Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification

Ji Fang
Intelligent Systems Laboratory
Palo Alto Research Center
Palo Alto, California 94304, USA
`fang@parc.com`

Bi Chen
College of Information Sciences and Technology
Pennsylvania State University
Pennsylvania, USA
`cb.chenbi@gmail.com`

Abstract

Two typical approaches to sentiment analysis are lexicon look up and machine learning. Even though recent studies have shown that machine learning approaches in general outperform the lexicon look up approaches, completely ignoring the knowledge encoded in sentiment lexicons may not be optimal. We present an alternative method that incorporates sentiment lexicons as prior knowledge with machine learning approaches such as SVM to improve the accuracy of sentiment analysis. This paper also describes a method to automatically generate domain specific sentiment lexicons for this learning purpose. Our experiment results show that the domain specific lexicons we constructed lead to a significant accuracy improvement for our sentiment analysis task.

1 Introduction

Two typical approaches to sentiment analysis are lexicon look up and machine learning. A lexicon look up approach normally starts with a lexicon of positive and negative words. The overall sentiment of a text is determined by the sentiments of a group of words and expressions appearing in the text (Liu, 2007; Zhou and Chaovalit, 2008). However, a significant challenge to this approach is that the polarity of many words is domain and context dependent. For example, *long* is positive in *long battery life* and negative in *long shutter lag*. Such words are associated with sentiment in a particular domain, but are not subjective in nature. Nevertheless, current sentiment lexicons do not capture such domain and context sensitivities of sentiment expressions. They either exclude such expressions or tag them with an overall polarity tendency based on statistics gathered from

certain corpus. While excluding such expressions leads to poor coverage, simply tagging them with a polarity tendency leads to poor precision.

Because of these limitations, machine learning approaches have been gaining increasing popularity in the area of sentiment analysis (Pang et al., 2002; Gamon, 2004). A machine learning approach such as Support Vector Machine (SVM) does not rely on a sentiment lexicon to determine the polarity of words and expressions, and can automatically learn some of the context dependencies illustrated in the training data.

Although recent studies have shown that machine learning approaches in general outperform the lexicon look up approaches for the task of sentiment analysis (Pang et al., 2002), completely ignoring the advantages and knowledge provided by sentiment lexicons may not be optimal. We present an alternative method that incorporates sentiment lexicons as prior knowledge with machine learning approaches such as SVM to improve the accuracy of sentiment analysis. This paper also describes a method to automatically generate domain specific sentiment lexicons for this learning purpose. Our experiments show that compared to general purpose domain independent sentiment lexicons, the domain specific lexicons lead to more significant accuracy improvement.

The sentiment analysis task performed in this paper is a fine grained product aspect level sentiment classification task for camera reviews. Namely, for each sentence in the camera reviews, we need to predict whether this sentence discusses any camera aspects, and if so, what is the associated sentiment.

2 Related Work

Given the task and the approaches of this study, we review the related works from three areas: 1. product aspect level sentiment analysis; 2. combining lexicon-based and machine learning approaches

for sentiment analysis; 3. sentiment lexicon generation.

Product aspect level sentiment analysis aims to determine both the product aspects/features and their associated opinion at the sentence level. Earlier works include Hu and Liu (2004) and Popescu and Etzioni (2005). Both of these works extract frequent noun phrases as product aspects. Therefore, they do not identify implicitly expressed product aspects, and they do not further categorize the extracted noun phrases.

In our study, we extract both the explicitly and implicitly expressed product aspects, and we further categorize the semantically related aspects. Zhao et al. (2010)'s work is close to ours in this sense. However, in terms of opinion extraction, they only extract opinion words associated with product aspects, and they do not further identify the polarities of the opinion words. By contrast, we aim to identify the polarities associated with the product aspects. Our approach features incorporating lexicon information into machine learning. Thus we review studies that combine lexicon-based and machine learning approaches for sentiment analysis next.

