Automatic Identification of Pro and Con Reasons in Online Reviews

Soo-Min Kim and Eduard Hovy
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292-6695
{skim, hovy}@ISI.EDU

Abstract

In this paper, we present a system that automatically extracts the pros and cons from online reviews. Although many approaches have been developed for extracting opinions from text, our focus here is on extracting the reasons of the opinions, which may themselves be in the form of either fact or opinion. Leveraging online review sites with author-generated pros and cons, we propose a system for aligning the pros and cons to their sentences in review texts. A maximum entropy model is then trained on the resulting labeled set to subsequently extract pros and cons from online review sites that do not explicitly provide them. Our experimental results show that our resulting system identifies pros and cons with 66% precision and 76% recall.

1 Introduction

Many opinions are being expressed on the Web in such settings as product reviews, personal blogs, and news group message boards. People increasingly participate to express their opinions online. This trend has raised many interesting and challenging research topics such as subjectivity detection, semantic orientation classification, and review classification.

Subjectivity detection is the task of identifying subjective words, expressions, and sentences. (Wiebe et al., 1999; Hatzivassiloglou and Wiebe, 2000; Riloff et al, 2003). Identifying subjectivity helps separate opinions from fact, which may be useful in question answering, summarization, etc.

Semantic orientation classification is a task of determining positive or negative sentiment of words (Hatzivassiloglou and McKeown, 1997; Turney, 2002; Esuli and Sebastiani, 2005). Sentiment of phrases and sentences has also been studied in (Kim and Hovy, 2004; Wilson et al., 2005). Document level sentiment classification is mostly applied to reviews, where systems assign a positive or negative sentiment for a whole review document (Pang et al., 2002; Turney, 2002).

Building on this work, more sophisticated problems in the opinion domain have been studied by many researchers. (Bethard et al., 2004; Choi et al., 2005; Kim and Hovy, 2006) identified the holder (source) of opinions expressed in sentences using various techniques. (Wilson et al., 2004) focused on the strength of opinion clauses, finding strong and weak opinions. (Chklovski, 2006) presented a system that aggregates and quantifies degree assessment of opinions scattered throughout web pages.

Beyond document level sentiment classification in online product reviews, (Hu and Liu, 2004; Popescu and Etzioni, 2005) concentrated on mining and summarizing reviews by extracting opinion sentences regarding product features.

In this paper, we focus on another challenging yet critical problem of opinion analysis, identifying *reasons* for opinions, especially for opinions in online product reviews. The opinion reason identification problem in online reviews seeks to answer the question "*What are the reasons that the author of this review likes or dislikes the product?"* For example, in hotel reviews, information such as "found 189 positive reviews and 65 negative reviews" may not fully satisfy the information needs of different users. More useful information would be "This hotel is great for families with young infants" or "Elevators are grouped according to floors, which makes the wait short".

This work differs in important ways from studies in (Hu and Liu, 2004) and (Popescu and Etzioni, 2005). These approaches extract features

of products and identify sentences that contain opinions about those features by using opinion words and phrases. Here, we focus on extracting pros and cons which include not only sentences that contain opinion-bearing expressions about products and features but also sentences with reasons why an author of a review writes the review. Following are examples identified by our system.

```
It creates duplicate files.
Video drains battery.
It won't play music from all
music stores
```

Even though finding reasons in opinion-bearing texts is a critical part of in-depth opinion assessment, no study has been done in this particular vein partly because there is no annotated data. Labeling each sentence is a time-consuming and costly task. In this paper, we propose a framework for automatically identifying reasons in online reviews and introduce a novel technique to automatically label training data for this task. We assume reasons in an online review document are closely related to pros and cons represented in the text. We leverage the fact that reviews on some websites such as epinions.com already contain pros and cons written by the same author as the reviews. We use those pros and cons to automatically label sentences in the reviews on which we subsequently train our classification system. We then apply the resulting system to extract pros and cons from reviews in other websites which do not have specified pros and cons.

