A Hierarchical Aspect-Sentiment Model for Online Reviews

Suin Kim[1]*, Jianwen Zhang[2], Zheng Chen[2], Alice Oh[1], Shixia Liu[2]
[1]Department of Computer Science, KAIST, Republic of Korea
[2]Microsoft Research Asia, Beijing, P.R. China
[1]{suin.kim@kaist.ac.kr, alice.oh@kaist.edu}, [2]{jiazhan,zhengc,shliu}@microsoft.com

Abstract

To help users quickly understand the major opinions from massive online reviews, it is important to automatically reveal the latent structure of the aspects, sentiment polarities, and the association between them. However, there is little work available to do this effectively. In this paper, we propose a hierarchical aspect sentiment model (HASM) to discover a hierarchical structure of aspect-based sentiments from unlabeled online reviews. In HASM, the whole structure is a tree. Each node itself is a two-level tree, whose root represents an aspect and the children represent the sentiment polarities associated with it. Each aspect or sentiment polarity is modeled as a distribution of words. To automatically extract both the structure and parameters of the tree, we use a Bayesian nonparametric model, recursive Chinese Restaurant Process (rCRP), as the prior and jointly infer the aspect-sentiment tree from the review texts. Experiments on two real datasets show that our model is comparable to two other hierarchical topic models in terms of quantitative measures of topic trees. It is also shown that our model achieves better sentence-level classification accuracy than previously proposed aspect-sentiment joint models.

1 Introduction

Online reviews contain rich information on different aspects of a product and the sentiment polarities of users. For example, in laptop reviews, there are comments on aspects such as the overall design, battery, screen, and CPU. Before making purchases, consumers often seek opinions from other users by reading their reviews (Chen and Xie 2008). The key information a consumer wants to get from the reviews is: (1) whether the product is good, and (2) what aspects received positive or negative opinions. This task is quite challenging because it is difficult for a humanbeing to extract statistical aspect-sentiment information from a massive set of online reviews. Thus, in recent years there is surging interest on the research topic of *aspect-based sentiment analysis* (Moghaddam and Ester 2012). However, most of the research so far ignores the hierarchical structure of the aspects and sentiments, which is in fact of great importance and is the focus of this work.

We can understand the necessity of a hierarchical structure in sentiment analysis from the following two viewpoints, consumer and technology.

*This work was performed when Suin Kim visited MSR Asia.

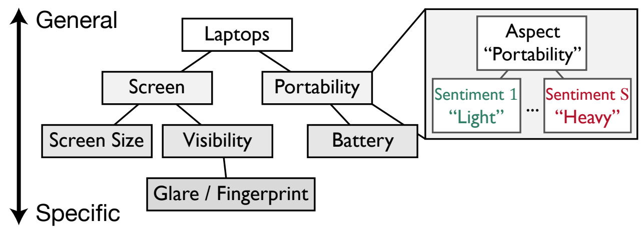


Figure 1: Part of the tree structure discovered by HASM, applied to the LAPTOPS review corpus. Each node itself is a two-level tree, whose root represents the aspect and the children represent the associated sentiment polarities.

From the viewpoint of consumers, different users need different information and hence are interested in different granularities of aspects and sentiments. For example, some consumers care about the opinions of general aspects such as screens and CPUs, while others may pay more attention to more specific aspects such as CPU frequency and cache size. Analysis of aspects and sentiments at some single granularity cannot satisfy all the users. Therefore it is desirable to convey a hierarchy of aspects and sentiments to users so they can easily navigate to the desired granularity. Additionally, a well-organized hierarchy of aspects and sentiments from the general to the specific is easy to understand by humanbeing.

From the viewpoint of technology, the tree structure helps sentiment analysis. Well-identified sentiment words contribute much to the accuracy of sentiment analysis. In real world applications, the polarities of many words depend on the aspect (Li, Huang, and Zhu 2010), and the differentiation of prior and contextual polarity is crucial (Wilson, Wiebe, and Hoffmann 2009). For example, the word "fast" is positive when used to describe a CPU, but it would be negative when describing a battery. This problem is a great challenge for aspect-based sentiment analysis, especially for unsupervised models. Commonly, we provide only general sentiment seed words such as "good" and "bad" which of of little help in identifying aspect-specific sentiment expressions. Existing unsupervised models try to propagate the polarity of these general words to the aspect-specific sentiment words by their co-occurrences within some context, but it is difficult to do so from the most generic sentiment words to the fine-grain aspect-specific words since the co-occurrences can be quite sparse. Instead, by discovering the hierarchical structure, we can propagate the polarities along the hierarchy from the general to the specific so that most aspect-specific sentiment words can be identified.

In this paper, we aim to extract an aspect-sentiment hierarchy from reviews. It provides users (1) an overall evaluation of the products, (2) aspect-based sentiments, (3) sentiment-based aspects, and it also supports (4) navigation of aspects and the associated sentiments over the tree at different granularities. However, it is not easy to jointly model aspect hierarchy and sentiments because it needs to learn both the hierarchical structure and the aspect-sentiment topics from an unlabeled corpus. Although there is related research on modeling topic hierarchies (Blei, Griffiths, and Jordan 2010), incorporating sentiment makes it even more challenging.