In previous studies, the lexicon-based and machine learning approaches have been incorporated in two ways. The first way is to develop two weighted classifiers using these two approaches and then integrate them into one system. Andreevskaia and Bergler (2008)'s work falls into this category. The second way is to incorporate lexicon knowledge directly into learning algorithms. Our work falls into this category.

In the second category, Wilson et al. (2005), melville et al. (2009), Dang et al. (2010) and Sindhwani and Melville (2008) all use a general purpose sentiment dictionary to improve polarity classification. Our work differs from these previous studies in that we incorporate not only a general purpose sentiment lexicon but also *Domain Specific Sentiment Lexicons* into SVM learning, and we use this method for identifying both product aspects and their associated polarities. More importantly, our experiment results show that while a general purpose sentiment lexicon provides only minor accuracy improvement, incorporating domain specific dictionaries leads to more significant improvement.

Regarding the construction of sentiment lexicon, earlier studies have focused on generating general purpose dictionaries. These methods range from manual approaches (Wiebe et al., 2005) to semi-automated (Hu and Liu, 2004; Kim and Hovy, 2004; Zhuang and Jing, 2006) and automated approaches (Mohammad et al., 2009). More attention has been devoted to domain specific lexicon construction recently. For example, Fahrni and Klenner (2008) present a method to identify polarity adjectives specific to food targets extracted from wikipedia. Jijkoun et al. (2010) generate a topic-specific lexicon from a general purpose polarity lexicon. In this paper, we present a method to build domain specific sentiment lexicons from scratch using a combination of corpus filtering, web searching using linguistic patterns and dictionary expansion techniques. Among these techniques, web searching using linguistic patterns was first introduced by Hatzivassiloglou and Sebastiani (1997) to generate domain independent sentiment adjectives. Kobayashi et al. (2004) designed patterns to extract co-occurring aspect nouns and opinion adjectives. Fahrni and Klenner (2008) also used this technique and their lexicon is also limited to adjectives . By contrast, we use this technique to generate domain specific lexicon not limited to adjectives and nouns. Our method is described in detail below.

3 Generating Domain Specific Lexicons

As discussed above, the sentiments of many words or phrases are context or domain dependent. For example, *long* is positive if it is associated with the camera aspect of 'Battery Life'. However, the same word carries negative sentiment when it is associated with the camera aspect of 'Shutter Lag'. Therefore, it is critical to know the topic/domain being discussed when we try to determine the associated sentiment.

Based on this observation, we aim to build domain/topic specific lexicons covering both expressions indicating a specific domain and expressions indicating different sentiments associated with that particular domain. For example, our lexicon regarding 'Camera Picture Quality' would consist of two sub-lexicons. One includes words and phrases such as *picture, image, photo, close up* etc, which are good indicators for the topic of 'Picture Quality' in the area of digital cameras. The other one includes words and expressions that carry positive or negative sentiments if the associated topic is camera picture quality. For exam-

ple, this second sub-lexicon would indicate that while *sharp* and *clear* are positive, *blurry* is negative when they are associated with camera picture quality. We achieved our goal by using a combination of corpus filtering, web search with linguistic patterns and dictionary expansion. Each of these techniques is described in detail in the following subsections.

3.1 Corpus Filtering

We first use a training corpus, in which each camera review sentence is annotated with a camera aspect as well as the associated sentiment, to build a foundation for our domain specific lexicons. Our approach is as follows.

First, for each camera aspect such as *Durability*, we extract all of the content words and phrases that occur in the training sentences labelled as expressing that aspect. The content words and phrases we extracted include nouns, verbs, adjectives, adverbs as well as their negated forms. This step produces an initial list of lexicon for each camera aspect.

Second, for each word and phrase in the list for each of the camera aspects, we check to see if that word or phrase also occurs in any other camera aspect lexicon. If yes, we remove it from the lexicon. After this step of filtering, we obtained a list of lexicon for each camera aspect, which contains only words and phrases unique to that camera aspect.

The quality of the lexicons produced using this approach is in general very high. For example, the following lexicon regarding the camera *Durability* was generated based on our relatively small training corpus with 2131 sentences covering 23 categories (22 camera aspects and a category of 'none', meaning that none of the 22 camera aspects was discussed).

