Mining Aspect-Specific Opinion using a Holistic Lifelong Topic Model

Shuai Wang, Zhiyuan Chen, Bing Liu
Department of Computer Science
University of Illinois at Chicago
shuaiwanghk@gmail.com, czyuanacm@gmail.com, liub@cs.uic.edu

ABSTRACT

Aspect-level sentiment analysis or opinion mining consists of several core sub-tasks: aspect extraction, opinion identification, polarity classification, and separation of general and aspect-specific opinions. Various topic models have been proposed by researchers to address some of these sub-tasks. However, there is little work on modeling all of them together. In this paper, we first propose a holistic fine-grained topic model, called the JAST (Joint Aspect-based Sentiment Topic) model, that can simultaneously model all of above problems under a unified framework. To further improve it, we incorporate the idea of lifelong machine learning and propose a more advanced model, called the LAST (Lifelong Aspect-based Sentiment Topic) model. LAST automatically mines the prior knowledge of aspect, opinion, and their correspondence from other products or domains. Such knowledge is automatically extracted and incorporated into the proposed LAST model without any human involvement. Our experiments using reviews of a large number of product domains show major improvements of the proposed models over state-of-the-art baselines.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text analysis; I.7.0 [**Document and Text Processing**]: General

Keywords

Opinion Mining; Aspect-Specific Opinion; Topic Model; Lifelong Machine Learning

1. INTRODUCTION

Aspect-level sentiment analysis or opinion mining is a comprehensive task that aims to extract aspects, identify opinions, classify opinion polarity, and recognize general opinions and aspect-specific opinions. In this paper, we refer to these four sub-tasks as *four dimensions* of aspect-level sentiment analysis. To give an example about these four dimensions, let us say a review about a cellphone product mentions "The screen is very clear and great."

WWW 2016, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4143-1/16/04.
http://dx.doi.org/10.1145/2872427.2883086.

1. For aspect extraction, "screen" should be extracted as an aspect.
2. For opinion identification, "clear" should be identified as an opinion word (or simply opinion). Likewise, "great" should also be identified.
3. For polarity classification, "clear" and "great" should be recognized as expressing positive opinions about the "screen".
4. For general and aspect-specific opinion separation, "clear" is an aspect-specific opinion as it indicates the clarity of the aspect *screen*. On the contrary, "great" is a general opinion as it can be used to modify many other aspects. In this paper, we call the characteristic of an opinion (word) expressing a general or aspect-specific opinion as *opinion generality*.

The first three dimensions are clearly useful as they are core problems of sentiment analysis [26]. The fourth dimension is also important because it allows the system to discover opinion reasons, which are interesting to users (e.g., consumers and businesses) too as they almost always want to know what aspects are liked and disliked, and the reasons behind the sentiments/opinions. For example, the review sentence "The picture is bad" expresses a negative sentiment/opinion, but it does not say why the picture is bad, i.e., no reason is given, because the opinion word *bad* is a general opinion word. However, the sentence "The picture is blurry" clearly gives the reason of the negative sentiment because *blurry* is aspect-specific to the aspect *picture* indicating a specific (negative) property. Thus, opinion generality is important and is considered in our work.

Existing research has attempted to tackle some of the above dimensions of aspect-level sentiment analysis. Topic modeling has been popularly applied recently. For example, [25] proposed a joint sentiment/topic (JST) model to identify sentiment polarities of aspects. [19] extended the work and proposed an aspect and sentiment unification model (ASUM) which assumes that all the words in a single sentence are generated from one aspect. [52] separated opinions and aspects by using a maximum entropy model. There are also some other related works, which will be discussed in Section 2. However, the existing models do not have the capacity to model all four dimensions simultaneously. We believe that the unified joint modeling can benefit each dimension through their correlation. In this paper, we take a major step forward and present a holistic solution to jointly model all the four dimensions using a unified framework. Following the existing works, in our paper, an aspect corresponds to a topic in topic modeling.

We first propose a fine-grained topic model, called the JAST (Joint Aspect-based Sentiment Topic) model, to jointly model all four dimensions in a holistic manner. The strength of JAST is that all the component dimensions can help improve each other during the joint modeling process. The rationale here is that we can model each dimension as latent variables in a graphical model, which captures their relationship simultaneously. Experimental results show that JAST achieves significant improvements over the baseline models (Section 5). However, on analysis of the results to gain insights of the JAST model, we found that there was still some room for further improvement.

The main issue with JAST is that it sometimes identifies some general opinion words as aspect-specific, and vice versa. For example, opinion "nice" might sometimes be mistakenly assigned as an aspect-specific opinion for aspect *screen*. One cause of this issue is that fully-unsupervised topic models are not guaranteed to generate coherent topics that are consistent with human judgment [6]. Due to the power law distribution of natural language words, most words do not co-occur with most other words [53]. That means, topic models, which are based on *higher-order word co-occurrences* [15], will suffer from low word co-occurrences. As a result, some coherent aspect-specific opinions cannot be identified while they are mixed with other general opinions within the same topic.

For illustration, let us use an example from our experiments. The word "smooth" should be an aspect-specific opinion word for aspect *screen*. However, in some reviews of the domain/product like Laptop, the co-occurrence for "smooth" and *screen* may not be high enough since not every laptop is equipped with touch-screen ("smooth" is usually more associated with touch-screens). Thus, "smooth" cannot be discovered as an aspect-specific opinion word for aspect *screen* in the JAST model, even though they are in fact being mentioned together in some reviews. On the other hand, the word "nice" is mistakenly identified as an aspect-specific opinion since many occurrences of "nice" happen in the same sentences with *screen*. Due to this high co-occurrence of "nice" and *screen*, the JAST model made the mistake by treating "nice" as an aspect-specific opinion for aspect *screen*.

In order to solve the above problem, we propose a more advanced model called the LAST (Lifelong Aspect-based Sentiment Topic) model. The LAST model incorporates the idea of lifelong machine learning (LML) [44, 8], which has the advantage of extracting and cumulating knowledge from the past learning and using the knowledge for future learning. In the context of the combination of topic modeling with LML, it was first realized in [8], which proposed the Lifelong Topic Model. However, the model is not for opinion mining and it did not jointly model the four dimensions as we do in our work. We believe that the idea of LML can be a promising direction for addressing the above issue, because a system (or a model) that has worked on many domains and retained the discovered knowledge should be able to utilize them to help opinion mining. It is like we humans gain experience from the past and it can guide our future behaviors.

Specifically, LAST is a knowledge-based topic model that extracts and incorporates knowledge from multiple products or domains. In other words, the knowledge is automatically mined from the model results in other domains, including the discovered aspects, opinions, and aspect-opinion pairings (e.g., aspect *screen* and opinion "smooth"), and then assists the modeling of the target domain or a new coming domain. The knowledge transfer is feasible because there is a considerable amount of aspect and opinion overlapping or sharing across domains. Note that we do not use the past results directly but will perform an additional mining to discover more reliable and general knowledge to be used in the new task/domain. The rationale is that when some words appear in the same topic across many past domains, it indicates that these words are likely to be related. Following the previous example, there are other domains like Tablet and Cellphone that are likely to have touch-screens, and the words "smooth" and *screen* may co-occur very frequently in those domains. Based on such domains, we can extract the knowledge indicating "smooth" is likely to be an aspect-specific opinion to *screen*. Back in the domain Laptop, such knowledge can be leveraged to guide the model to discover the similar relationship.