To tackle this challenge, we propose a hierarchical aspect sentiment model (HASM) to discover a hierarchical structure of aspect-based sentiments from unlabeled online reviews. In HASM, the whole structure is a tree. Each node itself is a two-level tree, whose root represents an aspect and the children represent the sentiment polarities associated with it. Each aspect or sentiment polarity is modeled as a distribution of words. To automatically extract both the structure and parameters of the tree, we use a Bayesian nonparametric model, recursive Chinese Restaurant Process (rCRP) (Kim et al. 2012), as the prior and jointly infer the aspect-sentiment tree from the review texts. Empirical studies on two real review datasets show the effectiveness of the proposed model.

2 Related Work

Overall, there are two types of work related to ours: aspect-based sentiment analysis and hierarchical topic modeling.

Much work has been devoted to effectively perform aspect-based sentiment analysis. **Joint modeling of Sentiment and Topic (JST)** (Lin and He 2009) is a flat topic model based on LDA (Blei, Ng, and Jordan 2003). For each polarity, a flat mixture of topics is associated with it and all the words with the polarity are generated from this mixture. The drawback of JST is that finding the different polarities for the same topic is difficult. The **ASUM** model proposed by (Jo and Oh 2011) assumes all words from a sentence are of the same polarity, hence it can be considered as a sentence-level JST model. **Reverse JST (RJST)** (Lin et al. 2012) reverses the association direction between topics and polarities in JST. It associates all the polarities to each semantic topic. While RJST makes it convenient to find the different polarities for the same topic, RJST performs poorly on document-level sentiment analysis. Similarly, (Kawamae 2012) discovers aspects and their corresponding sentiment by dividing the topics into aspect-sentiment mixtures. **Seeded aspect and sentiment model** (Mukherjee and Liu 2012) discovers aspect-based sentiment topics given seed words for aspect categories. However, all of the above models treat aspects as a flat mixture of topics, ignoring the natural hierarchical structure inside the aspects.

There are several models that effectively discover topic hierarchies from text (Blei, Griffiths, and Jordan 2010; Kim et al. 2012; Li and McCallum 2006; Adams, Ghahramani, and Jordan 2010), but they cannot perform sentiment analysis. Actually they provide us powerful modeling tools based on which we can design our sentiment models. **Multi-grain Topic Model** (Titov and McDonald 2008) and (Shi and Chang 2008) handle multiple granularities of aspects in the context of online review analysis, but these actually only deal with aspect extraction not sentiment modeling, which means they are closer to hierarchical topic modeling.

There has been a great deal of research into discovering both aspects and the related sentiments outside of the topic modeling framework. (Bloom, Garg, and Argamon 2007) extracts an appraisal expression that has attitude, target, and source. (Kessler and Nicolov 2009) and (Jin, Ho, and Srihari 2009) proposed a machine learning framework to discover both aspects and sentiment expressions semantically related to each aspect or target. To leverage the ability to discover sentiment-aspect pairs, (San Pedro, Yeh, and Oliver 2012) analyzed a set of user comments to infer opinions about the aesthetic value of an image. (Kennedy and Inkpen 2006) used three types of valence shifters: negations, intensifiers, and diminishers, to determine sentiment polarities expressed in movie reviews.

As far as we know, the only work on the hierarchical structure of aspects in sentiment modeling is the **Sentiment Ontology Tree** (Wei and Gulla 2010), which analyzes the reviews of one product by manually labeling the product's aspects and sentiments in the review text. It leverages a hierarchical ontology of aspects and sentiments, but the hierarchy is pre-specified rather than learned from data.

3 Hierarchical Aspect-Sentiment Model

We approach the problem of discovering a hierarchy of aspects and sentiments with an intuitive idea:

> A corpus of reviews contains a latent structure of aspects and sentiments that can naturally be organized into a hierarchy, and each of those node is made up of an aspect and the sentiment polarities associated with it.

To implement the idea, we define an *aspect-sentiment tree* T to represent the hierarchical structure. Rather than specifying the width or depth of the tree beforehand, we adopt a Bayesian nonparametric approach to learn the tree from the data. The core components of this model are: (1) the definition of the tree T, (2) the likelihood $p(\boldsymbol{w}|T)$, how the review corpus $\boldsymbol{w}$ is generated from the tree, (3) prior $p(T)$, how the tree is generated in prior, and (4) the posterior $p(T|\boldsymbol{w})$ after the reviews are observed. The first three components are introduced in the following three subsections one by one. The last component is an inference task, which will be introduced in Section 4.

3.1 Definition of T: Aspect-Sentiment Tree

Figure 2 shows a magnified view to the proposed *aspect-sentiment tree*. To capture the idea that each aspect is associated with a set of sentiments, each node of the tree T, called an *aspect-sentiment node*, is itself a tree. Specifically, as shown in Figure 2, an aspect-sentiment node Φ_k for aspect k consists of an aspect topic φ_k^0 at the root and S sentiment-polar topics $\varphi_k^1 \cdots \varphi_k^S$ as the children. Each aspect/sentiment topic is a Multinomial distribution over the whole vocabulary.

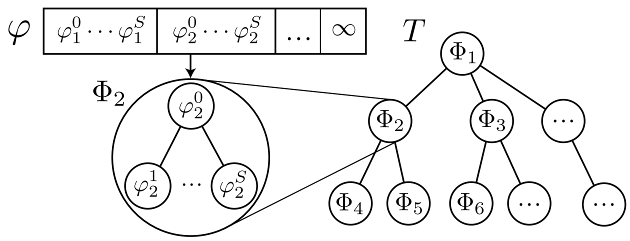


Figure 2: An aspect-sentiment hierarchy and a magnified aspect-sentiment node. An aspect-sentiment node is defined by two latent variables: An infinite-depth, infinite-breadth tree T and a topic set φ, which are independently drawn from Dirichlet(β) distribution. For each aspect-sentiment node Φ_k, we set a two-level topic tree, having aspect topic φ_k^0 and S sentiment polar topics $\varphi_k^1 \cdots \varphi_k^S$.