This paper is organized as follows: Section 2 describes a definition of reasons in online reviews in terms of pros and cons. Section 3 presents our approach to identify them and Section 4 explains our automatic data labeling process. Section 5 describes experimental and results and finally, in Section 6, we conclude with future work.

2 Pros and Cons in Online Reviews

This section describes how we define reasons in online reviews for our study. First, we take a look at how researchers in Computational Linguistics define an opinion for their studies. It is difficult to define what an opinion means in a computational model because of the difficulty of determining the unit of an opinion. In general, researchers study opinion at three different levels: *word level, sentence level*, and *document level*.

Word level opinion analysis includes word sentiment classification, which views single lexical items (such as *good* or *bad*) as sentiment carriers, allowing one to classify words into positive and negative semantic categories. Studies in sentence level opinion regard the sentence as a minimum unit of opinion. Researchers try to identify opinion-bearing sentences, classify their sentiment, and identify opinion holders and topics of opinion sentences. Document level opinion analysis has been mostly applied to review classification, in which a whole document written for a review is judged as carrying either positive or negative sentiment. Many researchers, however, consider a whole document as the unit of an opinion to be too coarse.

In our study, we take the approach that a review text has a main opinion (recommendation or not) about a given product, but also includes various reasons for recommendation or non-recommendation, which are valuable to identify. Therefore, we focus on detecting those reasons in online product review. We also assume that reasons in a review are closely related to pros and cons expressed in the review. Pros in a product review are sentences that describe reasons why an author of the review likes the product. Cons are reasons why the author doesn't like the product. Based on our observation in online reviews, most reviews have both pros and cons even if sometimes one of them dominates.

3 Finding Pros and Cons

This section describes our approach for finding pro and con sentences given a review text. We first collect data from epinions.com and automatically label each sentences in the data set. We then model our system using one of the machine learning techniques that have been successfully applied to various problems in Natural Language Processing. This section also describes features we used for our model.

3.1 Automatically Labeling Pro and Con Sentences

Among many web sites that have product reviews such as amazon.com and epinions.com, some of them (e.g. epinions.com) explicitly state pros and cons phrases in their respective categories by each review's author along with the review text. First, we collected a large set of <review text, pros, cons> triplets from epin-

ions.com. A review document in epinions.com consists of a topic (a product model, restaurant name, travel destination, etc.), pros and cons (mostly a few keywords but sometimes complete sentences), and the review text. Our automatic labeling system first collects phrases in pro and con fields and then searches the main review text in order to collect sentences corresponding to those phrases. Figure 1 illustrates the automatic labeling process.

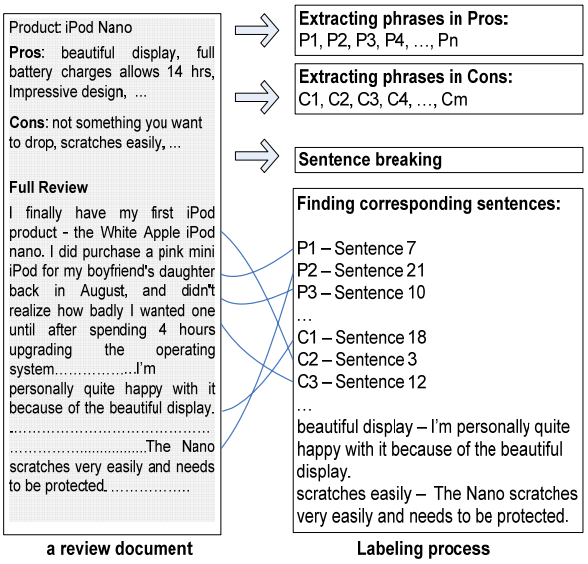

Figure 1. The automatic labeling process of pros and cons sentences in a review.