In term of the mined knowledge, there are 3 types that we consider in this paper. We use another aspect *shipping* as an example for explanation (this example will be further discussed in Section 5):

1. Aspect-opinion pair, e.g., {shipping, quick}.
2. Aspect-aspect pair, e.g., {shipping, delivery}.
3. Opinion-opinion pair, e.g, {quick, fast}.

Each type of knowledge comes from aspects, opinions, and aspect-opinion pairings respectively (see Section 4.1). To leverage the extracted knowledge, we use the *generalized Pólya urn* (GPU) model, which will be illustrated in Section 4.2. Briefly, the key advantage of LAST is that it is able to mine more aspect-specific opinions that are coherent with the corresponding aspect as well as higher quality aspects, by extracting and leveraging prior knowledge automatically without any human invention.

In summary, this paper makes three main contributions:

1. It proposes a novel fine-grained holistic topic model, called JAST, to deal with four dimensions in aspect-level sentiment analysis, i.e., to identify aspects, opinions, opinion polarity and opinion generality simultaneously.
2. It proposes a more advanced model called LAST that can extract and leverage aspect, opinion, and their correspondence knowledge from multiple domains to further generate better aspect-specific opinions and more coherent aspects. To our knowledge, this is the first work that learns aspect, opinion, and their correspondence knowledge from the results of many domains with lifelong machine learning.
3. It conducts experiments using reviews of 50 different types of products. The experimental results show significant improvements of the proposed models over state-of-the-art baselines.

2. RELATED WORK

Aspect-based opinion mining has been an important research direction [16]. In recent years, various researches have been conducted to perform different sub-tasks. Since our work focuses on topic modeling, we will mainly discuss the existing related works using topic modeling.

The most related works are the joint models that model aspects and opinions. [25] proposed a joint sentiment-topic model (JST). Rather than modeling topics (or aspects) only as in LDA, JST models both opinion/sentiment and aspect as random variables. However, JST does not separate aspects and opinions, and does not tackle the opinion generality problem. Later on, [19] proposed a model called ASUM assuming that *one sentence is generated by one topic or aspect*, i.e., ASUM assigns all words in a sentence to the same topic. It was shown in [19] that the ASUM model outperformed JST. Similar to JST, ASUM does not separate opinions and aspects nor does it address the opinion generality issue. [33] utilized some topical word seed sets as the knowledge to improve the modeling of aspects and opinions. Each seed set consists of a set of seed words for a particular topic. However, their seed sets are manually provided while our proposed method is fully automatic.

[52] provided an approach to separating aspects and opinion words by integrating supervised learning into topic modeling. They also distinguished general and specific opinions, but they do not identify opinion polarity and their supervised component needs manu-

ally labeled data. Their supervised learning model also classifies a word as a background word or not. Some aspect and opinion terms may be lost if they are predicted as background words so we do not adopt it in our model. To address those problems, we utilize an opinion lexicon. On one hand, the opinion lexicon provides information to help identify opinion polarity. On the other, it helps separate aspect and opinion words with reliable prior information. We do not model the background topic explicitly as we observed that in our finer-grained models, background words usually do not have high probabilities in aspect and opinion topics. Thus, they do not cause much problem. Meanwhile, aspect and opinion information will not be lost by misclassifying words to background words in this way. Recently, [47] proposed a novel unsupervised approach for aspect (words) and opinion (words) extraction based on Restricted Boltzmann Machine [42]. However, apart from the opinion lexicon, it also relies on Parts-Of-Speech (POS) tagging and external Google n-gram corpus for prior information estimation, which we do not use. It also requires manual aspect-topic assignment, which we do not adopt.

Though not aimed at opinion mining, a related fine-grained model is reported in [10] for movie recommendation. It combines collaborative filtering and topic modeling. The model covers user, movie, review content, and review rating in a comprehensive manner. Since our work focuses on opinion mining, it does not include user, movie/product, or review rating information, nor is it concerned with recommendation. Thus our model is quite different.

There are also many topic models that have been used for the task of aspect extraction and categorization in, for example, [4, 11, 24, 28, 49, 27, 32, 33, 43, 10, 5, 14, 22, 23, 30, 45, 46]. Although related, their focuses are very different from ours because they do not target at full aspect-based opinion mining. Here we discuss some of the papers to indicate the type of differences. [11] aims to find informative sentences that are related to certain aspects. [49] proposed a joint optimization framework to identify the relationship between opinion, opinion holder and opinion target. [45] uses rating in their topic model, which is not used in our case. In the context of aspect-level sentiment analysis, [23] uses discourse structure to improve the performance. However, we do not consider discourse here. To address the sparsity issue of the cold start items, i.e., items that have less than 10 reviews, [32] proposed the Factorized LDA (FLDA) model with the consideration of ratings. Again, we do not consider rating in our work.

Additionally, there are some other existing generative approaches that model cross-collection and multi-faceted (or multi-dimensional) information or topics. [51] proposed a topic model for comparative text mining. It discovers common topics across multiple collections, and distinguishes the general cross-collection and collection-specific information under a discovered common topic. [38] extended the work and proposed a new cross-collection mixture model to identify cross-culture differences in blogs and forums. Despite the usage of multiple corpora or domains, these works are not for opinion mining and their models also function quite differently as our goal is not to find cross-collection (or cross-domain) commonalities and differences. On multi-faceted or multi-dimensional analysis, a two-dimensional model is reported in [39] that discovers different facets under one topic, e.g., extracting two different perspectives under a specific issue (topic), e.g., Israeli-Palestinian Conflict. A k-dimensional model called factorial-LDA was proposed in [37] and it improved the performance of [39]. [40] further enhanced the factorial-LDA model and adapt it to the task of summarization of drug experiences. Although, to some extent, these multiple-dimensional models are related to aspect extraction or opinion mining (if aspect and aspect-specific opinion are treated as two dimensions), they are not specialized models nor fine-grained models for opinion mining as [25, 19, 52]. [13] introduced a model that can model both the semantic and syntactic information based on a traditional topic model and a hidden Markov model (HMM). However, all these models are quite different from ours in both the goal and model composition. Additionally, none of these existing models is able to automatically learn prior knowledge and use it to improve its model inference and its modeling results.