All the $S+1$ topics within a node Φ_k are independently generated from a same Dirichlet(β) distribution, which means they share a same general semantic theme. With this nested definition of an aspect-sentiment node, we are able to distinguish between the aspect topics and the aspect-dependent sentiment-polar topics. Organizing these nodes into the tree T, we are able to represent the topics in a hierarchy from the general near the root to the specific near the leaves.

Notice that neither the width nor the depth of the tree T is specified beforehand. Instead, we adopt a Bayesian nonparametric approach. Without observing any data, the tree T can be considered of infinite-width and infinite-depth. After observing a review corpus, the structure of the tree T will be inferred from data. Such mechanism is explained in Section 3.3.

3.2 Likelihood $p(\boldsymbol{w}|T)$: Document Modeling

We describe the generative process of a set of reviews given an aspect-sentiment tree T. For each review document d, we first draw a sentiment distribution $\pi \sim$ Dirichlet(η). Each document, sentence, and word is conditionally independent. The graphical representation of HASM is shown in Figure 3, and the notations are explained in Table 1. The formal generative process for each sentence i in document d is as follows:

1. Draw an aspect-sentiment node $c \sim T$.
2. Draw a sentiment s ∼Multinomial(π).
3. Draw a subjectivity distribution $\theta \sim$ Beta(α).
4. For each word j,
   (1) Draw a word subjectivity p ∼Binomial(1, θ).
   (2) Draw the word w ∼Multinomial($\varphi_c^{s\times p}$).

We model each sentence to represent one of the S sentiment polarities. Each sentence is generated from the mixture of aspect topic and s-sentiment polar topic with the mixing proportion θ. For each word in a sentence, the model determines its polarity, $s \times p$, from the sentence sentiment s and the word subjectivity p. Subjectivity for each word indicates whether the word has sentiment polarity or not, and can have one of two values {0-*Non-subjective*, 1-*Subjective*}.

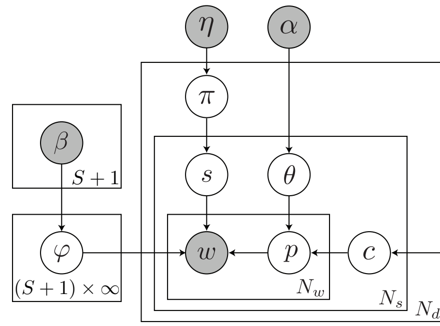


Figure 3: Graphical representation of HASM.

For example, a *subjective* word "glare" under the negative-sentiment sentence is generated from the negative polar topic, while a *non-subjective* word "screen" is an aspect word, regardless of the sentiment polarity of the sentence.

The hyperparameter α is the uniform Dirichlet prior controlling the mixing probability for the word subjectivity. Another hyperparameter η controls the sentiment mixing probability and can be interpreted as a prior observation count for the sentiment-polar sentences in a review.

3.3 Prior $p(T)$: rCRP Prior to Generate the Tree

We use a Bayesian nonparametric model, the recursive Chinese restaurant process (rCRP) (Kim et al. 2012), as a prior to generate the tree T. rCRP is a stochastic process which defines a probability $p(T)$ for a tree T with unbounded depth and width. Thus it is useful for automatically learning the structure of a tree from data. With rCRP as the prior, the formal generative process of the aspect-sentiment tree T is as follows:

1. Draw a tree T with unbounded width and depth from rCRP prior:
$$T \sim \mathrm{rCRP}(\gamma).$$
2. Draw aspect/sentiment-polar topics for each node:
$$\forall s \in \{0, \ldots, S\}, \forall k \in \{0, \ldots, \infty\}, \varphi_k^s \sim \mathrm{Dirichlet}(\beta_s).$$

The hyperparameter γ of rCRP controls the overall structure of the tree. A high value for γ increases the probability of generating children for each aspect-sentiment node, which favors a wide and deep tree in prior. In contrast, a small γ favors a narrow and shallow tree in prior.

Besides rCRP, another representative model which can be used as the prior for a tree is the nested Chinese restaurant process (nCRP) (Blei, Griffiths, and Jordan 2010). The key difference between these two models is: nCRP constrains that all the words in a document comes from a single path in the tree, while rCRP allows the words of a document coming from any subset of the nodes in the tree. For our problem, it is not reasonable to model a whole piece of review text from a single aspect path as a review often contains multiple aspects and sentiments. Moreover, we found that nCRP-based model on short documents (e.g., sentences from a review text) tends to assign most words in a document into a single node because a short document is not long enough to disperse the

