Exploring Weakly Supervised Latent Sentiment Explanations for Aspect-level Review Analysis

Lei Fang, Minlie Huang and Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology
Tsinghua University, Beijing 100084, PR China
fang-l10@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn,
zxy-dcs@tsinghua.edu.cn

ABSTRACT

In sentiment analysis, aspect-level review analysis has been an important task because it can catalogue, aggregate, or summarize various opinions according to a product's properties. In this paper, we explore a new concept for aspect-level review analysis, *latent sentiment explanations*, which are defined as a set of informative aspect-specific sentences whose polarities are consistent with that of the review. In other words, *sentiment explanations* best represent a review in terms of both aspect and polarity. We formulate the problem as a structure learning problem, and sentiment explanations are modeled with latent variables. Training samples are automatically identified through a set of pre-defined aspect signature terms (i.e., without manual annotation on samples), which we term the way *weakly supervised*.

Our major contributions lie in two folds: first, we formalize the use of aspect signature terms as weak supervision in a structural learning framework, which remarkably promotes aspect-level analysis; second, the performance of aspect analysis and document-level sentiment classification are mutually enhanced through joint modeling. The proposed method is evaluated on restaurant and hotel reviews respectively, and experimental results demonstrate promising performance in both document-level and aspect-level sentiment analysis.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural language processing—*Text Analysis*

Keywords

Opinion Mining; Sentiment Classification; Sentiment Analysis; Structural Learning; Text Mining

1. INTRODUCTION

The booming web gives an enormous impetus to the prosperity of online customer reviews. Such content tends to become a major resource from which users may find opinions or comments on the products or services they want to consume. However, users sometimes might be overwhelmed, and not be able to read reviews one by one when facing a considerably large number of reviews. Users may be not satisfied with numerical review statistics since textual opinions are more helpful. To address this issue, aspect-level review analysis may be a better option.

Recently, aspect-level review analysis has shown advantages over traditional document-level sentiment analysis[8, 9, 14, 31, 33, 38]. Many existing research works[6, 7, 20, 22] addressed the task of aspect extraction, considering it as a prerequisite for aspect-based sentiment analysis. Generally, the aspect extraction process starts from some given seed words for each aspect. Then, words that either have strong associations with the extracted ones or satisfy certain predefined rules are selected. Many other approaches[2, 10, 20, 12] that extend the topic models[1] are also widely studied. The seed words are sometimes termed *aspect signature terms*, which can be obtained by some simple methods with a small amount of manual annotation. Thus, aspect signature terms makes it easy to scale to other domains or expand to new aspects when new products or brands are introduced. For example, the word set {"value", "price", "cost", "worth"} is a set of signature terms for aspect "price" in hotel or restaurant review, while {"storyline", "story", "tale", "script", "storyteller"} signify the aspect "story" in movie review.

However, it should be noted that those given aspect signature terms are not fully utilized in these approaches. Such prior knowledge is only employed for model initialization: aspect seed initialization or prior distribution for latent topics. As prior knowledge has long been shown to play an important role on human brain in understanding the world[30], many research works in data mining or machine learning attempt to promoting the performance by incorporating prior knowledge. For instance, a variety of approaches have been proposed to encode prior knowledge into support vector machines[3, 5, 11, 34] and showed remarkable performance improvement. In this paper, we address the problem of aspect-level sentiment analysis by making full use of these aspect signature terms.

We give an exemplar hotel review in Figure 1. It mentions several aspects including room, service, food, and price. Each aspect covers several sentences. As can be seen, the reviewer enjoyed the room and service of the hotel, but complained about the food, and gave a negative overall rating to the hotel. It's worth noting that, though there are positive opinions on room and service, the major aspect (food) that the reviewer complained about leads to a nega-

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505538.

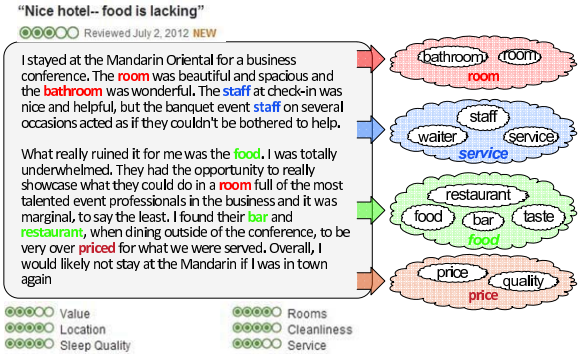

Figure 1: A sample hotel review

tive overall rating. We are inspired by the simple observation and propose the following conjectures:

- Sentences that are associated with aspects might be the *sentiment explanations* in predicting document-level polarity while other sentences that are not coherent with the overall rating may mislead the classifier.
- Finding aspect-specific information that is coherent with the overall rating would be more convincible and useful for aspect-level review analysis.

In addition, statistics on more labeled data also reveals that **aspect-associated sentences** may act as ***sentiment explanations***. We manually annotated about 450 restaurant reviews containing 4,405 sentences with 6 predefined aspects such as taste, ambience, etc. (see details in the experiment section). Each sentence of the review is labeled with aspect and polarity label. Figure 2 presents the aspect distribution of positive and negative reviews. Positive reviews mention much more about taste than negative reviews (22.2% vs. 12.3%), which may imply that taste is a major factor for giving a positive overall rating. In comparison, service is mentioned much more frequently in negative reviews than in positive ones (21.1% vs. 5.5%). This gives the signal that restaurants with delicious food may receive more positive reviews, while a poor service may lead to more negative reviews.

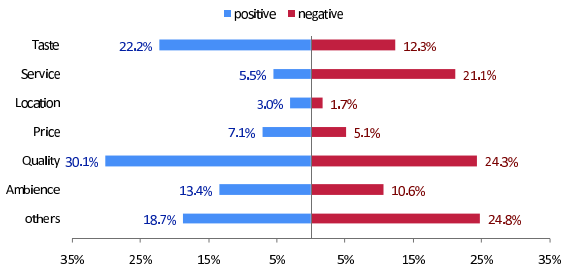

Figure 2: Different aspect distributions in positive and negative reviews

Furthermore, users are writing reviews to praise or criticize on some specific aspects about which they care, which results in a high correlation between aspects and opinions. Table 1 presents some statistics on the labeled data. Among sentences that are labeled with positive or negative polarity, 93.48% ($=\frac{58.64\%}{58.64\%+4.13\%}$) are tagged with aspect label. The statistics explains that aspect-level analysis shall be performed on the polarity part of the review, which is neglected in prior studies. However, it is difficult to obtain data with aspect labels to perform aspect-level analysis since aspect annotation is cost and time expensive.