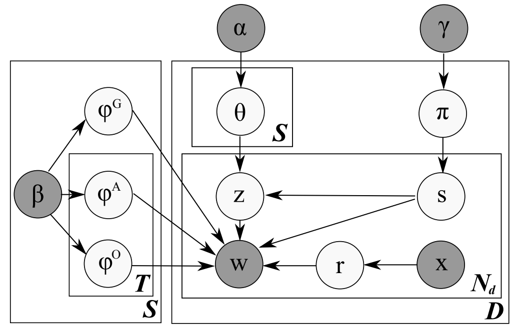


Figure 1: The graphical model of JAST

Since our LAST model can exploit prior aspect and opinion knowledge, it is thus related to knowledge-based topic models such as [2, 18, 41, 33, 17]. However, the knowledge used in these systems are all provided by the user. Our work is also related to transfer learning and lifelong machine learning. Topic models have been used to help transfer learning in [48, 36, 20, 50]. However, transfer learning in these works is for supervised classification and requires human labeling. Our work is more related to [8, 7] which combines topic modeling with lifelong machine learning. [8] considered the positive word correlation as the knowledge while [7] further utilized the negative word correlation. However, they did not separate aspects and opinions and nor did they consider polarity or generality as we do in our fine-grained modeling. Moreover, they did not consider aspect-opinion knowledge and they cannot identify aspect-specific opinions. Also in the context of sentiment analysis, [9] proposed the LSC (Lifelong Sentiment Classification) model that tackles the supervised polarity classification problem. However, it did not use topic modeling and the classification was at the document level, while our models are unsupervised and for aspect-level opinion mining.

3. JAST MODEL

We now present the proposed JAST model, which jointly models aspect, opinion, polarity, and generality. The graphical model is given in Figure 1 and the notations are explained in Table 1.

The generative process is shown as follows:

1. For each document d, we draw a sentiment distribution $\pi_d \sim Dir(\gamma)$;
2. For each sentiment s under document d, we draw a topic distribution $\theta_{d,\ s} \sim Dir(\alpha)$;
3. For each sentiment s, we draw three types of word distributions:
   (a) A general opinion word distribution under sentiment s, denoted as ${\varphi_s}^G \sim Dir(\beta_s)$;
   (b) An aspect distribution under sentiment s and topic k, which is $\varphi^A_{s,\ k} \sim Dir(\beta_s)$;
   (c) An aspect-specific opinion distribution under sentiment s and topic k, $\varphi^O_{s,\ k} \sim Dir(\beta_s)$;
4. For each word w_i in document d:
   (a) choose a sentiment $s_i \sim Multi(\pi_d)$;

S	the number of sentiment polarities
D	the number of documents
T	the number of aspect topics
V	the number of words or terms in vocabulary
N_d	the number of words in document d
s, d, z	sentiment polarity, document, topic
w, x, r	word, lexicon indicator, word type
π	multinomial distribution over sentiments
θ	multinomial distribution over topics or aspects
φ^G	multinomial distribution over general opinion words
φ^A	multinomial distribution over aspect words
φ^O	multinomial distribution over aspect-specific opinion words
α, β, γ	Dirichlet prior for θ, φ, π
w_i, z_i, s_i	word in position i (word i), topic of word i, sentiment polarity of word i
$\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{s}$	all the words or terms in all documents, all the assigned topics, sentiment polarity
$\boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}$	all the assigned topics, sentiment polarity excluding the one assigned to word i
$n_{d,\,l}^{-i}$	the number of words in document d and sentiment l except word i
$n_{d,\,k,\,l}^{-i}$	the number of words under document d and sentiment l and topic k except word i
$n_{k,\,l,\,v}^{-i}$	the number of vocabulary terms v under sentiment l and topic k except word i
$n_{l,\,v,\,c}^{-i}$	the number of words of vocabulary term v under sentiment l and word type c expect word i
$n_{k,\,l,\,v,\,c}^{-i}$	the number of words of vocabulary term v under topic k, sentiment l, word type c expect word i

Table 1: Definition of Notations

(b) choose a topic $z_i \sim Multi(\theta_{d,\,s})$;
(c) choose a word type r_i based on indicator x_i;
(d) emit a word $w_i \sim Multi(\varphi^{r_i}_{s_i,\,z_i})$ or $w_i \sim Multi(\varphi^{r_i}_{s_i})$.

The model has three types of word distributions: φ^G_s, $\varphi^A_{s,\,k}$ and $\varphi^O_{s,\,k}$. φ^G_s indicates a general opinion word distribution under sentiment s; $\varphi^A_{s,\,k}$ and $\varphi^O_{s,\,k}$ are respectively the aspect and the aspect-specific opinion word distributions under sentiment s and topic k. Here we use the opinion sentence "The screen is very clear and great" given in Section 1 again to illustrate. The term "screen" is drawn from $\varphi^A_{s,\,k}$, while the term "clear" and the term "great" are selected from $\varphi^O_{s,\,k}$ and φ^G_s, and they are all under the positive sentiment s.

To model the separation of aspect and opinion, the word type r_i and indicator x_i are introduced. There are several possible approaches to construct these factors. Here we utilize the opinion lexicon, because on the one hand, the opinion lexicon can provide reliable information for polarity and also the identification of aspect and opinion terms, and on the other hand, no manual labeling is needed. So in the JAST model, the observed factor x and the hidden factor r serve for aspect and opinion identification. $x \in \{0, 1\}$ denotes whether a word exists in the opinion lexicon. If w_i appears in lexicon, then $x_i = 1$; otherwise $x_i = 0$. $r \in \{0, 1, 2\}$ indicates the word type of w_i, being an aspect, an aspect-specific opinion, or a general opinion respectively.

JAST assumes that the lexicon words are more likely to be opinion words than non-lexicon words. However, this is a soft constraint. Thus, two supporting elements λ^O and λ^A (see Equation 3) are designed. They are viewed as the prior information for the determination of whether a word is an aspect or opinion. Specifically λ^O controls how much we rely on the lexicon for identifying opinion words (i.e., $x = 1, r = 1$ *or* 2), while λ^A controls how much we believe a non-lexicon word is an aspect word (i.e., $x = 0, r = 0$). Although treating the words that are not in the lexicon as likely aspect terms may not always be correct, our experiments show that the model still generates rational and good results (see Section 5). Based on our observation, simply relying on the lexicon does not cause much problem in a fine-grained model, since the irrelevant words (or background works) are often ranked low in topics due to the naturally pairing of opinion and aspects in the opinion text.

Note that there could be other alternatives to model the identification process of aspect and opinion in JAST. Instead of fully relying on the lexicon, we can estimate the prior information λ^O and λ^A in the JAST model using supervised learning. In other words, those priors can be learned in a supervised manner without the direct auxiliary of the opinion lexicon. Following the works in [52, 33], we also proposed a semi-supervised model which uses a Maximum Entropy classifier as the supervised component. In particular, for each word, we use the surrounding three words as the window. Inside the window, we use the parts-of-speech as features for learning. The labeled data is obtained by checking the words in each sentence with the auxiliary of the opinion lexicon, i.e., if the word appears in the lexicon, it is labeled as an opinion; otherwise, an aspect. This approach saves us from obtaining expensive human labeled data. The advantage of this method over the simply relying on the lexicon is that it can provide more information in terms of identifying other opinion or aspect terms not appearing in the lexicon. We refer to this JAST model variant that integrates with a supervised component as JAST-S. We will see its performance in Section 5.