HYPERPARAMETERS	
α	Uniform Dirichlet prior for θ
β	Non-uniform Dirichlet prior for β
γ	rCRP prior
η	Beta prior for π
LATENT & VISIBLE VARIABLES	
s	Sentence sentiment
p	Word subjectivity
θ	In-sentence subjectivity distribution
c	Aspect for each sentence
π	In-document sentiment distribution
Φ	Aspect-sentiment node
φ	Aspect and sentiment-polar topic
T	rCRP tree
w	Word
PLATE NOTATIONS	
S	Number of sentiment polarities
N_d	Number of documents
N_s	Number of sentences for each document
N_w	Number of tokens for each sentence
SAMPLING NOTATIONS	
m_n	# of sentences assigned to aspect node n
M_n	# of sentences assigned to aspect node n and its all subnodes, including leaf
$\hat{\beta}_s$	$= \sum_w \beta_{s,w}$
$n_{k,s}^{w,(v)}$	# of words having v^{th} vocabulary assigned to aspect k and sentiment s
$n_d^{s,(k)}$	# of k-polar sentences in document d
$n_{d,i}^{p,(k)}$	# of k-subjective words in sentence i of d
n_{-i}	Counter variable excluding index i

Table 1: Variables and notations.

words over different levels along the entire path. As a result, the discovered tree is often fragmented, where a child node does not have any connection with its parent in semantic. The tree-structured stick-breaking model (Adams, Ghahramani, and Jordan 2010), which generates a document from a single node, also conflicts to our assumption. In rCRP, this problem is conquered as rCRP enforces the semantic correlations over the hierarchy by constraining that a word can be assigned to a child node only when it is already assigned to its parent. In the experiments (Section 5), we compared the results of rCRP with nCRP. More technical comparisons among them in detail can be found in (Kim et al. 2012).

In addition, to encourage the discrimination between different sentiment polarities, we incorporate basic human knowledge on sentiments into the model. Based on some common sentiment seed words, we use an asymmetric Dirichlet prior and set weights according the seed words on β (Jo and Oh 2011). To balance the sum of Dirichlet prior, we use the same number of seed words for each sentiment and aspect.

4 Inference

The task of inference is to learn the posterior $p(T|\boldsymbol{w})$ of the tree from data. Using collapsed Gibbs sampling, we only need to infer the posterior $p(\boldsymbol{c}, \boldsymbol{s}, \boldsymbol{p}|\boldsymbol{w}, \alpha, \eta, \boldsymbol{\beta}, \gamma)$ of three groups of variables, 1) $\boldsymbol{c}$, the aspect-sentiment node of each sentence 2) $\boldsymbol{s}$, the sentiment of each sentence 3) $\boldsymbol{p}$, the subjectivity of each word in a sentence. Other variables are integrated out and do not need sampling. As this is the routine procedure of Gibbs sampling, limited by space, we only briefly describe its skeleton as follows:

Aspect Sampling (Sampling $\boldsymbol{c}$). We follow the sampling procedure of rCRP, beginning from the root dish and moving down along the path. Each sentence contains statistics about sentiment of the sentence and subjectivity of words. Such information is being kept while the sentence is moving along the path or assigned to a new node. There are three possibilities for aspect-sentiment node Φ_k of each sentence i in document d:

1. $P(\text{Select the node } \Phi_k) \propto m_k \times P(\boldsymbol{w}_{di}|s, \boldsymbol{p}, \Phi_k, \boldsymbol{\beta})$
2. $P(\text{Select a child } c \text{ of } \Phi_k) \propto M_c \times P(\boldsymbol{w}_{di}|s, \boldsymbol{p}, c, \boldsymbol{\beta})$
3. $P(\text{Create a new child}) \propto \gamma \times P(\boldsymbol{w}_{di}|s, \boldsymbol{p}, \phi, \boldsymbol{\beta})$,

and the recursive assign process for each sentence stops if a node is selected by the first or the third assignment.

The probability of generating sentence i in document d from sentiment s, subjectivity $\boldsymbol{p}$, and node Φ_k is

$$P(\boldsymbol{w}_{di}|s, \boldsymbol{p}, \Phi_k, \boldsymbol{\beta}) \propto \prod_{l=0}^{1} \left(\frac{\Gamma(n_{k,s_i\times l,-i}^{w,(\cdot)} + \hat{\beta}_{s_i\times l})}{\prod_w \Gamma(n_{k,s_i\times l,-i}^{w,(w)} + \beta_{s_i\times l,w})} \times \frac{\prod_w \Gamma(n_{k,s_i\times l}^{w,(w)} + \beta_{s_i\times l,w})}{\Gamma(n_{k,s_i\times l}^{w,(\cdot)} + \hat{\beta}_{s_i\times l})} \right).$$

Sentiment Sampling (Sampling $\boldsymbol{s}$). We then sample the sentiment polarity s_{di} of each sentence. Beta hyperparameter η controls the sentiment distribution for each document. Higher η implies more prior confidence that distribution of sentiment polarities are likely to be even. The probability of assigning k^{th} sentiment for sentence i in document d is

$$P(s_{di} = k|\boldsymbol{w}, \boldsymbol{s}, \boldsymbol{p}, \boldsymbol{c}, \boldsymbol{\beta}) \propto (n_{d,-i}^{s,(k)} + \eta) P(\boldsymbol{w}_{di}|k, \boldsymbol{p}, \boldsymbol{c}_{di}, \boldsymbol{\beta}).$$