#.(sentences)	with polarity	without polarity
with aspect	**58.64%**	19.59%
without aspect	4.13%	17.64%

Table 1: Correlation between aspect and opinion (restaurant review)

To address the aforementioned issues, we present a structural learning model for jointly performing aspect- and document-level sentiment analysis. The major departure from prior studies is two-fold:

- We formalize the use of a handful set of aspect signature terms as weak supervision in a structural learning framework. And in order to avoid heavy manual annotation, aspect assignment for training samples is identified automatically with these signature terms.
- Aspect analysis and document-level polarity prediction are modeled jointly (see Section 3.4), which endows the model the capability of jointly predicting the document-level sentiment polarity and extracting aspect-specific sentiment explanations, with mutually improved performance.

The proposed method is evaluated on restaurant and hotel reviews respectively, and experimental results show remarkable improvements both in document-level sentiment classification and in aspect analysis. The remainder of this paper is organized as follows: in Section 2, we briefly introduce related work on multi-level sentiment classification and aspect analysis. In Section 3 we present the formulation of the model. In Section 4 we discuss the experimental settings and results. Finally, we summarize our work in the last section.

2. RELATED WORK

There has been much work focused on multi-level sentiment classification. For document-level and sentence-level sentiment analysis, Mao and Lebanon[16] extended the standard conditional random fields to model the sequential flow of sentiment throughout the document. They also demonstrated that it is useful to employ local sequential sentiment representation for document-level sentiment analysis. McDonald *et al.*[17] proposed a sentence-document model to perform fine-to-coarse sentiment analysis which aims to jointly classify sentiment on multiple levels of granularity. Also, they treated the inference of the sentence-level sentiment as a sequential labeling problem. Täckström and McDonald[29] proposed to discover fine-grained sentiment with hidden-state CRF[23] only by the document-level coarse-grained supervision. Yessenalina *et al.*[35] deployed the framework of latent structural SVMs[36] for multilevel sentiment classification jointly. They both treated the sentence-level sentiment as latent variables, which is trained in a structural learning model. Many of the previous works[4, 16, 17, 29, 35] claimed that document-level sentiment analysis can benefit from finer level classification.

Many works promoted the performance of sentiment analysis by incorporating prior knowledge as weak supervision. Li and Zhang[13] introduced lexical prior knowledge to non-negative matrix tri-factorization. Shen and Li[26] further extended the matrix factorization framework to model dual supervision from document and word labels. Melville *et al.*[18] proposed a generative background model to leverage lexical information in terms of word la-

bels. Silva *et al*.[27] proposed a self-augmentation training procedure incorporating sentiment rules which can be easily obtained by projecting the training data for effective sentiment stream analysis.

There is also much work on aspect-level analysis, such as aspect rating, ranking, extraction, or summarization. Hu and Liu[8] applied frequent itemset mining to extract product feature. Adjectives that are close to feature words are considered as opinion words. Reviews are then summarized according to product feature. Qiu *et al*.[22] proposed to iteratively extract aspect feature words and opinion words with predefined rules using "Double Propagation". Hai *et al*.[7] proposed to incorporate statistical association analysis in a bootstrapping framework to mine aspect features. Snyder and Barzilay[28] employed the good grief algorithm for multiple aspect ranking. Their algorithm jointly learns ranking models for individual aspects by modeling the dependencies between assigned ranks.

Besides, various extensions of generative topic models are also widely studied for aspect analysis[2, 10, 12, 15, 19, 20]. Titov and McDonald[31] proposed a multigrain topic model to discover local rateable aspects. Wang *et al*.[33] proposed to predict aspect rating using the generative probabilistic model. Zhao *et al*.[38] employed MaxEnt-LDA for jointly modeling aspects and opinions. By modeling the aspect-document structure and document generative process, they all[31, 33, 38] used the mined aspect-specific knowledge for further aspect analysis.

However, previous research works do not fully utilize aspect-specific supervision (aspect signature terms). In this paper, we leverage such weak aspect-specific supervision to extract sentiment explanations for both document-level sentiment classification and aspect-oriented review analysis.

3. A STRUCTURAL LEARNING MODEL EXPLORING LATENT SENTIMENT EXPLANATION

3.1 Definitions: Aspect and Sentiment Explanation

We define *aspect* as a set of signature terms that signify the occurrence of a product's property. For example, {"storyline", "story", "tale", "script", "storyteller"} defines the aspect "story" for movie review. A sentence is considered as a *sentiment explanation* if it is describing a certain aspect and its polarity is coherent with that of the review.

We assume that each sentiment explanation contributes to the overall rating of a review. In this paper, we further assume that each sentiment explanation is only associated with one aspect. This is practical in our problem as we can extract sub-sentences by separating the review document with punctuation marks such as semicolon, period, exclamation point, or interrogation mark. Without ambiguity, we term the separated text segments as "*sentence*".

A review document mainly consists of two parts: opinion part and non-opinion part. Each sentence of the opinion part characterizes a certain aspect or describes some opinion, while the non-opinion part is about the background or factual information. Among the opinion part, we only consider the *sentiment explanations* of the review, i.e., those sentences whose polarity are consistent with the overall polarity. In other words, sentiment explanations are the informative part that best represent the original review in both aspect and polarity. Thus, aspect analysis shall be performed on the sentiment explanations of a review. We believe that this would make the mined opinions more coherent, representative, and meaningful.

3.2 Problem Formulation

Let document be denoted by x, $y \in \{+1, -1\}$ represents the positive/negative polarity of the document, and $\mathcal{H}$ is the set of informative sentences representing the *sentiment explanations*, in which each sentence is attached with a certain aspect $a_i \in A = \{a_1, ..., a_k\}$. The task here, is to learn a function $\mathcal{F}(x, (y, \mathcal{H}))$ that jointly models the document polarity and the aspect assignment of the sentiment explanations, as follows:

$$(y^*, \mathcal{H}^*) = \underset{y \in \{+1,-1\}, \mathcal{H} \in \mathcal{P}(x)}{\operatorname{argmax}} \mathcal{F}(x, (y, \mathcal{H}))$$

where $\mathcal{P}(x)$ is the power set of all the sentences in x, and each sentence in $\mathcal{H}$ is predicted with an aspect label. Let x^j denote the j-th sentence of document x, and a^j is the attached aspect of x^j. Note that document-level polarity is the only supervision we used while aspect-level annotation on sentence is not required.