Inference: We use Gibbs Sampling [12], which is a standard inference technique for topic modeling. The conditional distributions are shown in Equations 1, 2 and 3 (see Table 1 for notations). In our Gibbs sampler, for each word position i in each document d, a topic k and a sentiment l are sampled first and a word type c is sampled after that.

$$\begin{aligned} &P(z_i = k, s_i = l | \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma) \\ &\propto \frac{n_{d,\,l}^{-i} + \gamma_l}{\sum_{l'}^{S} (n_{d,\,l'}^{-i} + \gamma_{l'})} \times \frac{n_{d,\,k,\,l}^{-i} + \alpha_k}{\sum_{k'}^{T} (n_{d,\,k',\,l}^{-i} + \alpha_{k'})} \\ &\times \frac{n_{k,\,l,\,w_i}^{-i} + \beta_{w_i,l}}{\sum_{v}^{V} (n_{k,\,l,\,v}^{-i} + \beta_{v,\,l})} \end{aligned} \tag{1}$$

$$\begin{aligned} &P(r_i = c | x_{w_i}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \alpha, \beta, \gamma) \\ &\begin{cases} g(c, x_{w_i}) \times \frac{n_{l,\,w_i,\,c}^{-i} + \beta_{w_i,\,l}}{\sum_{v}^{V} (n_{l,\,v,\,c}^{-i} + \beta_{v,\,l})} & c = 2 \\ g(c, x_{w_i}) \times \frac{n_{k,\,l,\,w_i,\,c}^{-i} + \beta_{w_i,\,l}}{\sum_{v}^{V} (n_{k,\,l,\,v,\,c}^{-i} + \beta_{v,\,l})} & \text{otherwise} \end{cases} \end{aligned} \tag{2}$$

$$g(r, x) = \begin{cases} \lambda^A & x = 0, r = 0 \\ \lambda^O & x = 1, r = 1 \textit{ or } 2 \\ 1 - \lambda^A & x = 0, r = 1 \textit{ or } 2 \\ 1 - \lambda^O & x = 1, r = 0 \end{cases} \tag{3}$$

4. LAST MODEL

This section introduces the LAST Model. It incorporates aspect and opinion knowledge learned/mined from multiple past domains in our proposed Gibbs sampler using the *generalized* pólya urn model.

Algorithm 1 LAST Learning Algorithm

Input: Target domain corpus $D_{(i)}$ and other domain corpora $\boldsymbol{D}_{(-i)}$

1: $\boldsymbol{A}_{(i)}, \boldsymbol{O}_{(i)}, \boldsymbol{G}_{(i)} \leftarrow$ JAST($D_{(i)}$)
2: /* Step 1. Aspect Matching */
3: **for** each domain $D_{(j)} \in \boldsymbol{D}_{(-i)}$ **do**
4: $\quad \boldsymbol{A}_{(j)}, \boldsymbol{O}_{(j)}, \boldsymbol{G}_{(j)} \leftarrow$ JAST($D_{(j)}$)
5: $\quad$ **for** each sentiment s and topic $t_{(j)}$ **do**
6: $\quad\quad t^*_{(i)} = \min_{t_{(i)}} \text{SKL}(A_{(i)s,\, t_{(i)}}, A_{(j)s,\, t_{(j)}})$;
7: $\quad\quad$ **if** $\text{SKL}(A_{(i)s,\, t^*_{(i)}}, A_{(j)s,\, t_{(j)}}) < \pi$ **then**
8: $\quad\quad\quad S_{s,\, t^*_{(i)}} \leftarrow S_{s,\, t^*_{(i)}} \cup \{(A_{(j)s,\, t_{(j)}}, O_{(j)s,\, t_{(j)}})\}$;
9: $\quad\quad\quad \boldsymbol{S} \leftarrow \boldsymbol{S} \cup S_{s,\, t^*_{(i)}}$;
10: $\quad\quad$ **end if**
11: $\quad$ **end for**
12: **end for**
13: /* Step 2. Knowledge Mining */
14: **for** each $S_{s,\, t_{(i)}} \in \boldsymbol{S}$ **do**
15: $\quad K_{s,\, t_{(i)}} \leftarrow$FIM($S_{s,\, t_{(i)}}$);
16: $\quad \boldsymbol{K} \leftarrow \boldsymbol{K} \cup K_{s,\, t_{(i)}}$;
17: **end for**
18: /* Step 3. Knowledge Utilization*/
19: $\boldsymbol{K}' \leftarrow$KnowledgeFiltering($\boldsymbol{K}$, $\boldsymbol{G}_{(i)}$)
20: $\boldsymbol{A}'_{(i)}, \boldsymbol{O}'_{(i)}, \boldsymbol{G}'_{(i)} \leftarrow$LAST($\boldsymbol{K}'$, $D_{(i)}$);

4.1 LAST Learning Algorithm

As introduced in Section 1, we apply multi-domain knowledge to improve JAST. The overall learning algorithm is shown in Algorithm 1. LAST has the same graphical model as JAST, but the model inference is very different.

The target domain or a new coming domain, is denoted by index i, while other domains or the existing domains (past data prepared for lifelong learning) are indicated by $-i$. The review corpus $D_{(i)}$ and the corpora $\boldsymbol{D}_{(-i)}$ are the inputs.

Step 1: Aspect Matching (lines 1 - 12). This step detects similar aspects generated from existing domains to each aspect in the target domain. With the aspects identified from our proposed fine-grained model, similar aspect matching become easier to realize. Specifically, by running the JAST model, the aspects $\boldsymbol{A}$, aspect-specific opinions $\boldsymbol{O}$ and general opinions $\boldsymbol{G}$ for each domain are extracted. They are represented by their top words ranked by probability. Then, we measure the aspect difference using the Symmetrised KL Divergence (short in SKL) [21]. Given two aspects A_x and A_y, the aspect difference is calculated with equation 4 and we filter the unlikely aspects with a threshold π (line 7). That is, the aspect from other domains j that has no matched aspect in target domain i will not be used. After all aspects generated from $\boldsymbol{D}_{(-i)}$ are processed, we obtain an aspect-opinion set $\boldsymbol{S}$. Each $S_{s,\, t_{(i)}} \in \boldsymbol{S}$ contains a set of matched aspects and their corresponding opinions.

$$SKL(A_x, A_y) = (KL(A_x, A_y) + KL(A_y, A_x))/2 \qquad (4)$$

Step 2: Knowledge Mining (lines 13 - 17). This step mines the knowledge from each $S_{s,\, t_{(i)}}$. We apply Frequent Itemset Mining (FIM) [1] to find those frequently co-occurring words or terms. The reason for using FIM is that a piece of knowledge that appears only in one domain might not be reliable or transferable to other domains. Those pieces of knowledge occurring in multiple domains are more likely to be correct and useful to other domains.

With matched aspects and corresponding aspect-specific opinions, three types of aspect-opinion knowledge are mined from $S_{s,\, t_{(i)}}$: (1) aspect-opinion pair, e.g., {shipping, quick}; (2) aspect-aspect pair, e.g., {shipping, delivery}; (3) opinion-opinion pair, e.g, {quick, fast}. Each piece of knowledge basically says that the two words should belong to the same target topic under sentiment s and topic $t_{(i)}$, or its corresponding aspect and aspect-specific opinion topics. As aspect and opinion are jointly modeled in our framework, they can mutually improve the quality of each other in modeling. Consequently, all three types of knowledge lead to better topic quality. In this paper, we use frequent itemsets of length two, which give us the knowledge as word pairs. After mining, a knowledge set $\boldsymbol{K}$ (line 16) is generated.