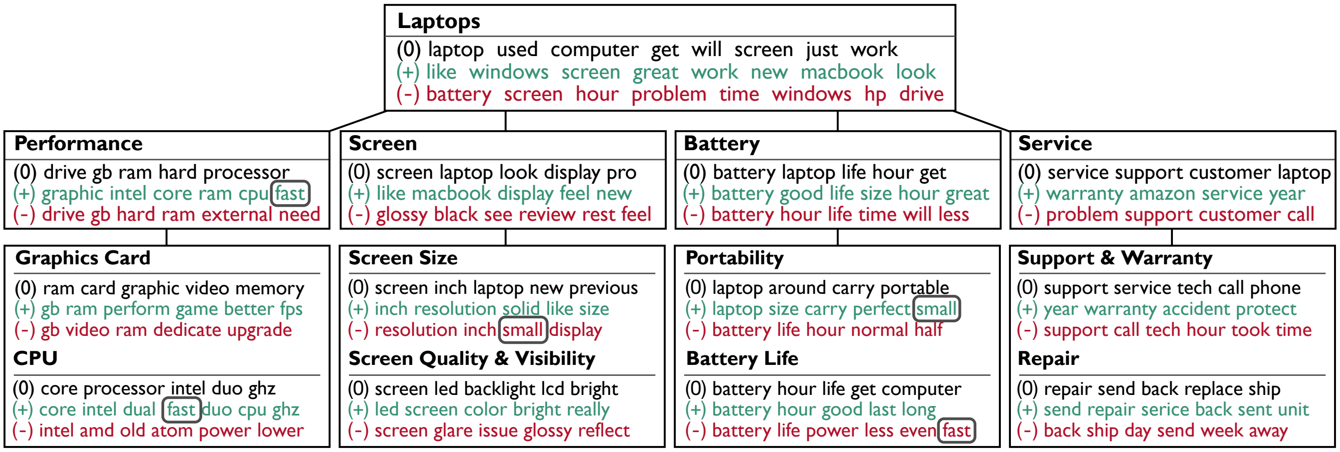
Subjectivity Sampling (Sampling $\boldsymbol{p}$). Finally we sample the subjectivity of each word. Subjectivity sampling is similar to the Gibbs sampling process in a basic LDA model with two topics: $\{$0-*Non-subjective*, 1-*Subjective*$\}$.

$$P(p_{dij} = k|\boldsymbol{w}, \boldsymbol{s}, \boldsymbol{p}, \boldsymbol{c}, \boldsymbol{\beta}) \propto (n_{d,i,-j}^{p,(k)} + \alpha) \times \frac{n_{c_{di},s_{di}\times k,-j}^{w,(v)} + \beta_{s_{di}\times k,v}}{\sum_{r=1}^{V} n_{c_{di},s_{di}\times k,-j}^{w,(r)} + \hat{\beta}_{s_{di}\times k}}.$$

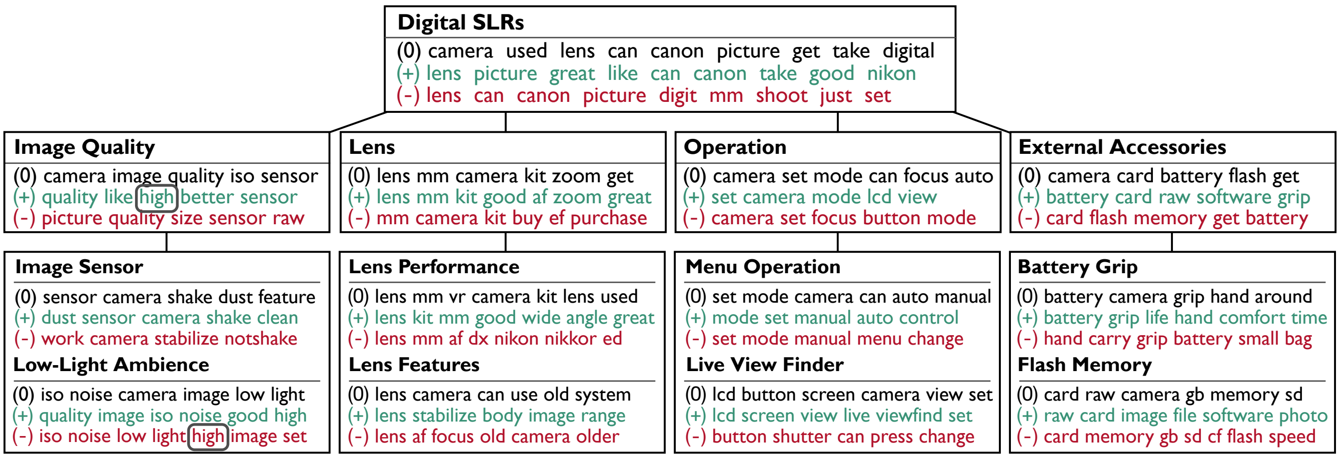
Estimating $\boldsymbol{\varphi}$. In our collapsed Gibbs sampling algorithm, some latent variables such as $\boldsymbol{\varphi}$ and $\boldsymbol{\theta}$, are integrated out. After the Gibbs sampling process, a topic $\boldsymbol{\varphi}$ can be obtained by Bayesian estimation: $\hat{\boldsymbol{\varphi}} = \int \boldsymbol{\varphi} \cdot p(\boldsymbol{\varphi}|\boldsymbol{w}, \boldsymbol{\beta}, \boldsymbol{c}, \boldsymbol{p}, \boldsymbol{s}) d\boldsymbol{\varphi}$.

5 Experiments & Results

In this section, we describe the experiments and analyze the results. We also quantitatively compare the performance of HASM and other hierarchical and aspect-sentiment joint models. A comprehensive comparison is difficult because there is no prior work that discovers a hierarchy of aspect-sentiment nodes. Instead, we divide the evaluation into two parts: hierarchy analysis and sentiment classification.