3.3 Loss Functions Encoding Aspect Information

In our model, we expect that each sentence in $\mathcal{H}$ characterizes one specific aspect. It should be noticed that for each aspect, we have a set of signature terms, which is critical in choosing an aspect if discriminate models are employed. For example, if "cost" is observed in a sentence, it is highly probable that it is talking about the "price" aspect.

To incorporate such weak supervision, we propose two types of loss functions for $\mathcal{H}$: sentence-level loss and document-level loss.

Sentence-level loss (SL)

The gold-standard aspect label of sentence x^j is $\hat{a}_{x^j}$ if the sentence contains signature terms of the accordant aspect [1], while the predicted aspect is a_{x^j}. This type of local loss measures the difference of aspect between the predicted aspect a_{x^j} and the reference aspect $\hat{a}_{x^j}$, with respect to a subset $\mathcal{H}_A$ of $\mathcal{H}$. It is denoted by $\Delta_{SL}(\hat{a}_{x^j}, a_{x^j})$ as follows:

$$\Delta_{SL} = \begin{cases} \frac{1}{|\mathcal{H}_A|} \sum_{x^j \in \mathcal{H}_\mathcal{A}} \Delta(\hat{a}_{x^j}, a_{x^j}) & \text{if} |\mathcal{H}_A| > 0, \\ 0 & \text{otherwise} \end{cases}$$

where $\mathcal{H}_A$ contains all the sentences automatically identified by aspect signature terms in $\mathcal{H}$.

Sentence-level loss is proposed to captures the local aspect feature, with the assumption that if sentence x^j contains the signature terms of aspect a, the aspect label of x^j is a.

Document-level loss (DL)

From the global perspective, we have a set of aspects $\hat{A}_x$ that for $a \in \hat{A}_x$, the document contains at least one aspect signature term for a. Once again, $\hat{A}_x$ can be obtained by a simple dictionary lookup process where the dictionary is the signature terms. The document-level loss measures the difference between the predicted aspect set A_x and the reference aspect set $\hat{A}_x$, as follows:

$$\Delta_{DL} = \begin{cases} 1 - \frac{|\hat{A}_x \bigcap A_x|}{|\hat{A}_x \bigcup A_x|} & \text{if} \hat{A}_x \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Document-level loss ensures that the predicted aspect set is not "far from" the aspect set obtained by signature terms.

3.4 Joint Modeling Aspect and Document Polarity

For the decision function $\mathcal{F}(x, (y, \mathcal{H}))$, we define the asymmetric loss function as

$$\Delta(\hat{y}, y, \mathcal{H}) = \Delta(\hat{y}, y) + \alpha \Delta_{SL} + \beta \Delta_{DL}$$

[1]The gold-standard is obtained by an automatic dictionary lookup process; that is why we call it *weak supervision*.

where α, β are the coefficients that balance the aspect loss and document-level sentiment classification loss, and $\hat{y}$ represents the gold-standard polarity of a review document.

Similar to Structural SVM [32], let $\Psi(x, y, \mathcal{H})$ denote the joint feature map that outputs the features describing the quality of predicting sentiment y using the sentence set $\mathcal{H}$. In order to obtain a model that is jointly trained, and that satisfies the condition that the overall polarity of document should influence the sentiment of extracted informative sentences, the document polarity shall also be encoded in $\Psi(x, y, \mathcal{H})$. In spirit to Yessenalina *et.al.*[35], we propose the following formulation of the discriminate function

$$\begin{aligned}\mathcal{F}(x,(y,\mathcal{H})) &= \vec{w}^T\Psi(x,y,\mathcal{H}) \\ &= \frac{1}{N(x)}\sum_{j\in\mathcal{H}}\left(y\cdot\vec{w}^T_{pol_{a^j}}\psi_{pol}(x^j)+\vec{w}^T_{subj_{a^j}}\psi_{subj}(x^j)\right) \\ &+ y\cdot\vec{w}^T_{doc}\psi_{pol}(x)\end{aligned}$$

where $N(x)$ is the normalizing factor, $\psi_{pol}(x^j)$ and $\psi_{subj}(x^j)$ represents the polarity and subjectivity features of sentence x^j respectively. $\vec{w}_{doc}$ denotes the weight vector modeling the overall polarity. $\vec{w}_{pol}$ and $\vec{w}_{subj}$ denote the weight for polarity and subjectivity features, respectively. More specifically, $\vec{w}_{pol_a}$ and $\vec{w}_{subj_a}$ represent the vectors of feature weight for aspect a to calculate the polarity and subjectivity score, respectively. That is, $\vec{w}_{pol_a}$ and $\vec{w}_{subj_a}$ are weight matrix and have the form as follows

$$\vec{w}_{pol} = \begin{bmatrix}\vec{w}^T_{pol_{a_0}}\\ \vdots \\ \vec{w}^T_{pol_{a_k}}\end{bmatrix}, \quad \vec{w}_{subj} = \begin{bmatrix}\vec{w}^T_{subj_{a_0}}\\ \vdots \\ \vec{w}^T_{subj_{a_k}}\end{bmatrix}$$

Document Polarity Prediction

To predict document polarity, we have the document-level sentiment classifier as

$$y^* = \underset{y\in\{+1,-1\}}{\operatorname{argmax}}\{\max_{\mathcal{H}\in\mathcal{P}(x)}\vec{w}^T\Psi(x,y,\mathcal{H})\} \quad (1)$$

In the experiment, we tune the size of $\mathcal{H}$ with respect to the number of sentences in x to obtain the optimal performance.

Aspect Assignment of Extracted Latent Sentiment Explanation

For each sentence x^j, we compute the joint subjectivity and polarity score with respect to aspect a and label y as

$$score(x^j,(a,y)) = y\cdot\vec{w}^T_{pol_a}\psi_{pol}(x^j)+\vec{w}^T_{subj_a}\psi_{subj}(x^j)$$

we then assign aspect a^j to sentence x^j if

$$a^j = \underset{a\in A}{\operatorname{argmax}}\{score(x^j,(a,y))\}$$

After sorting $score(x^j,(a^j,y))$ in decreasing order and taking summation by selecting the top $|\mathcal{H}|$ sentences (or fewer, if there are fewer than $|\mathcal{H}|$ that have positive joint score) as the total score for each $y\in\{+1,-1\}$, we then predict y with the higher joint scores as the sentiment of the whole document.