Since all the knowledge is generated automatically from the results of unsupervised models, inevitably there are errors, e.g., {shipping, nice} and {nice, quick}. Clearly, "nice" is not specific to *shipping*. As discussed in Section 1, general opinion words like "nice" may be identified as aspect-specific opinions in fully-unsupervised topic modeling. So the knowledge mining process based on the results of JAST may also suffer from it. At this stage, we keep all the knowledge $\boldsymbol{K}$ (including errors). We will deal with them in the next step.

Step 3: Knowledge Utilization (lines 18 - 20). This step uses $\boldsymbol{K}$ to improve modeling for the target domain. We first address the knowledge with errors, e.g., {shipping, nice} and {nice, quick}. Since general opinion words are also modeled in our fine-grained model, they can be used for identifying knowledge errors. Concretely, if the knowledge contains an opinion word found in $\boldsymbol{G}_{(i)}$, that knowledge will not be utilized for the target domain. For example, since "nice" is detected in our generated positive general opinion topic $G_{(i)\ positive}$, the knowledge containing "nice" will be discarded. In other words, we can handle the error by using $\boldsymbol{G}_{(i)}$ to acquire a filtered knowledge set $\boldsymbol{K}'$. The final task is to incorporate the clean knowledge $\boldsymbol{K}'$ into the LAST modeling process. We will illustrate how it works with our proposed sampler in the following sub-section.

4.2 Proposed Gibbs Sampler for LAST

This subsection shows the proposed Gibbs Sampler in LAST, which is different from that in JAST. To leverage the extracted opinion knowledge, we apply the *generalized* Pólya Urn model.

4.2.1 Pólya Urn Model

Pólya urn model [29] is a type of statistical model with self-reinforcing property, sometimes referred as "the rich get richer". It involves with an urn, in which there are balls of different colors. In the formulation of topic model, each color c represents each term/word $v \in V$.

In simple Pólya urn model, at each time, a ball is drawn from the urn. The color of this ball (say c) is recorded and then two balls of the color c are put back into the urn. As a result, the proportion of balls of the color c in the urn increases. The modeling of traditional topic models, such as LDA, is equivalent to the simple Pólya urn model [31]. The limitation of simple Pólya urn model is that it only involves the operation of the ball of one color at each time, i.e., only one word's proportion gets increased.

To overcome the above limitation, the *generalized Pólya urn* (GPU) model allows the procedure of putting back balls of multiple colors. In the GPU model, when a ball of a color is randomly drawn, balls of different colors can be returned to the urn according to the color matrix δ (which is usually specified by the user or by estimation). As a result, these additional balls of different colors added to the urn increase their proportions in the urn. The GPU model was first introduced in topic modeling in [31]. However, they did not use any knowledge. Later, the GPU model was utilized to incorporate knowledge in [8, 7]. In the LAST model, knowing the correlations of two words (say w_a and w_b) from the knowledge, we want to put back some balls of the color representing w_a when drawing w_b, and vice verse.

4.2.2 *Promotion Matrix Estimation*

To use the GPU model, one challenge is how to estimate the matrix δ. In LAST, the problem is how to incorporate the learned prior knowledge into the target domain with proper values, which we call *promotion matrix estimation*. Pointwise Mutual Information (PMI), known as an useful approach to measuring word association in documents [34], is suitable for our task. Here we only use the positive PMI values, as the mined knowledge from multi-domains implies positive semantic correlation. It is finally used to guide the knowledge utilization in LAST with a constraint factor μ (μ>0) which controls how much we believe its indicated values. Now we can compute the promotion rate $PR(w_a, w_b)$ for words w_a and w_b, with the definition given in Equation 5.

$$PR(w_a, w_b) = \mu \times log \frac{P(w_a, w_b)}{P(w_a)P(w_b)} \tag{5}$$

$$P(w) = \frac{\#D(w)}{\#D} \tag{6}$$

$$P(w_a, w_b) = \frac{\#D(w_a, w_b)}{\#D} \tag{7}$$

$P(w)$ indicates the probability of word w occurring in a random document of the target corpus, while $P(w_a, w_b)$ is the probability of co-occurrence of words w_a and w_b in a random document of the target corpus. They are estimated using Equations 6 and 7 where $\#D(w)$ is the number of documents in the target corpus that contain the word w and $\#D(w_a, w_b)$ is the number of documents that contain both words w_a and w_b. $\#D$ is the total number of documents in the target corpus. We can then estimate and leverage the learned knowledge with the promotion matrix for the target domain (Equation 8). Here s denotes the sentiment polarity. t is the topic while i is the domain index.

$$\delta_{s,t_{(i)},w_a,w_b} = \begin{cases} 1 & w_a = w_b \\ PR(w_a, w_b) & (w_a, w_b) \in K'_{s,\, t_{(i)}} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

4.2.3 *Inference*

The conditional distributions for the new Gibbs sampler are given in Equations 9 and 10. The $g(c, x_{w_i})$ in Equation 10 is from Equation 3. The notations are shown in Table 1 except $\delta_{l,\, k\,, w_j,\, w_i}$, which is the matrix defined in Equation 8.

$$\begin{aligned} & P(z_i = k, s_i = l | \boldsymbol{z}^{-i}, \boldsymbol{s}^{-i}, \boldsymbol{w}, \alpha, \beta, \gamma) \\ & \propto \frac{n_{d,\, l}^{-i} + \gamma_l}{\sum_{l'}^{S} (n_{d,\, l'}^{-i} + \gamma_{l'})} \times \frac{n_{d,\, k,\, l}^{-i} + \alpha_k}{\sum_{k'}^{T} (n_{d,\, k',\, l}^{-i} + \alpha_{k'})} \\ & \times \frac{\sum_{w_j}^{V} \delta_{l,\, k\,, w_j,\, w_i} \times n_{k,\, l,\, w_i}^{-i} + \beta_{w_i,\, l}}{\sum_{v}^{V} (\sum_{w_j}^{V} \delta_{l,\, k\,, w_j,\, v} \times n_{k,\, l,\, v}^{-i} + \beta_{v,\, l})} \end{aligned} \tag{9}$$

$$\begin{aligned} & P(r_i = c | x_{w_i}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{w}, \alpha, \beta, \gamma) \\ & \begin{cases} g(c, x_{w_i}) \times \frac{n_{l,\, w_i,\, c}^{-i} + \beta_{w_i,\, l}}{\sum_{v}^{V} (n_{l,\, v,\, c}^{-i} + \beta_{v,\, l})} & c = 2 \\ g(c, x_{w_i}) \times \frac{n_{k,\, l,\, w_i,\, c}^{-i} + \beta_{w_i,\, l}}{\sum_{v}^{V} (n_{k,\, l,\, v,\, c}^{-i} + \beta_{v,\, l})} & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

4.3 Discussion

One may argue that we actually do not need to go through the proposed process to mine and use prior knowledge. Instead, a simple PMI of all pairs of words over all the domains could be used, i.e., for each pair of words, if its PMI value over all the domains is higher than a certain threshold, it is treated as a piece of prior knowledge. This is a valid approach. However, this approach is inferior due to two main reasons. First, as mentioned above, most words do not co-occur with most other words due to the power law distribution in the natural language text. For example, the PMI value of words *price* and *expensive* is small as their co-occurrence is very small. As a result, we will not be able to discover their semantic correlation as a piece of knowledge. However, topic models are able to discover the pair via higher-level co-occurrences. For example, word *buy* may co-occur frequently with *price* in some documents while in some other documents, *buy* may have a high co-occurrence with *expensive*. In such cases, the transitive higher-level co-occurrences can be captured by topic models to produce topics with *price* and *expensive* together under the same topic.