(a) LAPTOPS dataset



(b) DIGITALSLRS dataset

Figure 4: A part of the aspect-sentiment hierarchy for LAPTOPS and DIGITALSLRS. We define root as a first-level node, and show the second-level aspect-sentiment nodes for which at least 15,000 sentences are assigned to each node. We also show their third-level children. We do not show stopwords or words that occur in both aspect and sentiment-polar topics. Words *fast*, *small*, and *high* occur in both positive and negative polarities depending on the aspect at various levels of granularity.

5.1 Data and Model Setup

We conduct experiments on two publicly available datasets, Amazon.com reviews of laptops (LAPTOPS) and digital SLRs (DIGITALSLRS)[1] (Jo and Oh 2011). Table 2 shows the detailed statistics for each dataset. We use sentiment seed words based on PARADIGM+ (Turney and Littman 2003).

We set up the model to find two sentiment-polar topics $\{$*positive, negative*$\}$ for sentences with the subjectivity value of 1 and an aspect topic for sentences with the subjectivity value of 0. The model hyperparameters are set as $\gamma = 0.1, \alpha = 10.0, \beta = \{10^{-6}, 0.01, 2.0\}$, and $\eta = 1.0$. We pre-process multi-word sentiment expressions "not X", such as "not good" by flipping the sentiment of "X" (Eguchi and Lavrenko 2006). To make most sentences to have single aspect and sentiment polarity, we split sentences by contrasting conjunctions, including "but", "however", and "yet".

[1]http://uilab.kaist.ac.kr/research/WSDM11

5.2 Sentiment-Aspect Hierarchy

We designed HASM to produce a hierarchical structure of aspect and aspect-based sentiment topics such that the hierarchy can be easily traversed to find certain aspects and corresponding opinions at the granularity that the user wants. Figure 4 shows a part of the discovered hierarchical structures for LAPTOPS and DIGITALSLRS. Indeed the hierarchy of aspects matches the intuition that the root node is the most general aspect with the general positive and negative opinions, and as the depth increases, aspects become specific features of the product. Also, the children of an aspect are closely related to their parent. For example, the aspects *Image Sensor* and *Low Light Ambience* are under the parent *Image Quality*. These two characteristics of the hierarchy allow a user to readily identify the opinions for specific aspects and the associated opinions. In a later section, we quantitatively analyze the general-to-specific nature as well as the parent-child relatedness of the aspect-sentiment hierarchy.

	LAPTOPS	DIGITALSLRS
Reviews	10,014	20,862
Targets	825	260
Avg. Reviews/Product	12.14	80.24
Avg. Review Length	1,160.1	1,029.6

Table 2: Statistics for the datasets used in the experiments.

Dataset	Model	Level 2	Level 3	Level 4
LAPTOPS	HASM	0.210	0.624	0.933
	rCRP	0.288	0.696	0.830
	nCRP	0.267	0.464	0.703
DIGITALSLRS	HASM	0.139	0.406	0.935
	rCRP	0.243	0.646	0.822
	nCRP	0.224	0.352	0.550

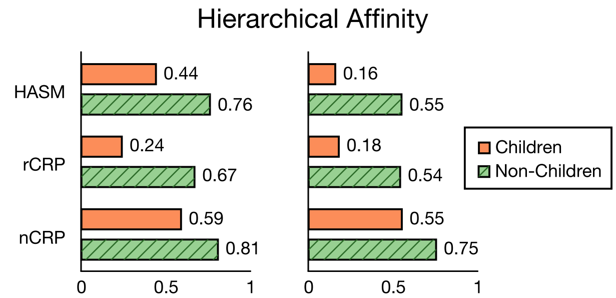
Table 3: The node specialization score for each model. For all three models, as the level increases, the specialization scores increase, which means the model assumptions are correctly reflected in the results.

The discovered aspect-sentiment hierarchy shows the polarities of some sentiment words depend on the aspect at various granularities. For LAPTOPS, "fast" is positive under *Performance* aspect node and its descendants, while it is negative under *Battery Life*. The word "small" also changes its sentiment polarity according to the aspect. For DIGITALSLRS, we found that the word "high" is positive for the medium-grain aspect *Image Quality*, but it turns negative for the fine-grain aspect *Low-Light Ambience*. This result confirms that sentiment words should be analyzed for aspects at different granularities, and HASM is capable of such analysis.

The overall size of a tree depends on both the complexity of data and the rCRP hyperparameter γ, which controls the probability of generating a new child for each aspect-sentiment node at the sampling process. A higher γ leads to a wider and deeper aspect-sentiment tree. We adjusted γ to discover the human-interpretable size of the tree. From the LAPTOPS dataset, HASM discovered 13 second-level aspect-sentiment nodes, which contain more than 50 sentences. For each second-level node, at most 4 third-level children are found.

5.3 Hierarchy Analysis

Held-out perplexity is a commonly used metric for topic models, but it is not appropriate here because neither the semantic quality of topics nor the quality of a hierarchy is reflected in perplexity score. In literatures, there is little work to quantitatively measure the quality of topic hierarchy either. Following the prior research on hierarchical topic modeling (Kim et al. 2012), we use the metrics for *hierarchical affinity* and *topic specialization* to evaluate the hierarchical sentiment-aspect structure. Additionally, we introduce a new metric to measure consistency of the sentiment-polar topics and the aspect topic within a single aspect-sentiment node. Since there is no prior hierarchical aspect-sentiment joint model, we use these three metrics to compare the results from HASM with two hierarchical topic models, rCRP and nCRP.