3.5 Model Training

3.5.1 Optimization Problem

With the problem formulation in previous section, the solution is to solve an optimization problem as follows:

OP_1 **:**

$$\begin{aligned}&\min_{\vec{w},\xi\geq 0}\frac{1}{2}||w||^2+\frac{C}{N}\sum_{i=1}^{N}\xi_i \\ &\quad s.t.\forall i: \\ &\max_{\mathcal{H}_i\in\mathcal{P}(x)}\vec{w}^T\Psi(x_i,y_i,\mathcal{H}_i)\geq\max_{\mathcal{H}'_i\in\mathcal{P}(x)}\vec{w}^T\Psi(x_i,-y_i,\mathcal{H}'_i) \\ &\qquad\qquad +\Delta(y_i,-y_i,\mathcal{H}'_i)-\xi_i\end{aligned}$$

As OP_1 is non-convex, we employ the framework of structural SVMs with latent variables[36]using CCCP algorithm [37]. According to the formulation, the true informative sentence set (sentiment explanation) is never observed, and thus is modeled as a hidden or latent variable. Thus, we keep $\mathcal{H}_i$ fixed to compute the upper bound for the concave part of each constraint, and rewrite the constraints as

$$\xi_i\geq\max_{\mathcal{H}'_i\in\mathcal{P}(x)}\vec{w}^T\Psi(x_i,-y_i,\mathcal{H}'_i)-\vec{w}^T\Psi(x_i,y_i,\mathcal{H}_i)+\Delta(y_i,-y_i,\mathcal{H}'_i)$$

After that, we have y_i completed with the latent variable $\mathcal{H}_i$ as if it is observed. For each training example, starting with an initialized sentence set in which each sentence is tagged with an aspect label, the training procedure alternates between solving an instance of the structural SVM using the $\mathcal{H}_i$ and predicting a new sentence set until the learned weight vector $\vec{w}$ converges.

In our work, we use bag-of-words features, and use the performance on a validation set to trigger the halting condition, which is a commonly adopted strategy when an optimization problem is non-convex.

3.5.2 Model Initialization

The normalizing factor is set as $N(x) = \sqrt{|\mathcal{H}|}$ since Yessenalina *et.al.*[35] demonstrates that the square root normalization can be useful, where the size of the extracted sentiment explanations $|\mathcal{H}|$ will be further discussed in the experimental section. To analyze the aspect of each sentence, we need to give an initial guess of the aspect and polarity for each sentence.

Sentence-level Polarity Initialization

To initialize the sentence level polarity, we employ a rule based method that counts positive and negative sentiment terms, with adversative relation considered. The decision rule is that if there are more positive terms than negative ones, the polarity of a sentence is positive, otherwise negative.

Sentence-level Aspect Assignment Initialization

Obviously, if a signature term of aspect a occurs in sentence x^l, we assign aspect a to x^l, and add x^l to an aspect specific sentence set S_a. For sentence x^l without any aspect term, we set a as the aspect label if

$$a = \underset{a'\in A}{\operatorname{argmax}}\{similarity(x^l, S_{a'})\}$$

We use cosine similarity and select the sentences whose polarity is consistent with the overall rating of a review as the initial guess of the sentiment explanation ($\mathcal{H}$).

4. EXPERIMENTS

4.1 Data Preparation

We crawled thousands of reviews from some social review sites such as dianping.com and daodao.com (Chinese version of tripAdvisor) to evaluate the proposed model. Each of these reviews has an overall rating ranging from one to five stars. We consider a review as positive if its rating is greater than or equal to 4 stars, or negative

if less than or equal to 2 stars, and leave neutral reviews as future work. Table 2 presents some statistics of the training corpus. It should be noted that our model is trained on this dataset only with a handful set of aspect signature terms. To further evaluate aspect analysis, we also manually labeled 884 reviews, in which each sentence is labeled in terms of polarity and aspect. Table 3 shows the statistics about the evaluation data.

domain	# Reviews for Training	
	Positive	Negative
restaurant	5,000	5,000
hotel	1,500	1,500

Table 2: Statistics of the data for training

domain	Aspect Annotated Reviews		
	Positive	Negative	#.sentences
restaurant	228	221	4,405
hotel	218	217	3,643

Table 3: Statistics of the labeled data for evaluation

The training corpus is then split into 10 folds. Two folds are left out for test, 7 folds for training, and 1 fold for development, the performance is averaged over 5 runs. For each domain, we pre-defined several aspects, each of which is represented by some signature terms that can be easily obtained by manual annotation on top frequent aspect words. The average number of signature terms for the pre-defined aspects is around 10, and table 4 presents several samples of the aspect signature terms used in this paper.

Domain	Aspect	Signature Terms
Restaurant	Taste	味道"taste", 口味"flavor"
	Ambience	环境"environment", 装修"decoration"
	Location	位置"location"
	Service	服务"service", 服务员"waiter", 态度"attitude"
	Price	价格"price" , 价钱"cost"
	Quality	质量"quality", 份量"quantity"
Hotel	Room	房间"room", 套房"suite"
	Service	服务"service", 前台"Front desk"
	Food	食物"food", 早餐"breakfast"
	Location	位置"location", 交通"traffic"
	Price	价格"price", 价钱"cost"
	Ambience	环境"environment", 气味"smell"
		装修"decoration", 隔音"soundproof"
	Facilities	设施"facility", 网络"Internet"

Table 4: Samples of aspect signature terms.