Second, even if we find a pair with a high PMI value, we do not know whether co-occurrences are from a single domain or multiple domains. Frequent co-occurrences in one domain may just indicate this pair of words is specific to that domain and may not be generally applicable. It could also be due to some idiosyncrasy of the data in that domain which causes the high and possibly spurious co-occurrences.

5. EVALUATION

5.1 Candidate Models for Comparison

This section evaluates the following models:

LDA [3]: The classic unsupervised topic model.

ASUM [19]: The aspect and sentiment unifications model. Since it is reported as achieving improvement over JST [25] and is the known closest work to us, it is regarded as our most important baseline. We downloaded the system from the authors' homepage.

ASUM-L: A variation of the ASUM model by applying the opinion lexicon that we use instead of the original seed words in ASUM.

JAST: Our proposed joint aspect-specific sentiment topic model, which models the identification of aspects, opinions, opinion polarity, and opinion generality simultaneously.

JAST-S: A semi-supervised variant of JAST using a Maximum Entropy classifier as the supervised component (see Section 3).

LAST: Our proposed lifelong aspect-based sentiment topic model. It automatically mines and leverages aspect, opinion, and their correspondence knowledge from multiple domains.

5.2 Experiment Setup

Datasets. We use the 50-domains online review corpus created by the authors of [8]. Each domain is a type of products and has 1,000 reviews. We follow their data pre-processing procedure with the standard lemmatization and stop word removal. However, we keep all general opinion words, e.g., *good, nice, great* (while they treated them as stop words and removed them), because general opinion topics are also one of our modeling components.

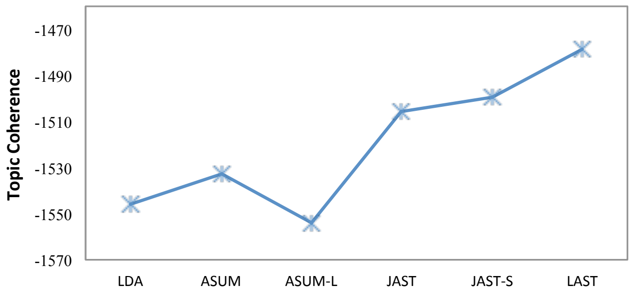


Figure 2: Average Topic Coherence of each model over 50 domains.

Lexicon. We use the opinion lexicon[1] of (Hu and Liu, 2004).
Parameter Setting. All the models are trained using 1000 iterations with 200 burn-in periods. The sentiment number is set as S = 2 for extracting positive and negative opinions. The common parameters are set as $\alpha = 0.1$, $\beta = 0.01$, $\gamma = 1$ and T = 15 for our proposed models based on our pilot experiments. For all baseline models, we try both our proposed parameters and the ones in their original papers, and select the better result for comparison. In LAST, for simplicity, we set λ^A and λ^O to 1, which already generates good results. For learning in LAST, we empirically set π to 7.0, μ to 0.3 and set minimum support for frequent itemset mining to $max(4, 0.7 \times |\boldsymbol{D_s}|)$ for aspect-opinion, $max(4, 0.3 \times |\boldsymbol{D_s}|)$ for aspect-aspect and $max(4, 0.2 \times |\boldsymbol{D_s}|)$ for opinion-opinion pairs, where $|\boldsymbol{D}_s|$ is the number of domains containing matched aspects for a target aspect. The top 15 aspect words and top 15 aspect-specific opinion words are selected to represent aspects $\boldsymbol{A}$ and opinions $\boldsymbol{O}$, which is intuitive as they are the top words for the representation of their topics. These words are used for aspect matching and knowledge mining. For general opinion $\boldsymbol{G}$, the top 25 words are used for representation, which should have more words than aspect-specific opinions by nature. It is also the similar size as the general sentiment seed words used in ASUM. Note that for LAST, each domain works as the target domain while the rest 49 domains serve as the past/existing domains used in mining prior knowledge.

5.3 Topic Coherence

This sub-section reports an objective evaluation based on Topic Coherence proposed in [31]. Topic models are conventionally evaluated using perplexity on held-out test data. However, as shown in [35], perplexity is unable to reflect the real semantic coherence for individual topics. The research in [6] showed that it sometimes even contradicts human judgment. Topic Coherence is now commonly used as a better alternative for assessing topic quality, as it evaluates the coherence and interpretability of topics, which is suitable for our task, as our goal is to make the opinions, along with aspects, more coherent in individual topics.

Figure 2 shows the comparison results. A higher Topic Coherence score indicates a higher topic quality, i.e., better topic interpretation. From Figure 2, we can make the following observations.

1. Our proposed second model LAST achieves the highest topic coherence score. The knowledge from the lifelong learning mechanism greatly benefits the model in discovering higher quality opinions and aspects. Since the aspect and aspect-specific opinion become more coherent, the topic quality is naturally improved. It also shows that the proposed approach in LAST is able to deal with wrong knowledge automatically.
2. Our proposed first model JAST and its variant JAST-S are inferior to LAST, but still outperform all the baseline models, i.e., LDA, ASUM and ASUM-L. As expected, with a supervised component in the JAST-S model, it is able to better identify other potential aspects or opinion words. Note that JAST is also comparable with JAST-S, which demonstrates the reliability of the lexicon as we discussed in Section 3. Comparing with the baseline models (LDA, ASUM, and ASUM-L), we can clearly see that the proposed fine-grained model JAST is able to better group aspects and opinions into interpretable topics, which shows that dealing with four dimensions simultaneously benefits the modeling.
3. ASUM-L has the lowest topic coherence score. This result indicates that it is not guaranteed to achieve improvements by simply using a bigger opinion lexicon. One main reason may be that a sentence could have two words with different opinion polarities or multiple aspects/opinions, which violates the assumption made by ASUM (i.e., one sentence has only one aspect). The results show that the assumption is not suitable for more fine-grained or aspect-specific opinion mining. ASUM and LDA do not perform as well as our proposed models. This again shows the effectiveness of our proposed models.

Statistical tests show that both the improvements for JAST and LAST are significant ($p < 0.001$) against the baselines using paired t-test.