(a) LAPTOPS (b) DIGITALSLRS

Figure 5: Hierarchical Affinity. The lower distance scores for *Children* indicate that a parent is more similar to its direct children, compared to non-children nodes at the same level. HASM and rCRP show similar patterns in the semantic distances for children and non-children, while nCRP does not.

Distance Metric. A basic metric needed in hierarchical consistency and topic specialization is the distance metric. We use consine similarty to measure the distance Δ between the topics and aspect-sentiment nodes,

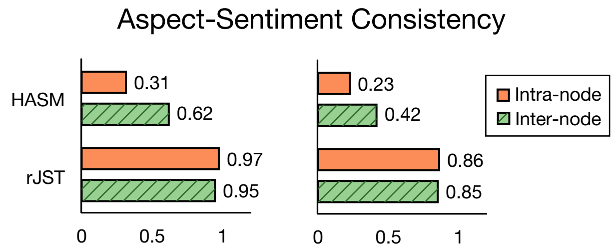
$$\Delta(\varphi_a^s, \varphi_b^s) = 1 - \frac{\varphi_a^s \cdot \varphi_b^s}{\|\varphi_a^s\|\|\varphi_b^s\|} \text{ and}$$

$$\Delta(\Phi_a, \Phi_b) = \frac{1}{S}\sum_{s}^{S} \Delta(\varphi_a^s, \varphi_b^s).$$

We also denote $\bar{\Delta}$ as the average distance between topics or nodes under the given condition.

Node Specialization. One important goal of HASM is to find aspects from the general aspects at the top of the tree to the more specialized aspects toward the leaves. To quantitatively measure how the discovered aspects become more specialized as the depth of the tree increases, we calculate the node distance for each aspect node from a reference node ϕ. Formally, node specialization score at depth l is $\bar{\Delta}_{\Phi_k \in \hat{\Phi}_l}(\phi, \Phi_k)$, where $\hat{\Phi}_l$ is the set of aspect-sentiment nodes at depth l. In our implementation, we select the reference node as the root (Kim et al. 2012). Table 3 shows the average distance from ϕ for each level of the tree, compared to rCRP and nCRP. All three models show similar patterns of increasing distance for increasing levels of the tree.

Hierarchical Affinity. Another important quality of an aspect-sentiment tree is *hierarchical affinity*, which reflects the intuition that an aspect-sentiment node is more similar to its direct descendants than other nodes. For an aspect-sentiment node Φ_k, we measure the average distance to its children, $\bar{\Delta}_{\Phi_c \in \mathrm{Children}(\Phi_k)}(\Phi_k, \Phi_c)$, and compare to the average distance to non-children nodes, $\bar{\Delta}_{\Phi_c \in \mathrm{Non-Children}(\Phi_k)}(\Phi_k, \Phi_c)$. We average distance to children and non-children nodes for all parent nodes at level 2. Figure 5 shows the hierarchical affinity of three models, HASM, rCRP, and nCRP. Both HASM and rCRP show significant differences between children and non-children, while nCRP does not.



(a) LAPTOPS (b) DIGITALSLRS

Figure 6: Aspect-Sentiment Consistency. The results show that HASM achieves lower intra-node average topic distance than inter-node average topic distance. In contrast, rJST shows high average topic distances for both intra- and inter-nodes. The comparison demonstrates that HASM achieves better aspect-sentiment consistency.

Aspect-Sentiment Consistency. In our model, each aspect-sentiment node itself is a tree consisting of aspect and sentiment-polar topics. For each node, HASM discovers the in-node tree with sentiment-polar topics that share a common aspect. We introduce a new metric, *aspect-sentiment consistency*, to measure how in-node topics are statistically coherent in compared to topics outside of the node. We compute and compare the average intra-node topic distance and the average inter-node topic distance. We define that a model with both low intra-node consistency and high inter-node divergence is aspect-sentiment consistent. For an aspect-sentiment node Φ_k at level L and Φ_l^c, the set of other aspect-sentiment nodes at level L, we define aspect-sentiment consistency of Φ_k at level L as

$$\forall \varphi_i \neq \varphi_j,$$
$$\text{INTRA-NODE CONSISTENCY: } \bar{\Delta}_{\varphi_i \in \Phi_k, \varphi_j \in \Phi_k}(\varphi_i, \varphi_j)$$
$$\text{INTER-NODE DIVERGENCE: } \bar{\Delta}_{\varphi_i \in \Phi_k, \varphi_j \in \Phi_k^c}(\varphi_i, \varphi_j).$$

We define aspect-sentiment consistency at level L as the averaged aspect-sentiment consistency for every Φ_k at level L. We compared HASM and rJST, the two aspect-sentiment joint models with the assumption of generating an aspect and then generating a sentiment, rather than vice versa (ASUM and JST). Figure 6 shows the comparison results at level 2.

5.4 Sentence-level Sentiment Classification

Another outcome of HASM is the analysis of the sentence-level aspect-specific sentiment. In this experiment, we compare the sentence-level classification accuracy of HASM with other aspect-sentiment joint models using two subsets of LAPTOPS and DIGITALSLRS. To simulate sentence-level sentiment classification without manual annotation, we choose single-sentence reviews and 100-character or shorter reviews for which we treat the whole review as one sentence. We experiment on two versions, a SMALL dataset containing reviews with 1 (strong negative) or 5 (strong positive) stars, and a LARGE dataset with reviews of 1 and 2 stars (negative) as well as 4 and 5 (positive) stars. We randomly sample reviews for both datasets to have 1,000 reviews for each rating score.