To evaluate our model in terms of both document-level sentiment classification and aspect prediction on the extracted "sentiment explanations", we design the following experiments:

- Document-level Sentiment Classification
  - We firstly compare the performance of our model over different size of "sentiment explanations" to obtain the optimal size of $\mathcal{H}$.
  - Secondly, we compare our model with standard SVM for document-level sentiment classification.
  - Thirdly, we evaluate the performance with different α and β to investigate the impact of aspect- and document-level loss functions on document-level sentiment classification.
- Case Studies on Aspect Analysis
  - We present case studies for aspect representative sentences and document-level "sentiment explanation" extraction, respectively.
- Quantitative Analysis
  - We firstly evaluate the polarity of the extracted sentences to verify whether the extracted sentences are coherent with the overall rating of a review.
  - Secondly, we compare the performance of aspect assignment on the extracted sentences with different α and β to verify whether it is effective to encode aspect information.
  - Thirdly, we compare our model with SVM-multiclass.
  - We then evaluate the performance of our model under different size of aspect signature terms.
  - Finally, we study the results and demonstrate the informativeness of the extracted "sentiment explanations".

4.2 Document-level Sentiment Classification

The Optimal Extraction Size

To perform document-level sentiment classification, we have to firstly determine the optimal number of extracted sentences (we term it *extraction size*). For simplicity, we set α and β to 1 and choose "Zero/One loss", which is the percentage of the wrong predictions, as the measure to evaluate document-level sentiment classification.

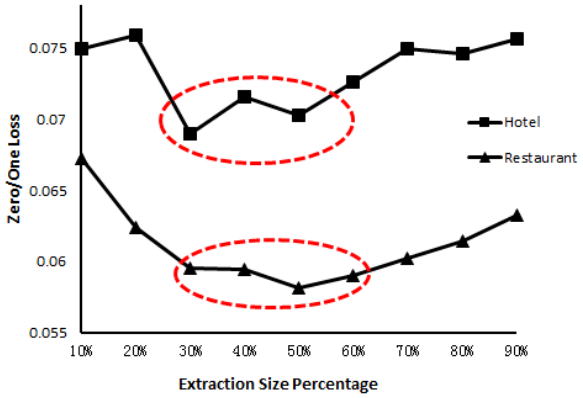

Figure 3: Zero/One loss of document-level sentiment classification by varying the extraction size

Figure 3 presents the performance of document-level sentiment classification by varying the extraction size. Initially, the Zero/One loss decreases when the extraction size increases, indicating that for the majority of extracted sentences, their polarities are coherent with the corresponding review's overall rating, which helps to improve the performance (we will further verify this later in Section 4.4.1). After the loss reaches a minimum, it increases when more sentences are extracted. The reason might be that sentences with inconsistent polarities (without opinion or even with opposite polarity) are extracted as sentiment explanations, which leads to the performance degradation.

Another observation is that when the extraction size is around 50%, our model reaches the best performance (See the area tagged by the eclipse in Figure 3). As is shown in Table 1, the percentage

of informative sentences is around 50%, and the coincidence may be the evidence that our model can extract those highly informative sentences. Due to these observations, we set the optimal extraction size to 50% to avoid losing possible opinionated sentences in the later experiments.

Sentiment Explanations for Document-level Sentiment Classification

To justify whether using the extracted *sentiment explanations* can improve document-level sentiment classification, we compared our method to standard SVM[21][2]. We set the extraction size to 50%. The major difference is that, in our model, document polarity is predicted over the extracted *sentiment explanations*, as illustrated in Equation 1, while standard SVM employs all of the sentences in a review.

domain	SVM	Our method
restaurant	92.55%	**94.19%**
hotel	91.91%	**93.11%**

Table 5: Accuracy of document-level sentiment classification

Table 5 clearly shows that, our model remarkably outperforms the baseline on document-level sentiment classification. Note that the SVM model is a very strong baseline for this task, as discussed in [21].

Impact of Encoding Weak Supervision on Document-level Sentiment Classification

In our model, α and β are designed for aspect analysis. We shall first investigate if α and β affect the performance of document-level sentiment classification before further aspect analysis. It can be done as follows: firstly β is set to 1 by varying α, and secondly α is set to 1 by varying β, both under the optimal extraction size (50%). Figure 4 and Figure 5 show that, the performance of document-level sentiment classification is fairly stable when changing α or β respectively, which also implies the robustness of our model on document-level sentiment classification.

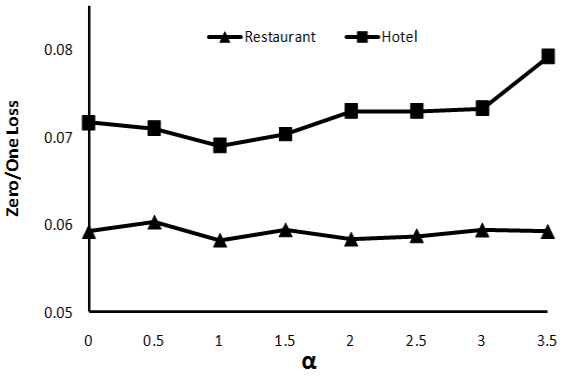

Figure 4: Zero/One loss of document-level sentiment classification by varying α (extraction size 50%)

4.3 Case Studies on Aspect Analysis

In this section, we present several case studies for both aspect representative sentences and document-level "sentiment explanation" extraction .

[2]http://svmlight.joachims.org/

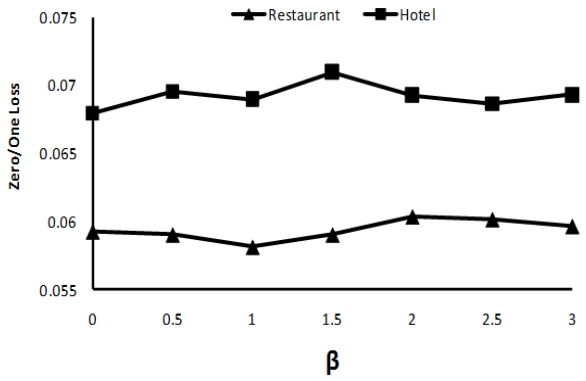

Figure 5: Zero/One loss of document-level sentiment classification by varying β (extraction size 50%)

Aspect Representative Sentences

Our model aims to assign the extracted *sentiment explanations* with polarity and pre-defined aspect labels. Table 6 presents some samples of the extracted *sentiment explanations*. It shows that even though some of these sentences do not explicitly contain aspect signature terms, our model provides the correct prediction of aspect assignments. Therefore, our model is capable of assigning the predefined aspect labels to the extracted sentences correctly.