5.4 Topic Quality Evaluation

Here we analyze the results using human judgment. Two human labelers who are familiar with Amazon product reviews are asked to label the results. In the above five models, LDA does not detect opinions, and its resulting topics are also not as coherent as those of the ASUM model. ASUM-L is worse in the objective evaluation than ASUM. The result of JAST-S is similar to JAST. Thus, we primarily compare the JAST and LAST models with ASUM. Note that since ASUM does not separate aspect and sentiment words in a topic, we manually identify and extract the top opinion words appearing in its generated topics. Results from four domains (types of products) are selected for manual evaluation based on the familiarity of the annotators towards the domains.

5.4.1 Opinion Precision

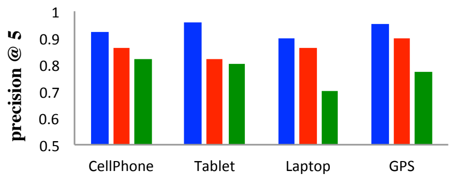
We first evaluate the precision of aspect-specific opinions. We define aspect-specific opinion precision based on the following: a correct opinion word should (a) have the correct polarity and (b) reasonably express opinion about the aspect. For example, for a negative opinion topic for aspect screen, both "fuzzy" and "bad" are correct for aspect *screen*, but "good" and "noisy" are incorrect. Note that here specific and general opinions are not distinguished (we will further evaluate them in the next sub-section).

Evaluation Measure. Since the aspect-based opinion words are generated by the topic model with ranking, we do not know the exact number of correct opinion words, a natural and commonly used metric for evaluation is $precision@n$ ($p@n$ for short), where n is a rank position. We give $p@n$ for n = 5 and 10.

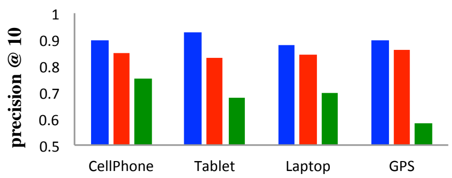
Topic Matching. As different models give different topic (aspect) distributions, we manually match ten best aspect topics for each domain, five positive and five negative respectively, and then compute the average opinion precision for each model.

Result Analysis. Figure 3a and Figure 3b give the average $p@5$ and $p@10$ for each labeled domain. LAST achieves the highest precision for all domains. JAST is also better than ASUM but not as good as LAST. On average, LAST improves ASUM by 15.8% in $p@5$, 22.5% in $p@10$. JAST also improves ASUM by 8.6% and

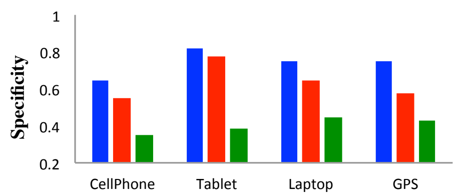
[1]http://www.cs.uic.edu/liub/FBS/sentiment-analysis.html



(a) Precision@5



(b) Precision@10



(c) Specificity

Figure 3: Opinion Evaluation - models in each figure from left to right are LAST, JAST and ASUM

16.8% respectively. Cohen's Kappa agreement scores for p@5 and p@10 are 0.848 and 0.804.

5.4.2 *Opinion Specificity*

We now evaluate whether the identified aspect-specific opinion words are indeed specific. After the previous sub-section, we filter out those incorrect opinion words for further evaluation. There are still two types of opinions, general and aspect-specific opinions. Two example opinion topics are shown for two aspects in Table 2. For example, "great" for aspect *shipping* is not really specific but "quick" is oppositely informative. The opinion words marked in blue are general opinion words, e.g., *problem, bad*. Here we evaluate whether an opinion word is specific enough to give meaningful description about the aspect. We call it *opinion specificity*. Besides the opinion words, the top 20 aspect words of each topic are additionally provided to the annotators (for reference), so that they can better understand what the corresponding aspect should be and then identify correct aspect-specific opinion words.

Evaluation Measure. We calculate the opinion specificity using Equation 11.

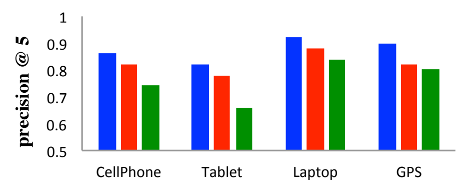
$$Specificity = \frac{n(specific@10)}{n(correct@10)} \tag{11}$$

The annotators evaluate the top 10 correct opinion words (denoted as n($correct$)@10) in every topic. The count of valid aspect-specific opinion words is n($specific$)@10. If n($correct$@10) is less than 5, we do not evaluate that topic, as a very small denominator may lead to a false high value.

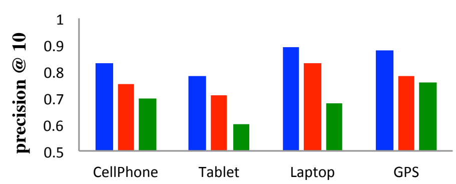
Results: Figure 3c gives the results. We can see that LAST and JAST improve 34.1% and 23.8% over ASUM respectively. A lot of general opinion words with high probabilities are found in ASUM,

Battery (Negative)			Shipping&Order (Positive)		
LAST	**JAST**	**ASUM**	**LAST**	**JAST**	**ASUM**
die	old	*problem*	new	free	*great*
dead	die	hot	free	*happy*	*good*
short	fail	*bad*	fast	fast	quickly
drain	*suck*	die	quick	*pleased*	*well*
fail	useless	*original*	refund	refund	*love*
old	hassle	old	promptly	*recommend*	*perfect*
hassle	*bad*	*new*	original	new	*nice*
wrong	*concern*	*long*	correct	works	*perfectly*
useless	bother	break	works	quick	new
complain	*nervous*	*hate*	accurate	promptly	fast

Table 2: Opinion words for Battery and Shipping&Order aspects. Incorrect opinion words are italicized and marked in red. Non-specific opinion words are italicized and marked in blue.



(a) Precision@5



(b) Precision@10

Figure 4: Aspect Precision - models in each figure from left to right are LAST, JAST and ASUM

e.g., *problem, great, good*, while the opinion words in JAST and LAST are more specific to the aspect. Cohen's Kappa agreement is 0.823.

Example Opinion Topics: Table 2 gives the aspect-specific opinion words of two example aspects. Incorrect opinion words are italicized and marked in red. Non-specific opinion words are italicized and marked in blue. For instance, for aspect $Battery$, *new, original*, and *long* are incorrect as they are not negative aspect-specific opinion words. The words in blue color like *problem, bad*, and *suck* are not aspect-specific, though correct in polarity. We can see that LAST discovers many aspect-specific and coherent opinion words in both example topics.

General Opinions. We also compute the average precision of the positive and negative general opinion words to see whether they are indeed general. The results are: p@10 = 83.8%, p@20 = 79.4% for JAST and p@10 = 85.0%, p@20 = 80.0% for LAST. We use more words here because the number of general opinion words is large. The polarities of top words (no filtering) are all correct. ASUM does not model general opinions.