Durability Lexicon: *[scratch, construct, build, rock, repair, damage, flimsy, not flimsy, junk, sturdy, sturdier, solid, durable, tough, bent, hard, not worth, firm, rug, broke, bulletproof]*

However, the drawback of this approach is that the coverage of the lexicons would completely rely on the coverage of the corpus, and annotating a broad coverage training corpus is time consuming, expensive and sometimes very difficult for a task such as sentiment analysis because of the richness of natural language.

We overcome this drawback by augmenting the initial domain specific lexicons we obtained from

the training corpus through web search and filtering using linguistic patterns as well as dictionary expansion. These two approaches are illustrated in the next two subsections.

3.2 Web Search and Filtering Using Linguistic Patterns

To improve the coverage of the domain specific lexicons we obtained from our training corpus, we designed two linguistic patterns and used them as searching queries to find more words and phrases conceptually associated with the camera aspects. The two linguistic patterns we used are as follows.

Pattern 1: "Camera Aspect include(s) *"

Pattern 2: Camera Aspect + "Seed Word and *"

In these two patterns, 'Camera Aspect' refers to expressions such as *camera accessories* and *camera price*. 'Seed Word' refers to seed words for a particular camera aspect. For example, *cheap* and *expensive* can serve as seed words for camera aspect *price*. Note that in Pattern 1, the camera aspect name is included as part of an exact search query, whereas in Pattern 2, the camera aspect name serves as the context for the search query.

Depending on the semantic nature of a camera aspect, we choose one of these two patterns to find expressions conceptually related to that aspect. For example, while "camera accessories include *" is very effective for finding accessory expressions, 'camera picture + "clear and *"' is better for finding expressions related to camera pictures.

When we use Pattern 1, we send it as a query to a search engine such as Bing[1]. We then extract words following 'include' or 'includes' in the top 50 results returned by the search engine. In each returned result, we extract words that follow 'include' or 'includes' until we hit the sentence boundary. The final step is to remove common stop words such as *the* and function words such as *with* and *of* from the extracted words. As an example, the following lexicon for camera accessory is generated using this method.

Accessory Lexicon: *[chip, chips, case, bag, card, software, tripod, strap, cable, adapt, charger, port, storage, hood, connector, kit, accessory, glove, belt, usb, mic, beltloop, flash, program, leather, pack, connect, not belt, not strap, zipper]*

---

[1]In our experiments, we used Bing for convenience. However, our approach is applicable using other search engines such as Google as well.
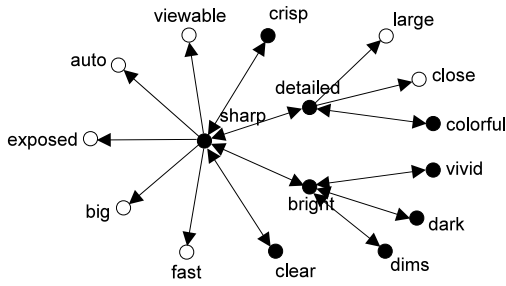
Figure 1: Noisy Words v.s. Non-noisy Words for Camera Picture Quality

When we use Pattern 2, we also extract words in the top 50 returned results. However, we adopt a different algorithm for filtering out noise in the returned results. For example, for finding expressions conceptually related to camera's picture quality, we use 'camera picture' as context words and 'clear' as a seed word. This pattern would match both 'clear and sharp' and 'clear and normal'. However, while 'sharp' is commonly used to describe picture quality, 'normal' is not. To filter noisy words such as 'normal', we use each of the candidate words as a new seed word in Pattern 2, and if the top 50 results returned by the new query include the original seed word 'clear', the candidate word is retained. Otherwise, it is discarded. For example, in our experiments, while 'camera picture + "sharp and *"' would return results matching 'sharp and clear', 'camera picture + "normal and *"' would not return results matching 'normal and clear'. Through this approach, we can distinguish 'sharp' from 'normal', and identify 'normal' as a noisy word. Figure 1 shows some of the noisy words identified by this approach when we extract expressions conceptually related to camera pictures. In this figure, words represented by hollow circles are identified as noise and removed from the camera picture quality lexicon. By contrast, words represented by solid circles are retained in our lexicon.

