 1 \text{ for } u \geq \gamma \end{cases}$$

where $u$ is the Net Overlap Score for the word and $\alpha, \beta, \gamma$ are the three adjustable parameters: $\alpha$ is set to 1, $\gamma$ is set to 15 and $\beta$, which represents a crossover point, is defined as $\beta = (\gamma + \alpha)/2 = 8$. Defined this way, the S-function assigns highest degree of membership (=1) to words that have the the Net Overlap Score $u \geq 15$. The accuracy vs. GI-H4 on this subset is 100%. The accuracy goes down as the degree of membership decreases and reaches 59% for values with the lowest degrees of membership.

## 5 Discussion and conclusions

This paper contributes to the development of NLP and semantic tagging systems in several respects.

- ***The structure of the semantic category of sentiment.*** The analysis of the category of sentiment of English adjectives presented here suggests that this category is structured as a fuzzy set: the distance from the core of the category, as measured by Net Overlap scores derived from multiple STEP runs, is shown to affect both the level of inter-annotator agreement and the system performance vs. human-annotated gold standard.

- ***The list of sentiment-bearing adjectives***. The list produced and cross-validated by multiple STEP runs contains $7,814$ positive and negative English adjectives, with an average accuracy of 66.5%, while the human-annotated list HM performed at 78.7% accuracy vs. the gold standard (GI-H4) [8]. The remaining $14,328$ adjectives were not identified as sentiment marked and therefore were considered neutral.

  The stratification of adjectives by their Net Overlap Score can serve as an indicator of their degree of membership in the category of (positive/negative) sentiment. Since low degrees of membership are associated with greater ambiguity and inter-annotator disagreement, the Net Overlap Score value can provide researchers with a set of volume/accuracy trade-offs. For example, by including only the adjectives with the Net Overlap Score of 4 and more, the researcher can obtain a list of $1,828$ positive and negative adjectives with accuracy of 81% vs. GI-H4, or $3,124$ adjectives with 75% accuracy if the threshold is set at 3. The normalization of the Net Overlap Score values for the use in phrase and text-level sentiment tagging systems was achieved using the fuzzy membership function that we proposed here for the category of sentiment of English adjectives.

  Future work in the direction laid out by this study will concentrate on two aspects of system development. First further incremental improvements to the precision of the STEP algorithm will be made to increase the accuracy of sentiment annotation through the use of adjective-noun combinatorial patterns within glosses. Second, the resulting list of adjectives annotated with sentiment and with the degree of word membership in the category (as measured by the Net Overlap Score) will be used in sentiment tagging of phrases and texts. This will enable us to compute the degree of importance of sentiment markers found in phrases and texts. The availability

[8] GI-H4 contains 1268 and HM list has 1336 positive and negative adjectives. The accuracy figures reported here include the errors produced at the boundary with neutrals.

of the information on the degree of centrality of words to the category of sentiment may improve the performance of sentiment determination systems built to identify the sentiment of entire phrases or texts.

- ***System evaluation considerations***. The contribution of this paper to the development of methodology of system evaluation is twofold. First, this research emphasizes the importance of multiple runs on different seed lists for a more accurate evaluation of sentiment tag extraction system performance. We have shown how significantly the system results vary, depending on the composition of the seed list.

  Second, due to the high cost of manual annotation and other practical considerations, most bootstrapping and other NLP systems are evaluated on relatively small manually annotated gold standards developed for a given semantic category. The implied assumption is that such a gold standard represents a random sample drawn from the population of all category members and hence, system performance observed on this gold standard can be projected to the whole semantic category. Such extrapolation is not justified if the category is structured as a lexical field with fuzzy boundaries: in this case the precision of both machine and human annotation is expected to fall when more peripheral members of the category are processed. In this paper, the sentiment-bearing words identified by the system were stratified based on their Net Overlap Score and evaluated in terms of accuracy of sentiment annotation within each stratum. These strata, derived from Net Overlap scores, reflect the degree of centrality of a given word to the semantic category, and, thus, provide greater assurance that system performance on other words with the same Net Overlap Score will be similar to the performance observed on the intersection of system results with the gold standard.

- ***The role of the inter-annotator disagreement***. The results of the study presented in this paper call for reconsideration of the role of inter-annotator disagreement in the development of lists of words manually annotated with semantic tags. It has been shown here that the inter-annotator agreement tends to fall as we proceed from the core of a fuzzy semantic category to its periphery. Therefore, the disagreement between the annotators does not necessarily reflect a quality problem in human annotation, but rather a structural property of the semantic category. This suggests that inter-annotator disagreement rates can serve as an important source of empirical information about the structural properties of the semantic category and can help define and validate fuzzy sets of semantic category members for a number of NLP tasks and applications.

## References


Eric Brill. 1995. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Computational Linguistics*, 21(4):543–565.

R.S. Burt. 1980. Models of network structure. *Annual Review of Sociology*, 6:79–141.

Philip Edmonds. 1999. *Semantic representations of near-synonyms for automatic lexical choice*. Ph.D. thesis, University of Toronto.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA.

Gregory Grefenstette, Yan Qu, David A. Evans, and James G. Shanahan. 2004. Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes. In Yan Qu, James Shanahan, and Janyce Wiebe, editors, *Exploring Attitude and Affect in Text: Theories and Applications*, AAAI-2004 Spring Symposium Series, pages 71–78.

Vasileios Hatzivassiloglou and Kathleen B. McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *35th ACL*, pages 174–181.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD-04*, pages 168–177.

Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. 2004. Using WordNet to measure semantic orientation of adjectives. In *LREC 2004*, volume IV, pages 1115–1118.

Soo-Min Kim and Edward Hovy. 2004. Determining the sentiment of opinions. In *COLING-2004*, pages 1367–1373, Geneva, Switzerland.

Adrienne Lehrer. 1974. *Semantic Fields and Lexical Structure.* North Holland, Amsterdam and New York.

Eleanor Rosch. 1978. Principles of Categorization. In Eleanor Rosch and Barbara B. Lloyd, editors, *Cognition and Categorization*, pages 28–49. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

P.J. Stone, D.C. Dumphy, M.S. Smith, and D.M. Ogilvie. 1966. *The General Inquirer: a computer approach to content analysis.* M.I.T. studies in comparative politics. M.I.T. Press, Cambridge, MA.

Pero Subasic and Alison Huettner. 2001. Affect Analysis of Text Using Fuzzy Typing. *IEEE-FS*, 9:483–496.

Peter Turney and Michael Littman. 2002. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada.

Alessandro Valitutti, Carlo Strapparava, and Oliviero Stock. 2004. Developing Affective Lexical Resources. *PsychNology Journal*, 2(1):61–83.

Hong Yu and Vassileios Hatzivassiloglou. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-03)*.

Lotfy A. Zadeh. 1975. Calculus of Fuzzy Restrictions. In L.A. Zadeh, K.-S. Fu, K. Tanaka, and M. Shimura, editors, *Fuzzy Sets and their Applications to cognitive and decision processes*, pages 1–40. Academic Press Inc., New-York.

Lotfy A. Zadeh. 1987. PRUF — a Meaning Representation Language for Natural Languages. In R.R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen, editors, *Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh*, pages 499–568. John Wiley & Sons.