The system first extracts comma-delimited phrases from each pro and con field, generating two sets of phrases: {*P1, P2, …, Pn*} for pros and {*C1, C2, …, Cm*} for cons. In the example in Figure 1, "beautiful display" can be P_i and "not something you want to drop" can be C_j. Then the system compares these phrases to the sentences in the text in the "Full Review". For each phrase in {*P1, P2, …, Pn*} and {*C1, C2, …, Cm*}, the system checks each sentence to find a sentence that covers most of the words in the phrase. Then the system annotates this sentence with the appropriate "pro" or "con" label. All remaining sentences with neither label are marked as "neither". After labeling all the epinion data, we use it to train our pro and con sentence recognition system.

3.2 Modeling with Maximum Entropy Classification

We use Maximum Entropy classification for the task of finding pro and con sentences in a given review. Maximum Entropy classification has been successfully applied in many tasks in natural language processing, such as Semantic Role labeling, Question Answering, and Information Extraction.

Table 1: Classes defined for the classification tasks.

Class symbol	Description
PR	Sentences related to pros in a review
CR	Sentences related to cons in a review
NR	Sentences related to neither PR nor CR

Maximum Entropy models implement the intuition that the best model is the one that is consistent with the set of constraints imposed by the evidence but otherwise is as uniform as possible (Berger et al., 1996). We modeled the conditional probability of a class c given a feature vector x as follows:

$$p(c \mid x) = \frac{1}{Z_x} \exp(\sum_i \lambda_i f_i(c,x))$$

where Z_x is a normalization factor which can be calculated by the following:

$$Z_x = \sum_c \exp(\sum_i \lambda_i f_i(c,x))$$

In the first equation, $f_i(c,x)$ is a feature function which has a binary value, 0 or 1. λ_i is a weight parameter for the feature function $f_i(c,x)$ and higher value of the weight indicates that $f_i(c,x)$ is an important feature for a class c. For our system development, we used MegaM toolkit [1] which implements the above intuition.

In order to build an efficient model, we separated the task of finding pro and con sentences into two phases, each being a binary classification. The first is an identification phase and the second is a classification phase. For this 2-phase model, we defined the 3 classes of c listed in Table 1. The identification task separates pro and con candidate sentences (CR and PR in Table 1) from sentences irrelevant to either of them (NR). The classification task then classifies candidates into pros (PR) and cons (CR). Section 5 reports system results of both phases.

[1] http://www.isi.edu/~hdaume/megam/index.html

3.3 Features

The classification uses three types of features: lexical features, positional features, and opinion-bearing word features.

For **lexical features**, we use unigrams, bigrams, and trigrams collected from the training set. They investigate the intuition that there are certain words that are frequently used in pro and con sentences which are likely to represent reasons why an author writes a review. Examples of such words and phrases are: "*because"* and *"that's why"*.

For **positional features**, we first find paragraph boundaries in review texts using html tags such as *
* and *<p>*. After finding paragraph boundaries, we add features indicating the first, the second, the last, and the second last sentence in a paragraph. These features test the intuition used in document summarization that important sentences that contain topics in a text have certain positional patterns in a paragraph (Lin and Hovy, 1997), which may apply because reasons like pros and cons in a review document are most important sentences that summarize the whole point of the review.

For **opinion-bearing word features**, we used pre-selected opinion-bearing words produced by a combination of two methods. The first method derived a list of opinion-bearing words from a large news corpus by separating opinion articles such as letters or editorials from news articles which simply reported news or events. The second method calculated semantic orientations of words based on WordNet[2] synonyms. In our previous work (Kim and Hovy, 2005), we demonstrated that the list of words produced by a combination of those two methods performed very well in detecting opinion bearing sentences. Both algorithms are described in that paper.

The motivation for including the list of opinion-bearing words as one of our features is that pro and con sentences are quite likely to contain opinion-bearing expressions (even though some of them are only facts), such as "The waiting time was horrible" and "Their portion size of food was extremely generous!" in restaurant reviews. We presumed pro and con sentences containing only facts, such as "*The battery lasted 3 hours, not 5 hours like they advertised*", would be captured by lexical or positional features.