5.4.3 *Aspect Precision*

For aspect topics, we also report $precision$@5 and $precision$@10 for the four domains. Figure 4a and Figure 4b give their corresponding results averaged over topics of each domain. We observe

Battery			Shipping&Order		
LAST	**JAST**	**ASUM**	**LAST**	**JAST**	**ASUM**
battery	battery	charge	order	arrive	*screen*
charge	charge	battery	receive	receive	receive
hour	life	recharge	arrive	order	arrive
life	hour	*iphone*	shipping	purchase	order
power	*device*	*sd*	ship	*expect*	*privacy*
charger	cable	*card*	today	send	cost
recharge	*phone*	receive	delivery	ship	money
night	*ipad*	replacement	usual	shipping	*monitor*
outlet	power	*purcharse*	*expect*	back	purchase
aaa	plug	*star*	*manner*	*seller*	*seller*

Table 3: Example aspect words for Battery and Shipping&Order. Errors are marked in red.

that LAST achieves dramatic improvements over ASUM. The margins of improvement of JAST over ASUM are also large. LAST is the best, which demonstrates that making use of knowledge learned from past domains is very helpful. Table 3 shows the aspect words of two example topics. We can see the superior performance of LAST. Cohen's Kappa agreement is 0.811. Note that since the objective of our models is essentially for opinion mining in a holistic manner, we do not target at outperforming the existing models that are specialized in the aspect extraction task. Here the results are for showing that, while mining more coherent opinions the joint modeling process can in fact improve the aspect quality as well.

6. CONCLUSION

This paper proposed to jointly model aspect, opinion, polarity and generality. The goal is to provide a holistic solution for the four dimensions and make the extracted aspect-specific opinions more coherent to aspects. For that we first presented a new joint model called JAST that can simultaneously model all the four dimensions, and then introduced a more advanced model called LAST, which can extract and leverage the prior knowledge from multiple domains to improve the performance of JAST, incorporating the idea of lifelong machine learning. Experimental results using reviews from 50 product types show significant improvements over state-of-the-art baseline models.

7. ACKNOWLEDGMENTS

This work was supported in part by a grant from National Science Foundation (NSF) under grant no. IIS-1407927, a NCI grant under grant no. R01CA192240, and a gift from Bosch. The content of the paper is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, NCI, or Bosch.

8. REFERENCES

[1] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[2] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 25–32. ACM, 2009.

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

[4] S. Branavan, H. Chen, J. Eisenstein, and R. Barzilay. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research*, pages 569–603, 2009.

[5] 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*, pages 804–812. Association for Computational Linguistics, 2010.

[6] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

[7] Z. Chen and B. Liu. Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1116–1125. ACM, 2014.

[8] Z. Chen and B. Liu. Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data. In *ICML*, pages 703–711, 2014.

[9] Z. Chen, N. Ma, and B. Liu. Lifelong Learning for Sentiment Classification. In *ACL*, pages 750–756, 2015.

[10] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 193–202. ACM, 2014.

[11] L. Fang and M. Huang. Fine granular aspect analysis using latent structural models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 333–337. Association for Computational Linguistics, 2012.

[12] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[13] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544, 2004.

[14] Y. He, C. Lin, and H. Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 123–131. Association for Computational Linguistics, 2011.

[15] G. Heinrich. A generic approach to topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 517–532. Springer, 2009.

[16] 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*, pages 168–177. ACM, 2004.

[17] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. Interactive topic modeling. *Machine learning*, 95(3):423–469, 2014.

[18] J. Jagarlamudi, H. Daumé III, and R. Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213. Association for Computational Linguistics, 2012.

[19] 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*, pages 815–824. ACM, 2011.

[20] J.-H. Kang, J. Ma, and Y. Liu. Transfer topic modeling with ease and scalability. In *SDM*, pages 564–575. SIAM, 2012.

[21] N. Kawamae. Latent interest-topic model: finding the causal relationships behind dyadic data. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 649–658. ACM, 2010.

[22] S. Kim, J. Zhang, Z. Chen, A. H. Oh, and S. Liu. A hierarchical aspect-sentiment model for online reviews. In *AAAI*, 2013.

[23] A. Lazaridou, I. Titov, and C. Sporleder. A bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *ACL (1)*, pages 1630–1639, 2013.

[24] 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*, pages 1137–1146. Association for Computational Linguistics, 2011.

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

[26] B. Liu. Sentiment analysis and opinion mining (synthesis lectures on human language technologies). *Morgan & Claypool Publishers*, 2012.

[27] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th international conference on World Wide Web*, pages 121–130. ACM, 2008.

[28] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140. ACM, 2009.

[29] H. Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC Texts in Statistical Science, 2008.

[30] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.

[31] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[32] S. Moghaddam and M. Ester. The flda model for aspect-based opinion mining: addressing the cold start problem. In *Proceedings of the 22nd international conference on World Wide Web*, pages 909–918. International World Wide Web Conferences Steering Committee, 2013.

[33] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 339–348. Association for Computational Linguistics, 2012.

[34] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *HLT-NAACL*, pages 100–108, 2010.

[35] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 215–224. ACM, 2010.

[36] S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359, 2010.

[37] M. Paul and M. Dredze. Factorial lda: Sparse multi-dimensional text models. In *Advances in Neural Information Processing Systems*, pages 2582–2590, 2012.

[38] M. Paul and R. Girju. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1408–1417. Association for Computational Linguistics, 2009.

[39] M. Paul and R. Girju. A two-dimensional topic-aspect model for discovering multi-faceted topics. *Urbana*, 51:61801, 2010.

[40] M. J. Paul and M. Dredze. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *HLT-NAACL*, pages 168–178, 2013.

[41] J. Petterson, W. Buntine, S. M. Narayanamurthy, T. S. Caetano, and A. J. Smola. Word features for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 1921–1929, 2010.

[42] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 448–455, 2009.

[43] C. Sauper and R. Barzilay. Automatic aggregation by joint modeling of aspects and values. *Journal of Artificial Intelligence Research*, 2013.

[44] S. Thrun. Lifelong Learning Algorithms. In S. Thrun and L. Pratt, editors, *Learning To Learn*. Kluwer Academic Publishers, 1998.

[45] I. Titov and R. T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, pages 308–316. Citeseer, 2008.

[46] 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*, pages 783–792. ACM, 2010.

[47] L. Wang, K. Liu, Z. Cao, J. Zhao, and G. de Melo. Sentiment-aspect extraction based on restricted boltzmann machines. 2015.

[48] G. Xue, W. Dai, Q. Yang, and Y. Yu. Topic-bridged PLSA for cross-domain text classification. In *SIGIR*, pages 627–634, 2008.

[49] B. Yang and C. Cardie. Joint inference for fine-grained opinion extraction. In *ACL (1)*, pages 1640–1649, 2013.

[50] S.-h. Yang, S. P. Crain, and H. Zha. Bridging the language gap: topic adaptation for documents with different technicality. In *International Conference on Artificial Intelligence and Statistics*, pages 823–831, 2011.

[51] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 743–748. ACM, 2004.

[52] 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*, pages 56–65. Association for Computational Linguistics, 2010.

[53] G. Zipf. Selective studies and the principle of relative frequencies in language, 1932.