Aspect	Aspect Representative Sentence
Taste	POS: 手切雪花牛肉很嫩 "The hand-curved beef is very tender." NEG: 墨鱼红烧肉非常咸。 "The braised squid and meat is too salty."
Ambience	POS: 喜欢这里的装饰 "I love the decoration here." NEG: 里面烟蒙蒙的！ "The room is very smoky!"
Service	POS: 服务员也挺勤快，挺专业的。 "The waiters are all hardworking and professional." NEG: 又催了服务员N 次才上的。 "We requested the waiter many times to serve us."
Price	POS: 每次吃完总会有抵扣券，可以和网上的优惠券一起使用的。 "We usually get the coupon after payment, and it can be used immediately with the online coupons." NEG: 靠近窗口的位子是要设最低消费的，根本不合理. "Tables by the window have minimum order amount, which is not acceptable."
Quality	POS: 吃了干锅牛蛙，感觉比霸王蛙更入味。 "Griddle cooked bullfrogs are much tasty than other dishes." NEG: 担担面绝对不是正宗的做法。 "Tan-Tan Noodles are absolutely not cooked in the authentic way."

Table 6: Samples of aspect specific sentence.

Document-level "Sentiment Explanation"

Here, we present an example of the extracted "sentiment explanation" from the document perspective. Table 7 shows a sample of 4-star restaurant review, mentioning price, taste, ambience, and ser-

vice. It can be seen that our model correctly assigns aspect labels to the three extracted sentences, all of which are coherent with the polarity of the review. Obviously, the reviewer praised the restaurant on all the mentioned aspects except price, as "too expensive" was even mentioned twice. However, "one flaw cannot obscure the great virtues", the taste, service, and environment are so impressive that the reviewer gave a positive overall rating for this review. From this point of view, our model is capable of extracting sentences that best represents the document in terms of both polarity and aspect.

* 什么都很好的，就是贵啊。"Everything is great except that it is very expensive."

** **川菜，挺合口味的，做得也精致。[Taste]**
"The Szechwan style of cooking is very tasteful and good looking."
环境不错的，去两次都是在靠窗的沙发坐。[Ambience]
"The environment is very nice, and I would like to sit next to the window."
服务员也挺勤快，挺专业的。[Service]
"The waiters are also very hardworking and professional."

就是贵。"But it is really expensive"

* The original review is at *www.dianping.com/review/26716397*

** The **black bold** part is set of sentences with aspects and polarities predicted by our model (sentiment explanations)

Table 7: Sample of a 4-star review document.

4.4 Quantitative Analysis

We present some quantitative analysis of our model in terms of both sentence-level polarity and aspect assignment.

4.4.1 *Sentence-level Polarity*

In this section, we evaluate whether the polarity of the extracted sentences is coherent with the overall rating of the corresponding review. For each labeled review, we obtain the ground-truth "sentiment explanations" based on manual annotation. After that, we compare the predicted sentence set with the ground-truth "sentiment explanations" in terms of precision, recall and F_1 score.

Extraction Size*	Restaurant			Hotel		
	P	R	F_1	P	R	F_1
10%	0.87	0.16	0.27	0.83	0.17	0.29
20%	0.84	0.25	0.38	0.78	0.24	0.36
30%	0.82	0.39	0.53	0.77	0.38	0.51
40%	0.79	0.48	0.59	0.75	0.46	0.57
50%	0.77	0.61	0.68	0.73	0.61	0.66
60%	0.75	0.69	0.72	0.71	0.67	0.69
70%	0.72	0.78	0.75	0.69	0.77	0.73
80%	0.70	0.85	0.77	0.67	0.83	0.74
90%	0.68	0.90	0.78	0.65	0.89	0.75

* $\alpha = \beta = 1$, P is short for precision and R is for recall

Table 8: Performance of sentence-level polarity prediction

Table 8 shows the performance of sentence-level polarity prediction on different extraction size. It can be seen that the recall and F_1 score increase rapidly when the extraction size grows from 20% to 60%, and the F_1 score then stays fairly stable. For both restaurant and hotel reviews, it only increases 6 percent when the extraction size increases from 60% to 90%. The precision shows that the majority of the extracted sentences are coherent with review overall rating. It should be noted that the sentence polarity is only inferred based on the document polarity. The results demonstrate that our model is capable of extracting *sentiment explanations* that are coherent with the polarity of a review.

4.4.2 *Aspect Assignment*

Our model is also capable of predicting aspect labels to the extracted informative sentences. To evaluate the performance of aspect assignment, we compare the aspect labels predicted by our model with manual annotation for the extracted sentences on the labeled reviews. Previous experimental results show the optimal extraction size is around 50%. Therefore, we set the extraction size to 50%.

Effectiveness of Encoding Aspect Signature Terms

We first evaluate whether the weak supervision (introduced by aspect signature terms) encoded in our model is effective for aspect analysis. We varied the value of parameter α and β to see how the encoded weak supervision impacts on the performance of aspect assignment. Note that we set β to 1 when evaluating the parameter α, and the evaluation for β is in the similar procedure. Figure 6 and Figure 7 show the precision of aspect assignment by varying α and β on restaurant and hotel reviews respectively. In addition, we also present the performance when $\alpha = \beta = 0$, indicating that aspect signature terms are only used for model initialization[3], as the dashed line shown in Figure 6 and Figure 7.

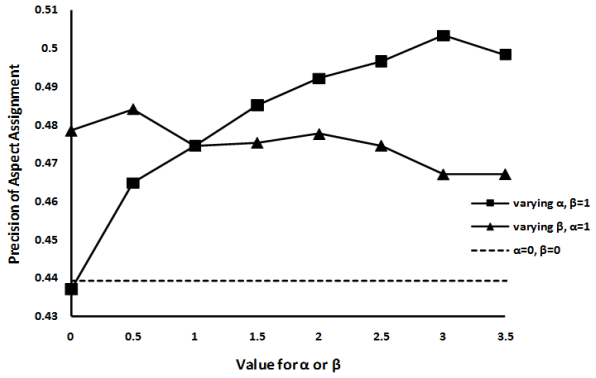

Figure 6: Precision of aspect assignment on restaurant reviews by varying α and β

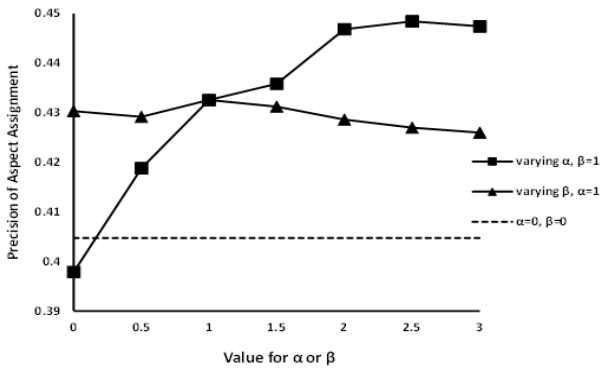

Figure 7: Precision of aspect assignment on hotel reviews by varying α and β

[3]The most common way that prior studies utilized aspect signature terms.