In Section 5, we report experimental results with different combinations of these features.

Table 2: Feature summary.

Feature category	Description	Symbol
Lexical Features	unigrams bigrams trigrams	***Lex***
Positional Features	the first, the second, the last, the second to last sentence in a paragraph	***Pos***
Opinion-bearing word features	pre-selected opinion-bearing words	***Op***

Table 2 summarizes the features we used for our model and the symbols we will use in the rest of this paper.

4 Data

We collected data from two different sources: epinions.com and complaints.com[3] (see Section 3.1 for details about review data in epinion.com). Data from epinions.com is mostly used to train the system whereas data from complaints.com is to test how the trained model performs on new data.

Complaints.com includes a large database of publicized consumer complaints about diverse products, services, and companies collected for over 6 years. Interestingly, reviews in complaint.com are somewhat different from many other web sites which are directly or indirectly linked to Internet shopping malls such as amazon.com and epinions.com. The purpose of reviews in complaints.com is to share consumers' mostly negative experiences and alert businesses to customers feedback. However, many reviews in Internet shopping mall related reviews are positive and sometimes encourage people to buy more products or to use more services.

Despite its significance, however, there is no hand-annotated data that we can use to build a system to identify reasons of complaints.com. In order to solve this problem, we assume that reasons in complaints reviews are similar to cons in other reviews and therefore if we are, somehow, able to build a system that can identify cons from

[2] http://wordnet.princeton.edu/

[3] http://www.complaints.com/

reviews, we can apply it to identify reasons in complaints reviews. Based on this assumption, we learn a system using the data from epinions.com, to which we can apply our automatic data labeling technique, and employ the resulting system to identify reasons from reviews in complaint.com. The following sections describe each data set.

4.1 Dataset 1: Automatically Labeled Data

We collected two different domains of reviews from epinions.com: product reviews and restaurant reviews. As for the product reviews, we collected 3241 reviews (115029 sentences) about mp3 players made by various manufacturers such as Apple, iRiver, Creative Lab, and Samsung. We also collected 7524 reviews (194393 sentences) about various types of restaurants such as family restaurants, Mexican restaurants, fast food chains, steak houses, and Asian restaurants. The average numbers of sentences in a review document are 35.49 and 25.89 respectively.

The purpose of selecting one of electronics products and restaurants as topics of reviews for our study is to test our approach in two extremely different situations. Reasons why consumers like or dislike a product in electronics' reviews are mostly about specific and tangible features. Also, there are somewhat a fixed set of features of a specific type of product, for example, ease of use, durability, battery life, photo quality, and shutter lag for digital cameras. Consequently, we can expect that reasons in electronics' reviews may share those product feature words and words that describe aspects of features such as *short* or *long* for *battery life*. This fact might make the reason identification task easy.

On the other hand, restaurant reviewers talk about very diverse aspects and abstract features as reasons. For example, reasons such as "You feel like you are in a train station or a busy amusement park that is ill-staffed to meet demand!", "preferential treatment given to large groups", and "they don't offer salads of any kind" are hard to predict. Also, they seem rarely share common keyword features.

We first automatically labeled each sentence in those reviews collected from each domain with the features described in Section 3.1. We divided the data for training and testing. We then trained our model using the training set and tested it to see if the system can successfully label sentences in the test set.

4.2 Dataset 2: Complaints.com Data

From the database [4] in complaints.com, we searched for the same topics of reviews as Dataset 1: 59 complaints reviews about mp3 players and 322 reviews about restaurants[5]. We tested our system on this dataset and compare the results against human judges' annotation results. Subsection 5.2 reports the evaluation results.

5 Experiments and Results

We describe two goals in our experiments in this section. The first is to investigate how well our pro and con detection model with different feature combinations performs on the data we collected from epinions.com. The second is to see how well the trained model performs on new data from a different source, complaint.com.