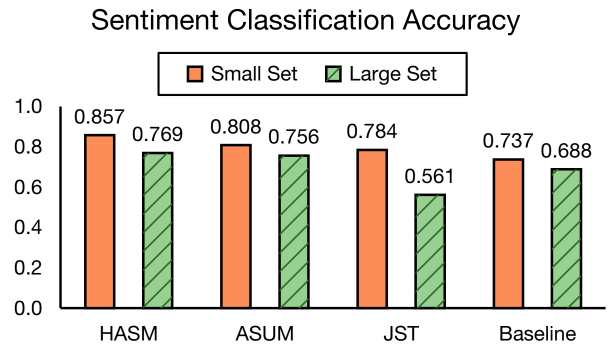


Figure 7: Sentence level sentiment classification. The Small set has reviews with 1 (negative) and 5 (positive) stars, and the Large set has reviews with 1 and 2 stars (negative) as well as 4 and 5 stars (positive). The sentiment classification accuracy of HASM is comparable to the other three models for both SMALL and LARGE datasets.

We use the default parameters for ASUM and JST and set the topic number as 30. For ASUM, which assigns a single sentiment to each sentence, we use the MAP estimation. For JST, we choose the sentiment with the largest proportion in the sentence. For both models, different setting of parameter does not affect much on their classification performance. In the baseline method, we simply count the seed word occurrences and select the sentiment with the largest proportion in the sentence. As shown in Figure 7, the sentiment classification accuracy of HASM is comparable to the other three models for both SMALL and LARGE datasets.

6 Conclusion and Future Work

We presented a Bayesian nonparametric model to discover an aspect-sentiment hierarchy from an unlabeled review corpus. Using a novel design in which each aspect-sentiment node is itself a tree, we built a new model based on rCRP for discovering aspect-sentiment topics over multiple granularities. We applied the model to two datasets from Amazon.com. Compared with the sentiment-aspect joint topic models and hierarchical topic models, the performance of our model is comparable to other models for both sentence-level sentiment classification accuracy and hierarchical topic modeling. Our model is flexible such that it can discover aspects with more than two sentiments, which can be useful for emotion or mood analysis. This framework can be further extended to discover a set of topics with shared features in a hierarchical structure.

7 Acknowledgements

We thank Yangqiu Song for inspiring discussions and Chengtao Li for implementing some baseline methods. Alice Oh was supported by the SW Computing R&D Program of KEIT(2012,10041313, UX-oriented Mobile SW Platform) funded by the Ministry of Knowledge Economy.

References

Adams, R. P.; Ghahramani, Z.; and Jordan, M. I. 2010. Tree-structured stick breaking for hierarchical data. *Advances in Neural Information Processing Systems* 23:19–27.

Blei, D. M.; Griffiths, T. L.; and Jordan, M. I. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM* 57(2):7.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Bloom, K.; Garg, N.; and Argamon, S. 2007. Extracting appraisal expressions. In *Proceedings of NAACL HLT*, 308–315. Association for Computational Linguistics.

Chen, Y., and Xie, J. 2008. Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science* 54(3):477–491.

Eguchi, K., and Lavrenko, V. 2006. Sentiment retrieval using generative models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 345–354. Association for Computational Linguistics.

Jin, W.; Ho, H. H.; and Srihari, R. K. 2009. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1195–1204. ACM.

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

Kawamae, N. 2012. Hierarchical approach to sentiment analysis. In *IEEE Sixth International Conference on Semantic Computing (ICSC)*, 138–145. IEEE.

Kennedy, A., and Inkpen, D. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence* 22(2):110–125.

Kessler, J. S., and Nicolov, N. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In *Proceedings of the third international AAAI conference on weblogs and social media*, 90–97. AAAI.

Kim, J. H.; Kim, D.; Kim, S.; and Oh, A. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 783–792. ACM.

Li, W., and McCallum, A. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, 577–584. ACM.

Li, F.; Huang, M.; and Zhu, X. 2010. Sentiment analysis with global topics and local dependency. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 1371–1376. AAAI.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384. ACM.

Lin, C.; He, Y.; Everson, R.; and Ruger, S. 2012. Weakly supervised joint sentiment-topic detection from text. *Knowledge and Data Engineering, IEEE Transactions on* 24(6):1134–1145.

Moghaddam, S., and Ester, M. 2012. Aspect-based opinion mining from online reviews. *Tutorial at SIGIR Conference*.

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

San Pedro, J.; Yeh, T.; and Oliver, N. 2012. Leveraging user comments for aesthetic aware image search reranking. In *Proceedings of the 21st international conference on World Wide Web*, 439–448. ACM.

Shi, B., and Chang, K. 2008. Generating a concept hierarchy for sentiment analysis. In *IEEE International Conference on Systems, Man and Cybernetics*, 312–317. IEEE.

Titov, I., and McDonald, R. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*, 111–120. ACM.

Turney, P. D., and Littman, M. L. 2003. Measuring praise and criticism: Inference of semantic orientation from association. In *ACM Transactions on Information Systems*, volume 21, 315–346. ACM.

Wei, W., and Gulla, J. A. 2010. Sentiment learning on product reviews via sentiment ontology tree. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 404–413. Association for Computational Linguistics.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics* 35(3):399–433.