3.3 Dictionary Expansion

Although expansion through looking up synonyms and antonyms recorded in dictionaries is a commonly used approach when a general purpose sentiment lexicon is built (Hu and Liu, 2004), we found this approach to be not always suitable for building domain specific lexicons. The reason is that building domain specific lexicons requires finding expressions that are conceptually related;

however expressions that are conceptually related are not necessarily synonyms or antonyms. For example, 'sharp' and 'clear' are conceptually related to camera picture qualities, but they are not true synonyms from a linguistic perspective.

However, in some cases, using dictionaries can still be very effective. For example, we built the following lexicon for camera price through web searching and filtering using Pattern 2.

Price Lexicon: *[cheap, lowest, discount, promo, coupon, promote, expensive, worthy, value]*

By including the synonyms of 'cheap' and 'expensive' in WordNet (Fellbaum, 1998), we are able to further expand the Price Lexicon.

3.4 Domain Specific Polarity Lexicon

So far we have described how we build domain specific lexicons for different camera aspects. The next step is to separate expressions that carry positive sentiment from those that carry negative sentiment in each domain lexicon.

For example, we want to be able to build the following sub-lexicons for 'Picture Quality'.

PictureQuality Positive Lexicon: *[clear, sharp, bright, sober, stable, tidy, vivid, sunny, crisp]*

PictureQuality Negative Lexicon: *[dark, dim, humid, fuzzy, gray, blurry, blur, indistinct, grainy, hazy, blurred]*

Our approach is as follows. For each expression in the Picture Quality Lexicon that we constructed through the combination of corpus filtering, web search and dictionary expansion, we check to see if it only appears in the training data labelled as expressing a positive opinion or a negative opinion about the camera's picture quality. If it is the former case, we include that expression into the PictureQuality Positive Lexicon, while if it is the latter case, we include that expression into the PictureQuality Negative Lexicon.

Having illustrated our approach for constructing domain specific sentiment lexicons, we next describe how we incorporate lexicon knowledge into SVM learning to improve sentiment classification.

4 Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification

Our sentiment classification task is as follows. For each review sentence about cameras, we need to predict both the camera aspect discussed in that sentence as well as the associated sentiment re-

garding that camera aspect. We achieve this goal by performing a two step classification. In step 1, we train a classifier to predict the camera aspect being discussed. In step 2, we train a classifier to predict the sentiment associated with that camera aspect. Finally, we aggregate the two step prediction results together to produce the final prediction.

In both steps, we incorporate the lexicon knowledge into conventional SVM learning. To illustrate our approach, we use sentence (1) as an example.

(1) *The case is rigid so it gives the camera extra nice protection.*

Using nouns, verbs, adjectives and adverbs as unigram feature words in a conventional SVM learning, this sentence can be represented as the following vector of words.

[case, rigid, give, camera, extra, nice, protection]

By incorporating the knowledge encoded in the lexicons, we automatically generate and insert additional features into the above representation.

For example, when we perform the step 1 aspect classification, because the feature word 'case' in the above representation is listed in our domain specific lexicon about camera accessories, we would insert an additional feature word 'accessory', and produce the following new representation.

[case, rigid, give, camera, extra, nice, protection, accessory]

By doing this, we promote the possibility of the camera aspect being 'accessory' if expressions of camera aspects occur in the sentence.

In the next step of polarity prediction, we incorporate both our domain specific sentiment lexicon and a general purpose domain independent sentiment lexicon extracted from the MPQA opinion corpus (Wiebe et al., 2005) [2].

For example, because 'nice' is indicated as a positive word in the MPQA lexicon, we would insert a feature word 'positive'. In addition, if the first step prediction result for sentence (1) is 'accessory', and 'rigid' is also a positive word in our domain specific lexicon regarding camera accessories, we would generate an extra feature word 'positive' in our final representation for sentence (1) for the second step polarity prediction as

_____
[2]We only extracted the words that are indicated as strongly subjective out of context from the MPQA opinion corpus

shown below.