For both datasets, we carried out two separate sets of experiments, for the domains of mp3 players and restaurant reviews. We divided data into 80% for training, 10% for development, and 10% for test for our experiments.

5.1 Experiments on Dataset 1

Identification step: Table 3 and 4 show pros and cons sentences identification results of our system for mp3 player and restaurant reviews respectively. The first column indicates which combination of features was used for our model (see Table 2 for the meaning of Op, Lex, and Pos feature categories). We measure the performance with accuracy (*Acc*), precision (*Prec*), recall (*Recl*), and F-score [6].

The baseline system assigned all sentences as reason and achieved 57.75% and 54.82% of accuracy. The system performed well when it only used lexical features in mp3 player reviews (76.27% of accuracy in Lex), whereas it performed well with the combination of lexical and opinion features in restaurant reviews (Lex+Op row in Table 4).

It was very interesting to see that the system achieved a very low score when it only used opinion word features. We can interpret this phenomenon as supporting our hypothesis that pro and con sentences in reviews are often purely

[4] At the time (December 2005), there were total 42593 complaint reviews available in the database.

[5] Average numbers of sentences in a complaint is 19.57 for mp3 player reviews and 21.38 for restaurant reviews.

[6] We calculated F-score by $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Table 3: Pros and cons sentences identification results on mp3 player reviews.

Features used	**Acc** (%)	**Prec** (%)	**Recl** (%)	**F-score** (%)
Op	60.15	65.84	57.31	61.28
Lex	**76.27**	66.18	**76.42**	**70.93**
Lex+Pos	63.10	**71.14**	60.72	65.52
Lex+Op	62.75	70.64	60.07	64.93
Lex+Pos+Op	62.23	70.58	59.35	64.48
Baseline	**57.75**			

Table 4: Reason sentence identification results on restaurant reviews.

Features used	**Acc** (%)	**Prec** (%)	**Recl** (%)	**F-score** (%)
Op	61.64	60.76	47.48	53.31
Lex	63.77	67.10	51.20	58.08
Lex+Pos	**63.89**	67.62	51.70	58.60
Lex+Op	61.66	**69.13**	**54.30**	**60.83**
Lex+Pos+Op	63.13	66.80	50.41	57.46
Baseline	**54.82**			

Table 5: Pros and cons sentences classification results for mp3 player reviews.

Features used	**Acc** (%)	Cons			Pros		
		Prec (%)	**Recl** (%)	**F-score** (%)	**Prec** (%)	**Recl** (%)	**F-score** (%)
Op	**57.18**	54.43	67.10	60.10	**61.18**	48.00	**53.80**
Lex	55.88	55.49	67.45	60.89	56.52	43.88	49.40
Lex+Pos	55.62	55.26	**68.12**	**61.02**	56.24	42.62	48.49
Lex+Op	55.60	55.46	64.63	59.70	55.81	46.26	50.59
Lex+Pos+Op	56.68	**56.70**	62.45	59.44	56.65	**50.71**	53.52
baseline	**53.87**	(mark all as pros)					

Table 6: Pros and cons sentences classification results for restaurant reviews.

Features used	**Acc** (%)	Cons			Pros		
		Prec (%)	**Recl** (%)	**F-score** (%)	**Prec** (%)	**Recl** (%)	**F-score** (%)
Op	**57.32**	54.78	51.62	53.15	**59.32**	**62.35**	**60.80**
Lex	55.76	55.94	52.52	54.18	55.60	58.97	57.24
Lex+Pos	56.07	**56.20**	**53.33**	**54.73**	55.94	58.78	57.33
Lex+Op	55.88	56.10	52.39	54.18	55.68	59.34	57.45
Lex+Pos+Op	55.79	55.89	53.17	54.50	55.70	58.38	57.01
baseline	**50.71**	(mark all as pros)					

factual. However, opinion features improved both precision and recall when combined with lexical features in restaurant reviews. It was also interesting that experiments on mp3 players reviews achieved mostly higher scores than restaurants. Like the observation we described in Subsection 4.1, frequently mentioned keywords of product features (e.g. durability) may have helped performance, especially with lexical features. Another interesting observation is that the positional features that helped in topic sentence identification did not help much for our task.