Figure 6 and Figure 7 both show that when β is fixed, the precision of aspect assignment firstly increases with respect to α, and reaches the maximum value ($\alpha = 3$ for restaurant and $\alpha = 3.5$ for hotel), and then decreases slowly. The reason might be that sentence-level local loss may slightly affect the performance of document-level sentiment classification. We have similar observation for β when α is fixed (the maximum value is reached when $\beta = 0.5$ for restaurant and $\beta = 1$ for hotel). In addition, it can also be observed that remarkable performance improvement can be obtained when aspect-specific knowledge is introduced ($\alpha > 0, \beta > 0$ compared to $\alpha = \beta = 0$), which demonstrates that it is effective and necessary to encode aspect-specific prior knowledge. In real applications, the parameters α and β shall be tuned according to the data. For the case here, we may let $\alpha = 3$ and $\beta = 0.5$ for restaurant reviews, and $\alpha = 2.5$ and $\beta = 1$ for hotel reviews, or use other grid-search methods such as optimization approach to obtain an optimal setting.

To evaluate the performance of aspect assignment, we employ SVM $multiclass$[4] as baseline. As for SVM $multiclass$, we select sentences with aspect signature terms for model training in the training corpus (see Table 2), and each sentence is treated as a training instance, in which the aspect signature term identifies the corresponding aspect label.

	Restaurant	Hotel
Our Model[#]	47.46%	43.26%
SVM-multiclass	35.00%	41.16%

[#] The extraction size is 50%, $\alpha = \beta = 1$

Table 9: Precision of aspect assignment

Table 9 presents accuracy of the aspect assignment for the extracted sentence. From the result, we can see that for restaurant reviews, our model outperforms much better than SVM $multiclass$, and for hotel reviews, our model shows a slight performance promotion.

As can be seen from Table 9, the performance difference on restaurant reviews is much more significant than that on hotel reviews. We take a further investigation on the data, and find that in restaurant domain, there are more "noisy" reviews than in hotel domain. It should be noted that "noisy" means irrelevant, as nowhere in the entire review is a mention of restaurant or hotel related property. For example, the customer received a parking ticket as his car was parked illegally, and only for this reason, he gave the restaurant a negative overall rating. Such unrelated reviews bring much noise during model training. As our model aims to extract *informative sentences* corresponding to a certain aspect, the results may also indicate that our model is capable of extracting highly informative sentences, without inclusion of those inconsistent sentences that are unrelated to any predefined aspects.

Impact of the Number of Aspect Signature Terms

We now study how the size of the signature term set affects the performance. In comparison, we also choose the SVM $multiclass$ as the baseline. After ranking all the aspect signature terms by document frequency on the training reviews in descending order, we select top 20%, 40%, 60%, 80%, and 100% of the total aspect signature terms to investigate how the performance changes with different sizes. Note that α and β are set to 1 as the default setting.

Figure 8 and 9 present the performance of precision on aspect assignment over different size of aspect signature terms for restaurant and hotel review respectively. It can be seen that for restaurant review, our model significantly outperforms the baseline, with gradual improvement when more signature terms are introduced. And for hotel review, our model outperforms the baseline when the seed size is over 60%. It should be noted that on average, 60% means for each aspect, only about 6 aspect signature terms are encoded in the model as weak supervision. Such a small number of aspect signature terms can be obtained at a fairly low cost for any domains in real application.

[4]http://svmlight.joachims.org/svm_multiclass.html

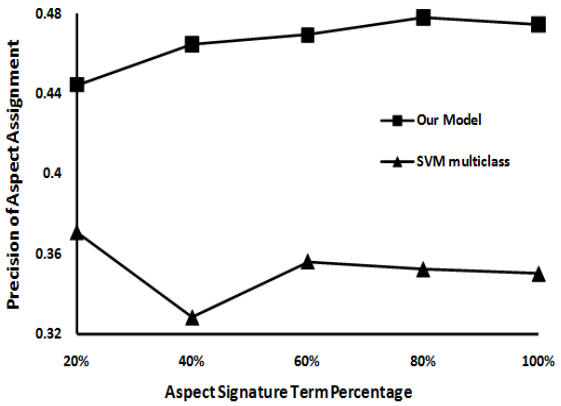

Figure 8: Precision of aspect assignment on restaurant reviews by varying the seed size

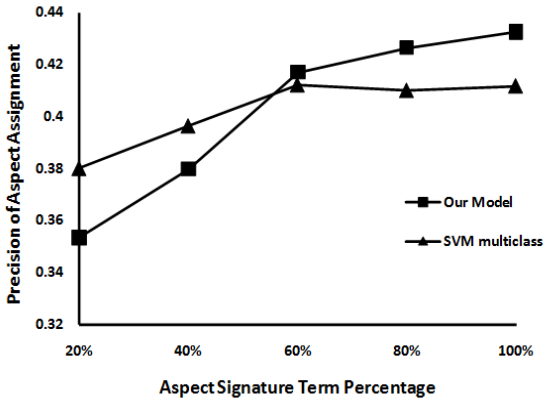

Figure 9: Precision of aspect assignment on hotel reviews by varying the seed size

Informativeness of the "Sentiment Explanations"

As mentioned before, restaurant reviews contain more noisy content than hotel reviews. Topic Models such as LDA[1] are capable of modeling topics in a collection of documents, and many models made use of the latent topics for classification. Examples include the Labeled LDA[24] and Partially Labeled Dirichlet Allocation (PLDA)[25]. The PLDA model is capable of capturing the noisy content with the "background topic", while other topic is learnt with human-interpretable labels, which is quite close to the idea of using signature terms in our work. Hence, we now compare our model with PLDA model on the precision of aspect assignment. We treat each sentence in the training corpus (see Table 2) as a document, and train a PLDA model on the sentences with aspect signature terms that identify the corresponding aspect labels. We

study the performance comparison by varying the size of aspect signature terms. Note that we also set $\alpha = \beta = 1$ with 50% as the extraction size.