[case, rigid, give, camera, extra, nice, protection, positive, positive]

We thus promote a 'positive' prediction regarding the aspect of 'accessory'.

Our experiments show that incorporating lexicon knowledge into SVM learning significantly improves the accuracy for our classification task; compared to the general purpose MPQA sentiment lexicon, the domain specific lexicon we constructed is more effective. Our experiment setting and results are reported in the next section.

5 Experiment Setting and Results

The sentiment analysis task we performed is a combined 45-way sentiment classification task. These 45 classes are derived from 22 aspects related to camera purchases such as *picture quality, LCD screen, battery life and customer support* and their associated polarity values *positive* and *negative*, as well as a class of *no opinion* about any of the 22 aspects. An example of such a class is *picture quality: positive*. The goal is to map each input sentence into one of the 45 classes.

As mentioned in the previous section, we performed a two step classification for our task. Namely, our final combined classifier consists of two classifiers. The first is an 'Aspect Classifier', which performs a 23-way camera aspect classification. The second is a 'Polarity Classifier', which performs a 3-way (*positive*, *negative* and *none*) classification. The final predictions are aggregated from the predictions produced by these two classifiers.

The classification accuracy is defined as follows.

$$Accuracy = \frac{Number\,of\,Sentences\,Correctly\,Classified}{Total\,Number\,of\,Sentences}.$$
(1)

In our experiment we labeled 2718 sentences randomly chosen from the Multi-Domain Sentiment Dataset created by Blitzer et al. (Blitzer et al., 2007); therefore, the classes in this data set are not balanced, and the majority class has 13% of the sentences.

As mentioned in the Related Work section, our task is different from those of the early studies on product aspect level sentiment analysis. Earlier works such as Hu and Liu (2004) and Popescu and Etzioni (2005) only extract explicitly expressed product aspects, and they do not identify implicitly

98

expressed product aspects. In addition, they do not further categorize the extracted noun phrases. By contrast, we need to extract both the explicitly and implicitly expressed product aspects and further categorize the semantically related expressions regarding product aspects. Zhao et al. (2010)'s work did extract both explicitly and implicitly mentioned product aspects, and they also further categorized the product aspects. However, in terms of opinion extraction, they only extracted opinion words associated with product aspects, and did not further identify the polarities of the opinion words. By contrast, we need to identify the polarities associated with the product aspects. Therefore, we cannot compare our results directly with those presented in the earlier works. Instead, we used the majority class (13%) as our baseline, and we compared our approach to incorporating lexicon knowledge with SVM learning mainly with a conventional SVM learning, because the latter is the state-of-the-art algorithm reported in the literature for sentiment analysis. Our results show that both the conventional SVM learning and our approach significantly outperform the majority class baseline.

We selected the Nouns, Verbs, Adjectives and Adverbs as our unigram word features. All of them are stemmed using the Porter Stemmer (Rijsbergen et al., 1980). Negators are attached to the next selected feature word. We also use a small set of stop words[3] to exclude copulas and words such as *take*. The reason that we choose these words as stop words is because they are both frequent and ambiguous and thus tend to have a negative impact on the classifier. The SVM algorithm we adopted is implemented by Chang and Lin (2001). We use linear kernel type and use the default setting for all other parameters.

We conducted 4 experiments. In experiment 1, we used the conventional SVM algorithm, in which no lexicon knowledge was incorporated; we refer to this experiment as SVM. In experiment 2, we incorporated only the knowledge encoded in the domain independent MPQA opinion dictionary into SVM learning; we refer to this experiment as 'MPQA + SVM'. In experiment 3, we incorporated only the knowledge encoded in the domain specific lexicons we constructed into SVM learning; we refer to this experiment as 'Domain-

---

[3]The stop words we use include copulas and the following words: *take, takes, make, makes, just, still, even, too, much, enough, back, again, far, same*

Lexicons + SVM'. In experiment 4, we incorporated both the knowledge encoded in the MPQA and the domain specific lexicons we constructed into SVM learning; we refer to this experiment as 'DomainLexicons + MPQA + SVM'. All of our results are based on 10-fold cross-validation, and they are summarized in Table 1.