Classification step: Tables 5 and 6 show the system results of the pro and con classification task. The baseline system marked all sentences as pros and achieved 53.87% and 50.71% accuracy for each domain. All features performed better than the baseline but the results are not as good as in the identification task. Unlike the identification task, opinion words by themselves achieved the best accuracy in both mp3 player and restaurant domains. We think opinion words played more important roles in classifying pros and cons than identifying them. Position features helped recognizing con sentences in mp3 player reviews.

5.2 Experiments on Dataset 2

This subsection reports the evaluation results of our system on Dataset 2. Since Dataset 2 from complaints.com has no training data, we trained a system on Dataset 1 and applied it to Dataset 2.

A tough question, however, is how to evaluate the system results. Since it seemed impossible to evaluate the system without involving a human judge, we annotated a small set of data manually for evaluation purposes.

Gold Standard Annotation: Four humans annotated 3 sets of test sets: Testset 1 with 5 complaints (73 sentences), Testset 2 with 7 complaints (105 sentences), and Testset 3 with 6 complaints (85 sentences). Testset 1 and 2 are from mp3 player complaints and Testset 3 is from restaurant reviews. Annotators marked sentences if they describe specific reasons of the complaint. Each test set was annotated by 2 humans. The average pair-wise human agreement was 82.1%[7].

System Performance: Like the human annotators, our system also labeled reason sentences. Since our goal is to identify reason sentences in complaints, we applied a system modeled as in the identification phase described in Subsection 3.2 instead of the classification phase[8]. Table 7 reports the accuracy, precision, and recall of the system on each test set. We calculated numbers in each A and B column by assuming each annotator's answers separately as a gold standard.

Table 7: System results on Complaint.com reviews (A, B: The first and the second annotator of each set)

	Testset 1		**Testset 2**		**Testset 3**		Avg
	A	B	A	B	A	B	
Acc(%)	65.8	63.0	67.6	61.0	**77.6**	**72.9**	68.0
Prec(%)	50.0	60.7	68.6	62.9	67.9	60.7	61.8
Recl(%)	56.0	51.5	51.1	44.0	**65.5**	**58.6**	54.5

In Table 7, accuracies indicate the agreement between the system and human annotators. The average accuracy 68.0% is comparable with the pair-wise human agreement 82.1% even if there is still a lot of room for improvement[9]. It was interesting to see that Testset 3, which was from restaurant complaints, achieved higher accuracy and recall than the other test sets from mp3 player complaints, suggesting that it would be interesting to further investigate the performance of reason identification in various other review domains such as travel and beauty products in future work. Also, even though we were somewhat able to measure reason sentence identification in complaint reviews, we agree that we need more data annotation for more precise evaluation.

Finally, the followings are examples of sentences that our system identified as reasons of complaints.

(1) Unfortunately, I find that I am no longer comfortable in your establishment because of the unprofessional, rude, obnoxious, and unsanitary treatment from the employees.

(2) They never get my order right the first time and what really disgusts me is how they handle the food.

(3) The kids play area at Braum's in The Colony, Texas is very dirty.

(4) The only complaint that I have is that the French fries are usually cold.

(5) The cashier there had short changed me on the payment of my bill.

As we can see from the examples, our system was able to detect con sentences which contained opinion-bearing expressions such as in (1), (2), and (3) as well as reason sentences that mostly described mere facts as in (4) and (5).