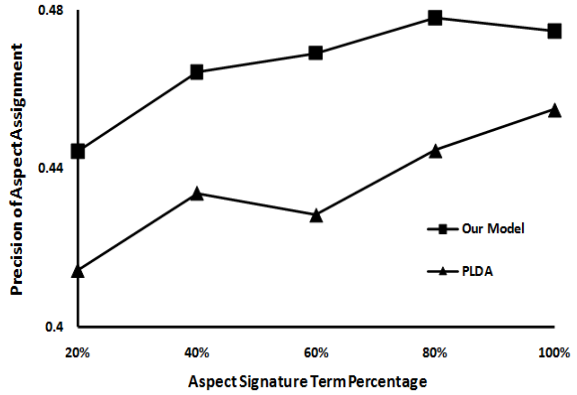

Figure 10: Precision of aspect assignment on hotel reviews by varying the seed size

Figure 10 illustrates results of PLDA and our model. It can be seen that our model outperforms the PLDA model under all sizes of aspect signature terms. Noticing that the PLDA model outperforms SVM $multiclass$ remarkably, partially due to the reason that noisy contents are grouped into the "background topic". From this point of view, our model is capable of extracting informative and representative sentences, filtering those sentences which are unrelated to any predefined aspects.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a structural learning model with a handful set of aspect signature terms that are encoded as weak supervision. Our model aims to extract *latent sentiment explanations* that are aspect-specific informative sentences whose polarity is consistent with the overall rating of a review. "*Sentiment explanations*" represent the original review in terms of both polarity and aspect, and are modeled with latent variables in this work. The proposed model is capable to perform both document- and aspect-level review analysis, and the performances of the two tasks are mutually enhanced through joint modeling. To summarize, the major contributions of this work are as follows:

- We incorporate aspect signature terms as weak supervision during model training, which remarkably promotes the performance of aspect-level analysis.
- We obtain mutually enhanced performance on document-level sentiment polarity prediction and aspect analysis through jointly modeling.

Experimental results on the extracted sentences also demonstrate that our model is capable of extracting latent *sentiment explanations* that are representative and informative in terms of both polarity and aspect. In addition, our model is a general approach which can be easily extended to other domain without re-implementation, except collecting a handful set of aspect signature terms as weak supervision.

As for future work, we plan to further improve aspect-level analysis by incorporating context information of aspect signature terms to obtain better semantic coherence. It may also be interesting to predict aspect rating by performing regression for the latent variables. We may also apply the results of our model for other sentiment analysis task such as aspect-oriented summarization.

6. ACKNOWLEDGMENTS

This paper was partly supported by the National Key Basic Research Program (also called 973 Program) with No.2013CB329403, the National Science Foundation of China with No.61272227 and Tsinghua University Initiative Scientific Research Program with No.20121088071.

7. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 804–812, 2010.

[3] D. Decoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 46:161–190, 2002.

[4] L. Fang and M. Huang. Fine granular aspect analysis using latent structural models. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 333–337, 2012.

[5] G. M. Fung, O. L. Mangasarian, and J. W. Shavlik. Knowledge-based support vector machine classifiers. *Advances in neural information processing systems*, 15:521–528, 2002.

[6] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su. Product feature categorization with multilevel latent semantic association. In *Proceedings of the 18th ACM conference on Information and Knowledge Management*, CIKM '09, pages 1087–1096, 2009.

[7] Z. Hai, K. Chang, and G. Cong. One seed to find them all: mining opinion features via association. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 255–264, 2012.

[8] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, 2004.

[9] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, AAAI '04, pages 755–760, 2004.

[10] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 815–824, 2011.

[11] F. Lauer and G. Bloch. Incorporating prior knowledge in support vector machines for classification: A review. *Neurocomputing*, 71:1578–1594, 2008.

[12] P. Li, Y. Wang, W. Gao, and J. Jiang. Generating aspect-oriented multi-document summarization with event-aspect model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1137–1146, 2011.

[13] T. Li, Y. Zhang, and V. Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*

Processing of the AFNLP: Volume 1 - Volume 1, ACL '09, pages 244–252, 2009.

[14] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351, 2005.

[15] B. Lu, M. Ott, C. Cardie, and B. K. Tsou. Multi-aspect sentiment analysis with topic models. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pages 81–88, 2011.

[16] Y. Mao and G. Lebanon. Isotonic conditional random fields and local sentiment flow. In *Proceedings of Advances in Neural Information Processing Systems*, NIPS '07, 2007.

[17] R. McDonald, K. Hannan, T. Neylon, M. Wells, and J. Reynar. Structured models for fine-to-coarse sentiment analysis. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL '07, pages 432–439, 2007.

[18] P. Melville, W. Gryc, and R. D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1275–1284, 2009.

[19] S. Moghaddam and M. Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 665–674, 2011.

[20] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 339–348, 2012.

[21] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of Empirical methods in natural language processing*, EMNLP '02, pages 79–86, 2002.

[22] G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37(1):9–27, Mar. 2011.

[23] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden-state conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1848–1852, Oct. 2007.

[24] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, 2009.

[25] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 457–465, 2011.

[26] C. Shen and T. Li. A non-negative matrix factorization based approach for active dual supervision from document and word labels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 949–958, 2011.

[27] I. S. Silva, J. Gomide, A. Veloso, W. Meira, Jr., and R. Ferreira. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 475–484, 2011.

[28] B. Snyder and R. Barzilay. Multiple aspect ranking using the good grief algorithm. In *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference*, HLT-NAACL, pages 300–307, 2007.

[29] O. Täckström and R. McDonald. Discovering fine-grained sentiment with latent variable structured prediction models. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 368–374, 2011.

[30] J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *Science*, 331:1279–1285, 2011.

[31] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*, ACL '08, pages 308–316, 2008.

[32] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 104–, 2004.

[33] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 783–792, 2010.

[34] X. Wu and R. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 326–333, 2004.

[35] A. Yessenalina, Y. Yue, and C. Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1046–1056, 2010.

[36] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, 2009.

[37] A. L. Yuille and A. Rangarajan. The concave-convex procedure (cccp). *Neural Computation*, 15:915–936, 2003.

[38] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 56–65, 2010.