The results in Table 1 show that incorporating both the domain independent MPQA lexicon and the domain specific lexicons that we built achieves the best overall performance. Of these two types of lexicon, incorporating the domain specific lexicons is more effective, as they contributed the most to the improvement of the classification accuracy. The improvement achieved by our approach is statistically significant with $p < 0.000001$ according to paired t-test.

Learning Method	Accuracy
SVM	41.7%
MPQA + SVM	44.3%
DomainLexicons + SVM	46.2%
DomainLexicons + MPQA + SVM	**47.4%**

Table 1: Overall Performance Comparison

Our results reported in Table 2 further illustrate that incorporating lexicon knowledge with SVM learning significantly improves both the accuracy for camera aspect classification and the accuracy for polarity classification. Both improvements are statistically significant with $p < 0.000001$ and $p < 0.05$ respectively according to paired t-test.

Learning Method	Aspect Accuracy	Polarity Accuracy
SVM	47.1%	65.6%
DomainLexicons + MPQA + SVM	**56.2%**	**66.8%**

Table 2: Breakdown Performance Comparison

6 Conclusions

To summarize, we have shown that incorporating the knowledge encoded in sentiment lexicons, especially domain specific lexicons, can significantly improve the accuracy for fine-grained sentiment analysis tasks. We have also described how we constructed our domain specific sentiment lexicons for the domain of camera reviews through a combination of corpus filtering, web searching and filtering and dictionary expansion. In addition, we have developed a method to incorporate the lexicon knowledge into machine learning algorithms such as SVM to improve sentiment learning.

References

A. Andreevskaia and S. Bergler. 2008. When specialists and generalists work together: overcoming domian dependence in sentiment tagging. In *Proceedings of ACL*.

John Blitzer, Mark Dredze, Fernando Pereira. Biographies, Bollywood, Boom-boxes, and Blenders. 2007. Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

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

A. Fahrni and M. Klenner. 2008. Old wine or warm beer: target-specific sentiment analysis adjectives. In *Symposion on Affective Language in Human and Machine, AISB Convention*.

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

Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the COLING*.

V. Hatzivassiloglou and F. Sebastiani. 1997. Predicting the semantic orientation of adjectives. In *Proceedings of ACL*.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the KDD*.

V. Jijkoun, M.d. Rijke, and W. Weerkamp. 2010. Generating focused topic-specific sentiment lexicon. In *Proceedings of ACL*.

S.-M. Kim and E. Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the COLING*.

N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of IJCNLP*.

Bing Liu. 2007. *Web Data Mining*. Springer, New York.

Prem melville, Wojciech Gryc, and Richard D. Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the KDD*.

Saif Mohammad, Cody Dunne, and Bonnie Dorr. 2009. Generating high-coverage semantic orientation lexicons from overtly marked words and a theasaurus. In *Proceedings of the EMNLP*.

Bo Pang, Lillian lee, and shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the EMNLP*.

A.M. Popescu and O. Etzioni. 2005. Extracting product features and opinions from reviews. In *Proceedings of HLT/EMNLP*.

C.J. Van Rijsbergen, S.E. Robertson, and M.F. Porter. 1980. New models in probabilistic information retrieval. Technical report, British Library Research and Development Report, no. 5587.

Vikas Sindhwani and Prem Melville. 2008. Document-word co-regularization for semi-supervised sentiment analysis. In *Proceedings of the ICDM*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the HLT-EMNLP*.

W.X. Zhao, J. Jiang, H. Yan, and X. Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP*.

L. Zhou and P. Chaovalit. 2008. Ontology-supported polarity mining. *Journal of the American Society for Information Science and technology*, 59.

L. Zhuang and F. Jing. 2006. Movie review mining and summarization. In *Proceedings of the CIKM*.