6 Conclusions and Future work

This paper proposes a framework for identifying one of the critical elements of online product reviews to answer the question, "What are reasons that the author of a review likes or dislikes the product?" We believe that pro and con sentences in reviews can be answers for this question. We present a novel technique that automatically labels a large set of pro and con sentences in online reviews using clue phrases for pros and cons in epinions.com in order to train our system. We applied it to label sentences both on epinions.com and complaints.com. To investigate the reliability of our system, we tested it on two extremely different review domains, mp3 player reviews and restaurant reviews. Our system with the best feature selection performs 71% F-score in the reason identification task and 61% F-score in the reason classification task.

[7] The kappa value was 0.63.

[8] In complaints reviews, we believe that it is more important to identify reason sentences than to classify because most reasons in complaints are likely to be cons.

[9] The baseline system which assigned the majority class to each sentence achieved 59.9% of average accuracy.

The experimental results further show that pro and con sentences are a mixture of opinions and facts, making identifying them in online reviews a distinct problem from opinion sentence identification. Finally, we also apply the resulting system to another review data in complaints.com in order to analyze reasons of consumers' complaints.

In the future, we plan to extend our pro and con identification system on other sorts of opinion texts, such as debates about political and social agenda that we can find on blogs or news group discussions, to analyze why people support a specific agenda and why people are against it.

Reference

Berger, Adam L., Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing, *Computational Linguistics*, (22-1).

Bethard, Steven, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic Extraction of Opinion Propositions and their Holders, *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*.

Chklovski, Timothy. 2006. Deriving Quantitative Overviews of Free Text Assessments on the Web. *Proceedings of 2006 International Conference on Intelligent User Interfaces (IUI06).* Sydney, Australia.

Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. 2005. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. *Proceedings of HLT/EMNLP-05.*

Esuli, Andrea and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. *Proceedings of CIKM-05, 14th ACM International Conference on Information and Knowledge Management*, Bremen, DE, pp. 617-624.

Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. *Proceedings of 35th Annual Meeting of the Assoc. for Computational Linguistics* (ACL-97): 174-181

Hatzivassiloglou, Vasileios and Janyce Wiebe. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *Proceedings of International Conference on Computational Linguistics (COLING-2000).* Saarbrücken, Germany.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews". *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004)*, Seattle, Washington, USA.

Kim, Soo-Min and Eduard Hovy. 2004. Determining the Sentiment of Opinions. *Proceedings of COLING-04*. pp. 1367-1373. Geneva, Switzerland.

Kim, Soo-Min and Eduard Hovy. 2005. Automatic Detection of Opinion Bearing Words and Sentences. *In the Companion Volume of the Proceedings of IJCNLP-05, Jeju Island, Republic of Korea.*

Kim, Soo-Min and Eduard Hovy. 2006. Identifying and Analyzing Judgment Opinions. *Proceedings of HLT/NAACL-2006, New York City, NY.*

Lin, Chin-Yew and Eduard Hovy. 1997. Identifying Topics by Position. *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP97).* Washington, D.C.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques*, Proceedings of EMNLP 2002*.

Popescu, Ana-Maria, and Oren Etzioni. 2005. Extracting Product Features and Opinions from Reviews , *Proceedings of HLT-EMNLP 2005*.

Riloff, Ellen, Janyce Wiebe, and Theresa Wilson. 2003. Learning Subjective Nouns Using Extraction Pattern Bootstrapping. *Proceedings of Seventh Conference on Natural Language Learning (CoNLL-03). ACL SIGNLL.* Pages 25-32.

Turney, Peter D. 2002. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of ACL-02*, Philadelphia, Pennsylvania, 417-424

Wiebe, Janyce M., Bruce, Rebecca F., and O'Hara, Thomas P. 1999. Development and use of a gold standard data set for subjectivity classifications. *Proceedings of ACL-99.* University of Maryland, June, pp. 246-253.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of HLT/EMNLP 2005*, Vancouver, Canada

Wilson, Theresa, Janyce Wiebe, and Rebecca Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. *Proceedings of 19th National Conference on Artificial Intelligence* (AAAI-2004).